# Spatially explicit models to test climatic and land use influences on the distribution of Oklahoma grassland birds

Claire M. Curry

Jeremy Ross

Andrea Contina

Eli S. Bridge

# Introduction

Grasslands are one of the world’s most endangered ecosystems, with declines of 82.6-99.9% of tallgrass prairie, 30-99.9% of mixed-grass prairie, and 20-85.8% of short-grass prairie in the plains states and provinces of North America (Samson and Knopf 1994). Drivers of decline include land use conversion via agriculture and changes in fire and grazing regimes (Samson et al. 2004). Grassland report (Askin et al?). These habitat losses have resulted in grassland birds being one of the fastest-declining groups in North America (Samson and Knopf).

The already tenuous status of grassland birds is further threatened by conversion to new crops resulting in permanent land use changes (Wright and Wimberly 2013), generational changes in land use (Higgins et al. 2002), changes in conservation programs for grassland habitats (Klute et al. 1997), and alterations to vegetation (Alward 1999) and ecosystem structure (Brown et al. 1997) from climate change (McCarty 2001). Grassland bird species are declining faster than other groups of birds as well (Sampson and Knopf 1994) and thus continued to be imperiled by these new threats to their habitat.

In the southern Great Plains, the U.S. state of Oklahoma contains a wide variety of grassland birds as its ecoregions range from tallgrass prairies in the east on the edge of eastern deciduous forests to several types of grasslands in the central part of the state and westward. Historically, each type of grassland comprised ?, ?, and ? ha, totaling ? ha (Sampson and Knopf paper). Modern estimates by comparable schemes are not available. Agriculture accounts for over $2.8 billion in the state’s gross domestic product in the study years (U.S. Bureau of Economic Analysis, 2017). Major crops in the state use. Land use is also for ranching, with ?? animals. State plans for biofuels GOOGLE. This combination of the state’s agricultural importance and forecast impact by climate change makes Oklahoma’s grassland birds vulnerable, and models to predict what factors will affect their distribution are important to effective management.

Species distribution modeling uses predictors, usually climatic variables () but now expanding to biotic variables (), to predict what areas are most suitable for a given organism, and estimate what variables constrain a given species’ range. Range-wide predictions have been made for grassland birds (O’Connor et al. 1999) but some species with smaller ranges were not accurately modeled, perhaps because different drivers of distribution are important in different regions (Bakker et al. 2002). A new method called Spatiotemporal Exploratory Modeling (STEM) adds additional accuracy by using an ensemble of smaller, regional models to account for local differences in variables that drive distributions (Fink et al. 2010 and 2013). The original model was developed for data with a continent-wide scale and seasonal variation. Understanding impacts of changing land use and climate in a region with much ecosystem variety requires a dynamic approach to species distribution modeling, so we will use this approach of regional ensemble models at a smaller scale, within one state, within a single season (the breeding season). We will compare the spatially explicit ensemble’s effectiveness with a typical SDM at the statewide scale. The objectives of our study are to estimate the current distribution of Oklahoma grassland birds and understand what variables are important in their distribution. These data will allow managers to make decisions on what areas are important for populations, what land use practices and trends may impact populations, and how climate change interacts with these. Considering management at the local scale is also important because many land use changes are made there (citation??).

We will examine three aspects of Oklahoma grassland bird distribution. First, what is the current distribution of these species? We will use point count, transect, and citizen science data to create density estimates (from point count and transect data, comparing estimates because they are different habitats) and species distribution models (from all three data sources). We will create statewide distribution models and compare these with spatially explicit ensemble models. Second, what predictor variables determine the distribution of each species and as such what land use changes might make these species vulnerable? We look at land use, conservation easements, and climatic variables, and ask which variables were most important in the statewide species distribution models. Finally, how will distributions change with climate change forecasts and potential land use changes? We use predicted climate change forecasts and estimates of potential land use change (soil distribution maps??) to estimate risks for Oklahoma’s grassland birds.

# Methods

## Study area

Brief discussion of ecoregions in Oklahoma, range of precipitation and temperature across the state, and what types of grasslands (and what areas exist) are here.

