**Predicting Acadian Flycatcher (*Empidonax virescens*) territory locations in Iowan forests**

Benjamin M West

# Abstract

The Acadian flycatcher (*Empidonax virescens*) is a forest-dwelling bird that is a Species of Greatest Conservation Need in Iowa, USA. This designation means that the ability to predict territory locations for this species could be useful for spatially targeting management and conservation activities. The goal of this project was to compare the abilities of three types of model to predict the presence of Acadian flycatcher territories; each model employed 17 habitat metrics as predictors. Using bird point count data and habitat metrics at 493 collected in forested areas in south-central Iowa in 2019, I created and tuned a classification decision tree, a random forest, and a stochastic gradient boosting machine (LightGBM) for comparison. Models were compared using a hold-out set and were assessed with overall accuracy, sensitivity, specificity, and AUC. All model types had similar, fairly high accuracy; the decision tree and random forest models tied for highest accuracy (0.788), and the LightGBM had the lowest accuracy (0.768). The decision tree model had the highest sensitivity (0.846), and the random forest model had highest specificity (0.836) and largest AUC (0.899). The combination of high accuracy, high sensitivity, and greater interpretablity of the decision tree model made it the most preferable model in the set, but larger bird point count datasets may still benefit from more complex machine learning methods.

# 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 forests of south central Iowa, USA. The Acadian flycatcher is an avian Species of Greatest Conservation Need in Iowa, making it one of many priority species for monitoring and conservation in that state (Iowa Department of Natural Resources, 2015). To support the goal of predicting sites in Iowa 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. Data were collected as part of a larger project describing the distributions, densities, and habitat relationships of birds for the purpose of making forest management decisions. For all field surveys (bird and habitat), 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 point survey, an Acadian Flycatcher territory considered “present” at that point for the 2019 breeding season. 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). Between the point and landscape scales, there was a total of 17 habitat metrics.

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 test set. Due to class imbalances (105 present, 289 absent in the training set), 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; this resampling converted the original training set of 394 unique observations to a partially resampled training set of 578 observations. Equal sample sizes for each class were used in an attempt to balance model sensitivity and specificity. 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 ability across 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’ (Robin et al., 2011), and other metrics were generated from confusion matrices produced by ‘caret’ (Kuhn, 2021).

# 3. Results

The proportion of sites occupied by singing Acadian Flycatchers was 0.263, creating a non-information rate of 0.737. Overall accuracy was higher than the no information rate across all models; it ranged from 0.768 to 0.788 (Table 2) Of the three models, the random forest and decision tree models tied for highest accuracy, and the LightGBM had the lowest accuracy (Table 2). Sensitivity was highest for the decision tree model, and specificity was highest for the random forest model; sensitivity for the random forest model was markedly lower than the other two model types (Table 2). AUC was high across all models (Table 1), ranging from 0.869 (decision tree) to 0.899 (random forest).

Table 2. Loss metrics for three predictive models used to predict Acadian flycatcher territory presence in southern Iowa forests. These metrics were calculated by comparing model predictions to a holdout dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Type** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| Decision Tree | 0.788 | 0.846 | 0.767 | 0.869 |
| Random Forest | 0.788 | 0.654 | 0.836 | 0.899 |
| LightGBM | 0.768 | 0.808 | 0.753 | 0.889 |

# 4. Discussion

Given that the models were trained with a relatively small, partially resampled sample of 578 observations, it is not surprising that a simple decision tree model performed identically to a random forest model in terms of overall accuracy. However, the decision tree model did the best job of predicting presence of Acadian Flycatcher territories, i.e., it had the highest sensitivity of all three models. In contrast, the random forest model had markedly low sensitivity, but it had the highest specificity of all three models, i.e., it did the best job of predicting absences of territories. The low sensitivity of the random forest may be due to resampling from a small pool of unique positive observations, giving the model few unique observations within the positive class. In comparison to the other two models, the LightGBM model had slightly worse overall accuracy; this poorer performance by this metric may be due to the relatively small dataset and/or suboptimal tuning. Additional observations, both in terms of of points surveyed and number of predictors, may have improved performance of all models.

These models could be used to predict Acadian flycatcher territory presence in Iowa forests, allowing for informed management decisions for Acadian flycatchers without employing expensive bird monitoring programs. Given the small forest patch sizes and agriculturally dominated landscape of Iowa, these models may not be applicable to other areas. Further research could include other bird species, examine the effects of dropping more labor-intensive habitat metrics, (e.g., midstory density) or explore the predictive consequences of using only remote-sensed data. The combination of interpretability by managers, high accuracy, and high sensitivity of the decision tree made it the preferred model in this exercise, but machine learning methods may still be useful for larger bird point count datasets.

# References

Bakermans, M. H., & Rodewald, A. D. (2006). Scale-Dependent Habitat use of Acadian Flycatcher (Empidonax virescens) in Central Ohio. *The Auk*, *123*(2), 368–382.

Iowa Department of Natural Resources. (2015). *Iowa’s Wildlife Action Plan: Securing a future for fish and wildlife.*

Kuhn, M. (2021). *caret: Classification and Regression Training*. https://cran.r-project.org/package=caret

MacArthur, R. H., & MacArthur, J. W. (1961). On bird species diversity. *Ecology*, *42*(3), 594–598.

Mahoney, M. J. (2021). *Machine Learning Concepts and Applications*.

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*.

Reidy, J. L., Thompson III, 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*, 34–46.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Rodewald, A. D., & Abrams, M. D. (2002). Floristics and avian community structure: implications for regional changes in eastern forest composition. *Forest Science*, *48*(2), 267–272.

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., & others. (2019). Decline of the North American avifauna. *Science*, *366*(6461), 120–124.

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*.

Therneau, T., & Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*.

USDA National Agricultural Statistics Service Cropland Data Layer. (2019). *Published crop-specific data layer [Online]*. https://nassgeodata.gmu.edu/CropScape/

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