## Response data

* Survey methods
  + Point counts
  + Transects
* Citizen science data
  + eBird: All complete data (points and transects) for 2013 and 2014.
  + Because some observers entered sightings from before and during our surveys into eBird.org, we eliminated counts that were within two hours of the actual survey start time and within 15 km of the survey start location. This eliminated ?? of ?? ebird entries.

## Predictors

We used climatic variables, land use variables, and land cover variables to predict bird distribution. Table 1 shows the variables, sources, and their definitions. Neighborhood predictors were calculated by the values in rectangular areas around each point, at the scale of 5 x 5 pixels (150 x 150 m) and 15 x 15 pixels (450 x 450 m) (Fink et al. 2010). We looked at proportion of each land cover class and proportion of several summed variables: open space (grasslands, hay/pasture, cropland, herbaceous wetlands, and barren land) since grassland bird occupancy can be influenced by the total non-structural cover (McDonald 2017). Neighborhoods were created in QGIS 2.14 with GRASS’s r.neighbors processing tool.

## Analyses

### Density estimations

* Using distance sampling, possibly including detectability from repeated surveys
  + Comparison of point count vs transect effectiveness if sample size large enough for each and geographical overlap sufficient. However, point counts go along road and transects usually walking off-road. Alternative: comparison of estimations from road pcs vs “off road” transects?

### Species distribution models

To model species distributions based on our predictors, we created four models for each species. The first is a statewide model with random forest trees, which give results competitive to other machine learning results such as boosted regression trees and bagged decision trees, with minimal tuning parameters required (Caruana and Niculescu-Mizil 2006, Cutler et al. 2007). The statewide model will allow ranking of variable importance. This gives us interpretable models for which we can make specific predictions about what predictor variables are influencing distribution in what ways for each piece of the ensemble. The second are three spatiotemporal exploratory models (STEM) (Fink et al. 2010) at varying scales. This second model, while it may give more accurate predictions, is harder to interpret because of its increased local accuracy (James et al. 2013). This is important because what influences distribution may vary by region (Bakker et al. 2002). STEM are merged over different spatial extents per Fink et al. 2010. The original STEM was used on continent-scale survey data and can be used with any base model (Fink et al. 2010, Fink et al. 2013). We see whether it is useful at a smaller scale by adapting the scale of our support sets to the state extent. With the diverse habitats and climatic variables found across Oklahoma, it should provide better predictions than the statewide model. As our survey dataset covers the breeding season only, we did not specify temporal windows for this model (unlike Fink et al. 2010) and used data from April-June. Hence, we will refer to our models as “spatially explicit” in this paper while adapting their spatial averaging ensemble design. Using both the statewide and spatially-explicit models give us complementary information on factors affecting species distribution in Oklahoma.

We had 7605 complete checklists. We used caret’s createDataPartition which randomly samples within each factor (presence and absence) to create balanced split of data with 6085 checklists for training. The remaining 20% data (1520 checklists) were used to independently evaluate model performance after training (see below). For the spatially explicit ensemble model, we generated 1000 random locations across the region extent. Each was surrounded by a square of ?? m. Each polygon was a support set. Data from the training set within this support set polygon was used for each individual model. The support set was discarded if it contained fewer than 10 checklists. The statewide model was trained with the full training set (6085 checklists). This number of checklists is good (citation in one of the ml articles, caruana et al 2005 maybe?)

Models were created using boosted regression trees (as base models for the spatially explicit ensemble model and as a statewide model) in R’s ‘caret’ package following Elith et al. 2008. Each support set included all checklists from the training dataset within a randomly centered square of size X x X. Each support set model was trained with five-fold cross validation on the its subset of training data. Each support set model was predicted using the raster package’s predict function to create a support set map for each support set. These maps were stacked statewide using the mosaic() function to get the mean value of each pixel. This process was repeated at three support set spatial scales, resulting in three spatially explicit ensembles per species. The statewide model was also trained with five-fold cross validation on the training data set (6085 checklists). The model was again predicted using raster::predict on the model.

To evaluate model performance (three spatially-explicit ensembles and one statewide model per species), we created a statewide grid of ? x ? m cells. We randomly sampled no more than 5 observations from each grid cell (this maximum is to reduce spatial bias of areas with large numbers of checklists) in the evaluation dataset (1520 checklists). We repeated the spatial sampling procedure randomly 50 times to get a distribution of the model performance (Fink et al. 2010) for each model. We tested four models in this way for each species and chose the scale (small, medium, large, statewide) with best performance measures for each species.

To determine which predictors were important in species distributions, for each species we ranked variables at the statewide level. Hochaka et al 2007 article lists citations of Breiman 2001, Brieman et al. 1984, and Caruana et al. 2006 of how to rank important variables. Elith et al. 2008 shows how to do partial dependence plots nad variable importance ranks based on

### Study species

Table ? shows the species we analyzed.

# Results

Map for STEM-type, basic model, and distance sampling estimates for each species.

# Discussion

We should be careful extrapolating climate change to extinction (Schwartz et al. 2006).

# Acknowledgements

This work was funded by USDA-NIFA grant #2013-67009-20369 to ESB.

# Notes to self

* Survey vegetation data from 2014 transects (none from 2013 or point counts in 2014?)… only can be used for that year and for 1/3 of 2014 transects. Not sure what it can be used for in this context of state-wide multiple years.
* Data I need to find if exists:
  + Forecast changes in landuse in OK
    - <http://tethys.dges.ou.edu/main/?cat=12>

Links for ensemble model making:

* + - How to implement
      * <http://machinelearningmastery.com/non-linear-classification-in-r-with-decision-trees/>
      * <https://cran.r-project.org/web/packages/ipred/vignettes/ipred-examples.pdf>
      * <https://cran.r-project.org/web/packages/adabag/adabag.pdf>
      * <https://onlinecourses.science.psu.edu/stat857/node/181>
      * <http://mlwave.com/kaggle-ensembling-guide/>
      * Simple averaging ensemble pseudocode: <http://www.kdnuggets.com/2016/02/ensemble-methods-techniques-produce-improved-machine-learning.html>
      * using caret to assemble ensembles?? <http://amunategui.github.io/blending-models/>
      * <http://www.overkillanalytics.net/more-is-always-better-the-power-of-simple-ensembles/>: has code, I think I can start from this.

Table 1. Predictors used in models.

|  |  |  |
| --- | --- | --- |
| Predictor | Definition | Source |
| NRCS Conservation Easement Areas by State | Presence or absence of a conservation easement | <http://gws.ftw.nrcs.usda.gov/GWDL/3276698/easements_EASEAREA_ok_3276698_01.zip> |
| NRCS Conservation Easement Areas by State Calculated Area | Size of the conservation easement in which the pixel exists | <http://gws.ftw.nrcs.usda.gov/GWDL/3276698/easements_EASEAREA_ok_3276698_01.zip> |
| NLCD2011 Landcover | NLCD class | <http://gws.ftw.nrcs.usda.gov/GWDL/3276698/land_use_land_cover_NLCD_ok_3276698_02.zip> |
| NLCD 2.25 and 20.25 ha neighborhoods | 5 x5 and 15 x 15 pixel neighborhoods |  |
|  | Open space (11, 31, 71, 81, 82, 95) |  |
|  | Open water 11 |  |
|  | Developed open space |  |
|  | Low intensity development (22) |  |
|  | Medium intensity development (23) |  |
|  | High intensity development (24) |  |
|  | Barren (31) |  |
|  | Forest (41, 42, 43) |  |
|  | Scrub/shrubland 52 |  |
|  | Grasslands 71 |  |
|  | Pasture and hay 81 |  |
|  | Croplands 82 |  |
|  | Woody wetlands 90 |  |
|  | Herbaceous wetlands 95 |  |
| Year |  |  |
| Day of year |  |  |
| Hour |  |  |
| Longitude |  |  |
| Latitude |  |  |
| Human population density | Number per square km |  |
| Bioclim variables |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Table . Study species with their conservation status.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | Breeding bird survey trend since YEAR | Oklahoma status | Federal status | IUCN status |
| Dickcissel |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |