Spatial Modeling of Bird Populations using Citizen Science Data

Abstract

Observation count data from eBird can be used to model the relative abundance of bird species. We found that such data is generally overdispersed compared to a Poisson distribution and that a quasi-Poisson generalized additive model is appropriate for the data. Expanding on previous research for eBird data, we incorporated spatial dependence into the modeling task by performing hierarchical generalized additive modeling with a spatial conditional autoregressive structure for random effects. We found that our data contains moderate spatial dependence and that models that account for spatial dependence have superior predictive performance to those that do not. We conclude that quasi-Poisson hierarchical generalized additive models with spatial random effects provide the best representation of the relative abundance of bird populations. Moreover, our spatially explicit models are more realistic based on domain knowledge when regarding the impact of environmental covariates, which is important when considering conservation implications and future projections.

1 Introduction

Birds across the United States and Canada are disappearing at an alarming rate. 2.9 billion birds, over one-fourth of the total bird population in this region, have been lost since 1970. This is a crisis that affects even common species not generally thought to be under threat. The blue jay population has lost a quarter of its birds; the red-winged blackbird population has lost a third [9] [16]. It is critical to have methods for understanding the current population distributions of bird species so that we can better protect them.

1.1 eBird

eBird is a citizen science project managed by the Cornell Lab of Ornithology where registered users record the observations of birds they identify in the form of checklists. A checklist is a list of birds observed by one or more eBird users during a specific birdwatching session accompanied by relevant metadata. Such metadata includes the time duration of the event, the distance traveled during the event, and the number of observers. The survey protocol type of the checklist is also recorded; the two standard types are stationary and traveling [4]. Additionally, an eBird user must select a geographic point location to represent each checklist. Finally, an eBird user must mark each checklist as either complete or incomplete.

The complete checklist is a key concept of the eBird checklist system. A checklist is complete if an eBird user gave their full effort to noticing all the birds around them; they tried to identify every bird they observed to the best level of precision and accuracy possible; and they included every species that they noticed and identified on their checklist [2]. A complete checklist does not require an eBird user to identify every bird they encountered by species, since this would require expert-level ability. Instead, a checklist is incomplete if an eBird user intentionally does not record any wild bird species "that was present, detected, and identified" [2].

Checklists submitted to eBird are put through an automatic data verification process. Filters flag any problematic checklists or species observations; expert volunteers then manually review this flagged data. There are specific protocols for both individual species observations in a checklist, hereafter referred to as "observations", and checklists overall [3]. In our research, we will use only observations and checklists that have successfully passed through this process.

1.2 Study Area and Selected Species

Our study area was Bird Conservation Region (BCR) 31, which corresponds to peninsular Florida. BCRs are defined by the North American Bird Conservation Initiative as "ecologically distinct regions in North America with similar bird communities, habitats, and resource management issues" [1]. BCR 31 is biologically rich in species thanks in part to its humid and conducive climate and also in part to its position between the tropical Caribbean and temperate North American climates [26]. It contains a variety of saltwater, freshwater, and terrestrial habitats. In particular, there are many coastal and interior wetlands that provide habitat for wading birds [26]. Unfortunately, these wetlands have been harmed by agricultural and urban runoff that degrades water quality, in addition to drainage [26]. In general, the Floridian peninsula faces population and land development pressure from humans.

In our research, we will work with the following ten species: white ibis (*Eudocimus albus*), glossy ibis (*Plegadis falcinellus*), roseate spoonbill (*Platalea ajaja*), great egret (*Ardea alba*), cattle egret (*Bubulcus ibis*), snowy egret (*Egretta thula*), great blue heron (*Ardea herodias*), little blue heron

(*Egretta caerulea*), tricolored heron (*Egretta tricolor*), and green heron (*Butorides virescens*). Each species is resident in peninsular Florida year-round [7]. Since these species are fairly similar and all present across BCR 31, we are able to make reasonable comparisons of the models generated for each species. Furthermore, all species are easy to both find and identify correctly due to their large body sizes and distinct characteristics. The observation data for these species is likely to be more accurate than it would be for species that are easier to miss or identify incorrectly. It has been proposed that citizen science data may be of higher quality for species that have large body sizes and are easy to identify correctly [37].

1.3 Previous Research

In [24], Johnston et al. propose statistical processes to refine citizen science data, in particular eBird data, to estimate species distributions. The authors recommend the use of the following two strategies in conjunction: imposing a structured protocol onto citizen science data using filters and including covariates that account for variation in effort on the part of observers. Together, these strategies improve the predictive performance of models fit using citizen science data. While the models in [24] are made for the metrics of encounter rate and occupancy probability, the results of the article can be extended to similar metrics of species distribution.

Johnston et al. provide a supplement that explains how to implement the procedures of [24] using eBird data and the R software [25]. In this supplement, the authors also discuss the concept of relative abundance and techniques for modeling it, such as a quasi-Poisson generalized additive model (GAM). The abundance of a species is the true number of individuals of that species in a given area. However, we cannot measure abundance directly due to the nature of eBird data collection and birds themselves as animals that frequently move. We therefore have to use relative abundance to stand in for true abundance. In the context of eBird, relative abundance is the count of individuals of a species observed in an eBird checklist.

Given the spatial nature of bird observation data, it could be beneficial to explicitly account for spatial dependence when modeling such data. Spatial autocorrelation, a type of spatial dependence, exists when observations gathered from closer locations have either higher or lower similarity. When modeling species distributions with ecological data, incorporating spatial autocorrelation has been shown to improve both model fits and predictions for species occurrence; if spatial autocorrelation is present, failing to incorporate it into the modeling process will lead to a biased model fit [15] [28].

There are three categories of factors that introduce spatial autocorrelation into species occurrence data. The first category is internal factors; these originate from the true patterns of the species under consideration and can not be addressed with any non-spatial modeling method. Internal factors include species dispersal patterns and colonial breeding habits [15]. The second category is external environmental factors, which have their own pattern of spatial autocorrelation that they introduce into species data [12]. Potential external factors include humidity, rainfall, and soil type. If external factors can be included in the model as environmental covariates, then spatial autocorrelation in model residuals can be reduced or even eliminated. However, if these factors cannot be included in modeling due to a lack of data, incorporating spatial dependence into the modeling process is a good strategy. The final category is additional missing factors such as conservation management practices or uneven effort on the behalf of data collectors. Adding non-environmental covariates to the model, such as covariates to represent observer effort, can resolve spatial autocorrelation caused by these miscellaneous factors. As with external

factors, if we cannot include these factors as covariates, we can resolve the issue by incorporating spatial dependence into our modeling.

In practice, we can examine whether or not our data has spatial autocorrelation, but we cannot know what factors introduced this spatial dependence into our data. A good method is to include all environmental and non-environmental covariates relevant to our data and then examine if there is any remaining spatial autocorrelation. If there is remaining spatial autocorrelation, we should address it by incorporating it into our modeling.

In [28], Lee et al. used a quasi-Poisson hierarchical generalized linear model (HGLM) with a spatially correlated conditional autoregressive (CAR) structure for random effects to model count responses from species observation data with excess zeros. The authors achieved strong performance results from this model type. These results are applicable to research performed with eBird count data, since such data are also spatial population data with excess zeros. This is a suitable method for incorporating spatial autocorrelation into our modeling task.

1.4 Research Question

A slight limitation of the work performed in [25] is that the authors do not evaluate whether or not using a GAM improves predictive performance when compared to using a generalized linear model (GLM). Since GLMs are simpler than GAMs, they should be selected over GAMs when model performance is similar between the two. To address this, we will generate a GLM and a GAM for each of our selected distributions. We will then compare the predictive performances of these GLM and GAM model fits to determine if GAMs improves predictive performance.

A crucial limitation of the work performed by Johnston et al. in [24] and [25] is the lack of any mention of potential spatial dependence in the data. In particular, the model types used for relative abundance in [25] assume independence between observations. However, eBird observations are spatial ecological data and may potentially be spatially dependent. To address this limitation, we will create HGLMs with spatial random effects and evaluate the presence of spatial dependence in the data. We will provide more details about these methods in Section 2.

In this paper, we will investigate statistical methods for modeling the relative abundance of bird populations using eBird citizen science data. We will synthesize and evaluate the practices proposed in [24], [25], and [28] in the context of our data. A key focus of our research is to investigate the effects of incorporating spatial dependence into the modeling of relative abundance. We will begin in Section 2 by discussing best practices for preparing citizen science data, generalized additive modeling, and methods for incorporating spatial dependence into a modeling task. In Section 3 we will investigate the predictive performances of models and evaluate the best model for our data. Finally, in Section 4 we will discuss the meaning of our results in the context of our research task and any areas of future research.

2 Methods

2.1 Data Preparation

2.1.1 Data Sources

Our primary data source was the eBird database, which contains all validated observations and checklists. Each observation can be matched with its corresponding checklist. If a checklist is

complete and has no corresponding observation for a certain species, we can infer that the given species was not detected. If a checklist has a corresponding observation for a certain species but the eBird user did not record a count value for that observation, the species has an "X count" for that checklist. X counts must be removed before we can model relative abundance. This is unfortunate, since X counts exclusively correspond to species detections, and therefore removing X counts from a species' dataset will change the structure of that dataset. However, it is necessary.

eBird checklists do not contain any information on the land surrounding a checklist's location. We will use supplementary data sources to generate relevant environmental covariates. For elevation data, we will use the EarthEnv project. For information on land cover, we will use the MODIS Land Cover Type Product, also referred to as MCD12Q1. The MCD12Q1 product is in the form of annual grids of 500 meter by 500 meter tiles that map land cover classes. We will use the University of Maryland legend for land cover, which contains 16 land cover classes. For more information on the MCD12Q1 product, see [39].

2.1.2 Data Filtering

We performed the following procedure for each of the ten species in our analysis.

First, we created a set of filters to extract observations and checklists from the eBird database. We filtered for observations of the selected species in BCR 31 in the month of June that were from complete checklists with the stationary or traveling protocol. We also filtered for checklists that were in BCR 31 in the month of June and were complete with the stationary or traveling protocol. We then formatted the data by merging observations and checklists such that each row corresponded to a checklist and contained the observed count for the selected bird species. If no observation for the selected species existed for a checklist, that checklist was given a count of 0.

Next, we filtered the data to reduce variation in effort. We restricted checklists to those that were no more than five hours long, had no more than five kilometers in distance traveled, and had no more than 10 observers. These filters impose a standard method of data collection, which has been shown to improve the performance of models fit using eBird data [24]. This set of filters was proposed in [25].

2.1.3 Effort Covariates

Data from citizen science sources such as eBird can be challenging to model due to the uneven effort put forward by observers when recording checklists. For example, a checklist could be recorded over ten minutes or over three hours. Both of these checklists are given equal status in the eBird database. If not accounted for, this bias can interfere with the modeling process. Thankfully, eBird checklists come with relevant metadata that describe the effort put forth by users while recording each checklist. We can use this effort information while modeling to account for bias. Johnston et al. have found that adding effort covariates improves the performance of models built with eBird data [24]. Furthermore, Adde et al. used effort covariates when modeling with eBird data. They found that explicit modeling of the observational processes of eBird users by way of effort covariates was needed for the optimal modeling use of eBird data [5].

We will use four effort covariates in our modeling. They are as follows: the time a checklist was started, the time duration of a checklist in minutes, the distance traveled in kilometers, and the number of observers.

2.1.4 Environmental Covariates

We used [10] to obtain boundaries for BCR 31. These boundaries were then used to acquire the relevant land cover data from [19] and elevation data from [8].

In generating our environmental covariates, we followed the procedure laid out in [25]. First, we acquired the MODIS tiles corresponding to BCR 32 for 2016 and 2017. A MODIS tile for a given year is a 1200 km square tile composed of a grid of 500 m square cells, where each cell corresponds to a specific land cover type. MODIS data is recalculated each year using updated data.

We then found the full set of unique checklist locations for 2016 and the set for 2017. Next, we used the *st_buffer()* function from the *sf* package to establish a circular neighborhood with a radius of 2.5 km around each 2016 location and 2017 location [29]. The authors of [25] determined that a radius of 2.5 km is sufficient to account for the lack of precision in the spatial points that represent checklists and also ecologically relevant for many bird species. We were then able to calculate land cover covariates for each 2016 location and 2017 location. For each location's circular neighborhood, we calculated the proportion of land that corresponds to each of the sixteen land cover types. These sixteen land cover covariates necessarily sum to 1 for each location.

We then used an analogous technique with elevation data from the EarthEnv project. The elevation data is the median elevation of a certain area, calculated at a resolution of 1 km, where elevation is represented by meters above sea level. We generated covariates for elevation mean and the standard deviation of elevation for each unique location in the data, again using circular neighborhoods with a radius of 2.5 km. (Unlike the land cover covariates, the elevation covariates are not affected by year.)

The final data set was created by merging the filtered checklist data and the environmental covariates across all of the 10 species of interest, while disposing of X counts. The data from 2016 became training data for that species, while the 2017 data became test data. It is ideal to validate models built from citizen science data with data collected by standardized protocol. However, we did not have access to such high-quality validation data. We acknowledge this as a limitation of our research.

2.2 Generalized Linear Models

GLMs are extensions of the linear regression method that allow us to model an arbitrary distribution of the response variable. They are formed of three components: the random component, the systematic component, and the link function. The random component provides the conditional probability distribution of the response variable Y given a predictor X or vector of predictors $\vec{X}$. We assume that Y, conditional on X, belongs to a certain family of distributions. The systematic component is a linear combination of explanatory variables. The link function, known as $g(.)$, provides the link between the random and the systematic component of a given GLM. It applies a transformation to the expected value of Y, denoted $E[Y]$, such that the transformed mean is equal to the systematic component, the linear combination of explanatory variables [23].

2.2.1 Poisson GLM

A common model used for a response variable Y that takes on non-negative integer values is the Poisson GLM. The full Poisson GLM regression equation for a response variable Y and predictors x_1 through x_k, with λ_i representing the expected Y-count for the i^{th} observation given the corresponding x_i vector, and with $i = 1, 2, \ldots, n$, is:

$$\begin{cases} Y_i \sim Pois(\lambda_i), i = 1, 2, \ldots, n \\ \log(\lambda_i) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} \end{cases} \tag{1}$$

In the above formulation, $Y_i \sim \ldots$ means that the random variable Y_i follows the provided distribution. The Poisson distribution represents the count data for events, such as observing a bird of a certain species, that are assumed to happen at a fixed assumed rate λ_i. More specifically, it implies the following:

$$\mathrm{P}(Y_i = k) = \frac{e^{-\lambda_i}\lambda_i^k}{k!} \text{ for } k = 0, 1, 2 \ldots \tag{2}$$

The right-hand side of the second line in Equation 1, which contains the linear combination of predictors, constitutes the systematic component of Poisson GLM. The left-hand side, on the other hand, indicates $g(\lambda_i) = \log(\lambda_i)$ is the link function of Poisson GLM. The vector of parameters $\vec{\beta}$ is estimated using a maximum likelihood approach [23].

2.2.2 Overdispersion and Underdispersion

The Poisson distribution with random variable Y_i and the Poisson parameter λ_i requires that $E[Y_i] = V[Y_i] = \lambda_i$, where $E[Y_i]$ is the expected value of Y_i and $V[Y_i]$ is the variance of Y_i. If the variance of Y_i is notably greater than λ_i, then the data does not satisfy this assumption of the Poisson distribution; this is known as overdispersion. On the other hand, if the variance of Y_i is notably less than λ_i, then the data is underdispersed. In practice, we can diagnose a fitted model for potential overdispersion or underdispersion by checking if the residual deviance is higher or lower, respectively, than the residual degrees of freedom. If the Poisson model is a reasonable fit for the data, we expect the residual deviance to be approximately equal to the residual degrees of freedom.

When the $E[Y_i] = V[Y_i] = \lambda_i$ assumption is not satisfied, the legitimacy of inference procedures made using that Poisson model is called into question. Situations of underdispersion and overdispersion can be addressed with different variations on the original Poisson distribution. Here, we will discuss the quasi-Poisson distribution.

2.2.3 Quasi-Poisson GLM

In the quasi-Poisson model, instead of assuming $E[Y_i] = V[Y_i] = \lambda_i$, the dispersion parameter ϕ is introduced such that $E[Y_i] = \lambda_i$ and $V[Y_i] = \phi\lambda_i$. The dispersion parameter ϕ allows the modeling technique to accommodate either overdispersed ($\phi > 1$) or underdispersed ($\phi < 1$) data [17]. For the formula used for ϕ, see [17].

Because we set $V[Y_i] = \phi\lambda_i$ in quasi-Poisson regression, we cannot use the classic maximum likelihood technique to estimate parameters. In short, this is because we do not have a well-defined probability distribution available for it (see [17] for more details). Instead, we are required to use the quasi-likelihood method. Classic and quasi-likelihood methods calculate coefficient estimates using analogous techniques, which therefore results in near-identical coefficient estimates. The most important difference is present in the calculation of the standard errors for coefficient estimates. In particular, the $SE(\hat{\beta}_j)$ in quasi-likelihood is equal to $\sqrt{\hat{\phi}} * SE(\hat{\beta}_j)_{classic}$, where $SE(\hat{\beta}_j)_{classic}$ is the standard error for $\hat{\beta}_j$ in the classic maximum likelihood version of the model.

This allows a quasi-Poisson model to account for more uncertainty in coefficient estimates; it therefore generates more accurate values for inference procedures.

The quasi-Poisson random component can be written in shorthand as $Y_i \sim_{\text{ind.}} QPois(\lambda_i, \phi)$. Quasi-Poisson regression has the same systematic and link component as Poisson regression.

2.2.4 Other GLMs

We also used the negative binomial distribution and the zero-inflated Poisson distribution (classic and hurdle formulations) in our modeling of relative abundance data. However, none of these model types proved to be superior to the quasi-Poisson distribution. Therefore, we will not discuss them here. For details on how these distributions operate, refer to [17] and [36].

2.3 Nonlinear Modeling Techniques

2.3.1 Smoothing Splines

We can generate splines, piecewise polynomials that are both continuous and smooth at their knots, via the smoothing spline technique. Smoothing splines result from minimizing an residual sum of squares criterion that is subject to a smoothness penalty. This minimization process generates a smooth curve that is designed to fit the observed data well without overfitting. In particular, to generate a smoothing spline g, we find the function g that minimizes the following, where λ is a nonnegative parameter that controls the strength of the smoothness penalty, by [23]:

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g"(t)^2 dt \tag{3}$$

The term $\sum_{i=1}^{n}(y_i - g(x_i))^2$ is a loss function that encourages g to fit the data. $\lambda \int g"(t)^2 dt$ is a penalty term that encourages g to be smooth, where $g"(t)$ indicates the second derivative of the function g. If $\lambda = 0$, then the function g will exactly interpolate the provided training data. On the other hand, as λ approaches infinity, g will be the linear regression of Y on X [23]. The function $g(x)$ generated by the minimization process will be a natural cubic spline with a knot at every unique value of X.

2.3.2 Generalized Additive Models

GAMs are a framework for extending nonlinear modeling approaches such that we can incorporate multiple predictors. A GAM is written as follows:

$$y_i = \beta_0 + g_1(x_{i,1}) + \cdots + g_k(x_{i,k}) + \epsilon_i \tag{4}$$

where each g_j for $j = 1, 2, \ldots, k$ is a smooth nonlinear function [23]. In particular, smoothing splines can be used to fit a function g_j.

GAMs have several advantages. Because they fit a nonlinear g_j to each X_j, they will automatically model nonlinear relationships that would not be included in standard multiple linear regression [23]. Additionally, these nonlinear fits can potentially generate predictions that are more accurate than those made by an analogous multiple linear regression model. Finally, the additive nature of the model allows us to examine the partial effect of each covariate on the response [23].

All but one of our nonlinearly modeled covariates were modeled with smoothing splines where the effective degrees of freedom K was set at 5. We modeled the covariate for the time a checklist was started using a cyclic cubic regression spline with $K = 7$ knots.

While the description of GAMs above directly extends from multiple linear regression, we can apply GAM techniques to any GLM. In particular, it could be incorporated into the systematic component of a Poisson GLM from Section 2.2.1 or a HGLM, to be discussed in Section 2.4.2.

2.4 Spatial Dependence Modeling

2.4.1 Spatially Dependent Data

All distributions we use to model our data assume that the observations in the response are independent. However, we are working with data that was obtained at certain geographic locations and may therefore be spatially correlated. In general, observations in space can be considered not mutually independent [11]. If our data is spatially correlated, the modeling assumption of independence is violated; this presents a series issue that affects the quality of our fits and how much trust we should place in inferences based on our models [28].

To represent potential spatial autocorrelation in our data, we will generate neighbor objects. Each of our locations is represented by a set of longitude and latitude coordinates; note that there are often multiple checklists that correspond to a single location. For every location, we will categorize all other locations as either neighbors or non-neighbors of that location [11]. To build these neighborhoods, we will use the Sphere of Influence method. A Sphere of Influence neighbor object is created by taking a Delaunay triangulation neighbor object and removing links between two locations that are relatively long [11]. Two locations are Sphere of Influence neighbors if circles centered on each of the two locations' representative coordinates, with radii equal to each location's nearest neighbor distance, intersect in two places. The Sphere of Influence method generates symmetric neighbor objects; that is, if location i is a neighbor of location j, then j is a neighbor of i.

Once a neighbor object has been created, we can assign spatial weights to each relationship. In cases where we do not have much knowledge of the spatial process underlying our data, it is best to use the binary weight style, where all neighbor relationships have a weight of one and all non-neighbor relationships have a weight of zero [11]. This is what we chose for our analysis.

2.4.2 Hierarchical Generalized Linear Models

To model spatial dependence, we will use HGLMs. An HGLM is a GLM that allows for random effects; we can specify a certain distribution or model matrix for the random effects [34]. By specifying a particular type of distribution for the random effects, we can impose the desired spatial dependence structure on our model. In our research, we will be generating quasi-Poisson HGLMs with spatially correlated random effects. By this point in our research, quasi-Poisson will have been shown to be the preferred distribution out of those introduced in Section 2.2.

The systematic component of a quasi-Poisson HGLM is $\vec{\beta} * \mathbf{X}_i + v_i$, where v_i is a random effect following a certain specified distribution. For our research, in accordance with standard practice for spatially correlated random effects [33], we assumed that $v_j \sim N(0, \mathbf{\Sigma})$, where j indexes the geographic location of an observation. We then specify $\mathbf{\Sigma}$ with a certain structure. For our research, we will use the CAR structure. Altogether, the model formula for a quasi-Poisson HGLM with spatially correlated random effects under the CAR structure is as follows:

$$\begin{cases} Y_{i,j} \sim QuasiPois(\lambda_{i,j}, \phi) \\ \log(\lambda_{i,j}) = \vec{X_{i,j}}\vec{\beta}^T + v_j \\ v_j \sim N(0, \mathbf{\Sigma}) \\ \mathbf{\Sigma} = \rho(\mathbf{I} - \tau * \mathbf{D})^{-1} \\ \phi = \frac{1}{n-(k+1)} \sum_{i=1}^{n} \frac{(y_{i,j}-\lambda_{i,j})^2}{\lambda_{i,j}} \end{cases} \tag{5}$$

where k is the number of predictors, n is the number of observations, i is the observation index that goes from 1 to n, and j is the index for the geographic location of an observation. $\mathbf{I}$ is an identity matrix and $\mathbf{D}$ is a symmetric matrix of spatial weights [28]. ρ is a spatial correlation parameter; it represents the level of spatial dependence present in the response. τ is a spatial variance component. Parameters ρ and τ are estimated during the process of fitting an HGLM [33]. Estimation of an HGLM model is performed using h-likelihood theory. For more details, see [34].

When we incorporate GAM techniques into the model formula for an HGLM, we can consider that model to be a hierarchical GAM (HGAM). Techniques that apply to HGLMs can be extended to apply to HGAMs.

2.4.3 Evaluating Spatial Correlation

After having fit an HGLM model with the CAR structure for random effects, we can use the spatial correlation estimate $\hat{\rho}$ to consider the level of spatial dependence in the data. As it is a measure of correlation, it takes on values from -1 to 1, with larger absolute magnitudes indicating stronger levels of spatial autocorrelation.

2.4.4 Making Predictions with HGLMs

While the systematic component of an HGLM includes a random effect, we can still provide predictions for new data as follows. Here, we will discuss using a quasi-Poisson HGLM fit to make predictions. Since v_j is distributed according to the Normal distribution with mean 0, and keeping in mind that predictions are made on average, the contribution of $\vec{v}$ to the prediction $\vec{\hat{y}}_{new}$ will be 0. We are then able to predict $\vec{\hat{y}}_{new}$ using exclusively our matrix of new data $\mathbf{X}_{new}$ and our vector of coefficient estimates $\vec{\hat{\beta}}$. as below, where the exponent is present to reverse the log-link transformation of the quasi-Poisson HGLM:

$$\vec{\hat{y}_{new}} = e^{(\mathbf{X}_{new} * \vec{\hat{\beta}})} \tag{6}$$

2.5 Evaluating Predictive Performance

Mean absolute deviation (MAD) is a commonly used metric that represent the average distance of predicted values from their corresponding true values. It is preferable to the similar metric of root mean squared error because it is robust against extreme observations in the test data. Lower values of MAD indicate that a model has stronger predictive performance with regards to predicting exact values [21]. The formula for MAD is below, where $\vec{y} = y_1, y_2, \ldots y_n$ and $\vec{\hat{y}} = \hat{y}_1, \hat{y}_2, \ldots \hat{y}_n$:

$$\text{MAD} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{7}$$

3 Results

All analyses in this research were performed using the R software environment, version 4.1.1 [30]. The key R packages we used to complete our analysis are hglm [6] [32], sf [29], car [18], MASS [40], pscl [22] [45], spdep [38], and auk [38]. We also used the package mgcv [41] [42] [43] [44].

3.1 Preliminary Analysis

We began with an initial suite of effort and environmental covariates inspired by the methods proposed in [24]. We used the white ibis data as a reference case for instances where we needed to make a decision that would be applied to all species. First, we removed any covariates that had at least 95 percent zero values and a small spread of nonzero values, as they would have been not useful to our modeling. Next, we performed multicollinearity analysis. To resolve problematic multicollinearity in our data, we dropped one covariate. Our final suite of selected covariates is presented in Table 1. Following the above, we performed influential data analysis on each species. The minimum number of observations dropped was 0, and the maximum number was 12.

We then conducted exploratory data analysis on the count data for each of our ten species of interest. In all cases, the data was powerfully right-skewed with an excess number of zeroes. Refer to Figure 1 for the example of the white ibis species.

3.2 Selecting Initial Model Type

First, we fit a Poisson GLM on the filtered training data for all ten species using the covariates in Table 1. For all species except the roseate spoonbill, the residual deviance was greater than the residual degrees of freedom, indicating potential overdispersion. Given the overall evidence of overdispersion in our data, we decided to discard the Poisson GLM fits and move on to more suitable modeling techniques.

We next fit a quasi-Poisson GLM, a negative binomial GLM, and a zero-inflated Poisson GLM on each of the ten species. An investigation of the quasi-Poisson GLM fits revealed that all species had $\hat{\phi} > 2$. This clear evidence of overdispersion in the data provides support for our choice to model using the quasi-Poisson, negative binomial, and zero-inflated Poisson distributions.

After fitting our GLMs, we prepared to fit GAMs. We first investigated each covariate to see if it had enough unique values to be sensibly modeled with nonlinear techniques, using a threshold of 50 unique values. The covariates for number of observers, deciduous broadleaf, cropland, and mosaic all failed to meet the threshold; we therefore continued to model these covariates linearly. All other covariates were modeled as discussed in Section 2.3.2. We fit a quasi-Poisson GAM, negative binomial GAM, and zero-inflated GAM to each of the ten species as above.

We then evaluated each of the six model fits for the ten species in order to select the best model type for our data. First, we generated predictions for each of our six models types of interest and each of our ten species using the test data. Next, we calculated MAD values using those predictions for all sixty models.

Final Suite of Covariates		
Name	Type	Description
Time Checklist Started	Effort	Time of day checklist was started, with range of values from 0 to 24
Duration	Effort	Duration of checklist in minutes
Distance Traveled	Effort	Distance traveled while completing checklist in km
Number of Observers	Effort	Number of people observing birds while recording the checklist
Evergreen Broadleaf	Land cover	Percentage of evergreen broadleaf terrain *
Deciduous Broadleaf	Land cover	Percentage of deciduous broadleaf terrain *
Woody Savanna	Land cover	Percentage of woody savanna terrain *
Grassland	Land cover	Percentage of grassland terrain *
Wetland	Land cover	Percentage of wetland terrain *
Cropland	Land cover	Percentage of cropland terrain *
Urban	Land cover	Percentage of urban terrain *
Mosaic	Land cover	Percentage of mosaic terrain *
Mean Elevation	Elevation	Mean of elevation *
Standard Deviation of Elevation	Elevation	Standard deviation of elevation *

Table 1: The final set of selected covariates, with type and description. * indicates that the covariate was measured in the area around the checklist location.

Figure 1: Left: Histogram of white ibis counts. Right: Bar plot of white ibis counts less than or equal to 25.

MAD Values and Percent Change			
	Quasi-Poisson GLM	Quasi-Poisson GAM	Percent Change
White Ibis	4.145	3.904	−5.829
Glossy Ibis	1.018	0.996	−2.176
Roseate Spoonbill	0.325	0.334	+2.664
Great Egret	1.405	1.348	−4.096
Cattle Egret	2.570	2.556	−0.556
Snowy Egret	1.480	1.439	−2.791
Great Blue Heron	0.632	0.575	−8.936
Little Blue Heron	0.816	0.786	−3.651
Tricolored Heron	0.985	0.870	−11.623
Green Heron	0.485	0.431	−11.049

Table 2: MAD values by model type; Percent change when switching from GLM to GAM.

For the three GLM fits for each species, we observed that the negative binomial GLM fits generally had worse predictive performance than the other two GLM fits. Quasi-Poisson GLM and zero-inflated Poisson GLM fits had fairly similar performances across the ten species. Since the quasi-Poisson model type is less complex than the zero-inflated Poisson model type, we considered quasi-Poisson to be superior. Additionally, for the three GAM fits for each species, the quasi-Poisson GAM fit has the best or near-best performance. Therefore, we decided to select quasi-Poisson as our preferred distribution.

Our next task was to evaluate whether or not GAM techniques should be implemented, given their additional interpretation complexity and risks of overfitting. The first two columns of Table 2 show the MAD values for the quasi-Poisson GLM and quasi-Poisson GAM fits for each species. The final column of Table 2 shows the percent change in MAD with the addition of GAM techniques. Since a smaller MAD value indicates better predictive performance, nine of the ten species showed an improvement in model performance when GAM techniques were added. The average percent change was −4.804 percent. Overall, adding GAM techniques to a quasi-Poisson GLM improved predictive performance quality across species. We selected the quasi-Poisson GAM as our preferred modeling type.

3.3 Incorporating Spatial Dependence

Following from the above results, we decided to use a quasi-Poisson HGAM with a CAR structure from random effects to model relative abundance. First, we created a Sphere of Influence neighbor object and a corresponding binary weights object for each species. We then fit a quasi-Poisson HGAM as we fit our quasi-Poisson GAM in Section 3.2 but with the addition of a CAR structure for random effects using our binary weights object. Unfortunately, the model failed to converge for the roseate spoonbill and tricolored heron species, which is not atypical for complex models such as HGLMs [20]. For the rest of this paper, we will be exclusively working with the other eight species.

We then used the CAR $\hat{\rho}$ statistic for each species' HGAM fit to investigate the practical level of spatial correlation in the data. These $\hat{\rho}$ statistics are provided in Table 3. The minimum $\hat{\rho}$ is 0.174 and the maximum is 0.214. Our CAR $\hat{\rho}$ statistics indicate that we do have spatial correlation present in our data. While it is not extremely large, it is still notable.

CAR $\hat{\rho}$ Values	
Species	CAR $\hat{\rho}$
White Ibis	0.199
Glossy Ibis	0.204
Great Egret	0.198
Cattle Egret	0.174
Snowy Egret	0.192
Great Blue Heron	0.214
Little Blue Heron	0.184
Green Heron	0.196

Table 3: The CAR $\hat{\rho}$ values for each quasi-Poisson HGAM fit.

MAD Values and Percent Change			
	Quasi-Poisson GAM	Quasi-Poisson HGAM	Percent Change
White Ibis	3.904	2.790	−28.536
Glossy Ibis	0.996	0.518	−47.987
Great Egret	1.348	1.061	−21.308
Cattle Egret	2.556	1.632	−36.144
Snowy Egret	1.439	0.958	−33.391
Great Blue Heron	0.575	0.424	−26.191
Little Blue Heron	0.786	0.557	−29.141
Green Heron	0.431	0.307	−28.889

Table 4: MAD values by model type; Percent change when switching from GAM to HGAM.

3.4 Selecting Final Model Type

We generated predictions for the 2017 test data using each of our HGAM fits; we then used these predictions to calculate MAD values. We can now compare these MAD values to the preexisting MAD values for the relevant eight species' GAM fits; see Table 4. There was quite notable improvement in MAD when switching from the GAM to HGAM modeling type across all eight species, with an average percent decrease (improvement) of 31.448 percent. The percent decrease goes up to 47.987 percent and never drops below 21.308 percent. This clearly demonstrates the advantages of incorporating spatial autocorrelation into our modeling procedure.

We can further investigate this improvement in predictive performance by splitting the test set into nondetections (zero counts) and detections (nonzero counts). When considering just nondetections, MAD decreases by 65.737 percent on average, an almost two-thirds improvement. This shows that the HGAM model type is much better at predicting nondetections than the GAM model type. However, when considering just detections, MAD increases (worsens) by 11.087 percent on average. In a reversal of the previous result, the GAM models are slightly better at predicting for detections than the HGAM models. In our opinion, this presents an acceptable tradeoff in predictive performance; the HGAM fits, while being slightly worse at predicting exact counts for detections, are much better than the GAM fits at identifying true non-detections.

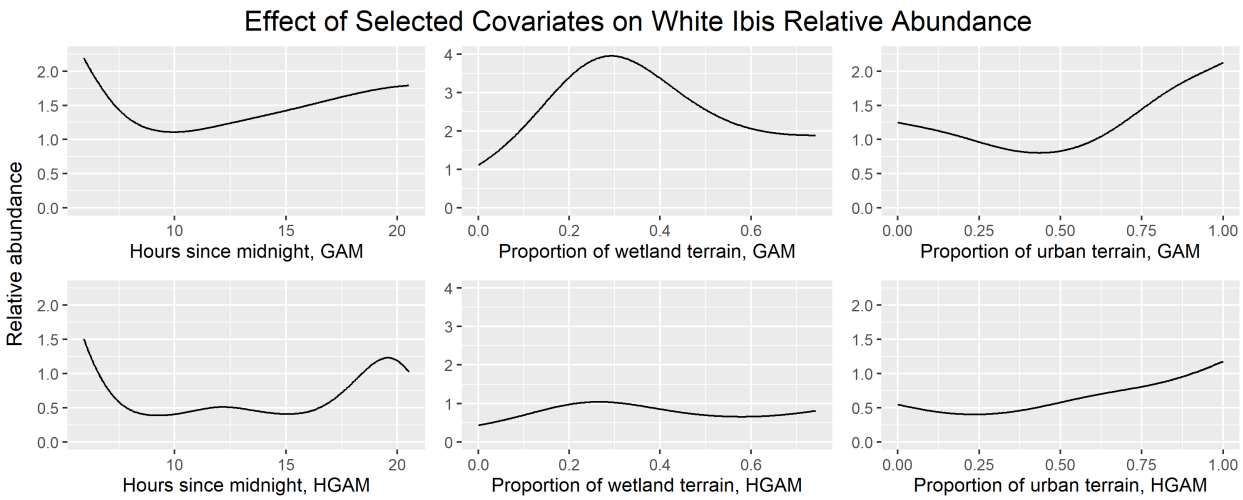

Figure 2: Effect displays of the selected covariates for the white ibis data. Above: quasi-Poisson GAM. Below: quasi-Poisson HGAM.

3.5 Results of the Final Model Type

Figure 2 features effect displays for three selected covariates in the context of the quasi-Poisson GAM and HGAM for the white ibis species. The effect displays were constructed in the following manner. First, we plotted histograms of each covariate to look for any skew in their distribution. If a covariate appeared to have extreme low (high) values, we set the minimum (maximum) value for its effect display to the 0.01 (0.99) quantile of that covariate. Otherwise, we did not adjust the minimum or maximum. The minimum and maximum were used to create a sequence of 300 evenly spaced values for the covariate. Next, we obtained the median values of all other covariates in the model. Following that, we used the sequence of 300 covariate values and the median values of the other covariates to make predictions for our selected model fit. Finally, we used those predictions to plot an effect display for that covariate. These plots represent the effect of each covariate on the response for the corresponding model fit.

The effect displays for the covariate representing the time a checklist was started are in the left column of Figure 2. For the GAM fit, relative abundance begins high, drops until approximately 10 a.m., then increases steadily. This suggests that the white ibis is most likely to be observed during early morning and evening hours and least likely to be observed in the late morning. For the HGAM fit, relative abundance drops rapidly, is stable from around 10 a.m. to 3 p.m., then increases again. The HGAM fit provides evidence that the white ibis is more likely to be observed during early morning and evening hours than in the middle of the day. The white ibis is known to roost in groups at night, then go foraging throughout the day [35]. Both of these effect displays are plausible in the context of the white ibis species.

The effect displays for the wetland covariate are in the middle column of Figure 2. For the GAM fit, relative abundance starts low, rises to a peak at approximately 30 percent wetland terrain, then falls before leveling off at around 60 percent. The effect display for the HGAM fit is more smooth and less extreme than that of the GAM fit; relative abundance has a slight peak around 30 percent and a slight dip around 60 percent. The white ibis is known to be associated with wetland habitats [27]. Therefore, we would expect that as the percentage of wetland terrain increases, relative abundance also increases. We find the effect display for the GAM fit to be not

particularly sensible, as it displays an exaggerated up-and-down relationship instead of a more steadily positive one. The effect display for the HGAM fit is comparatively less problematic; its depiction of the effect of wetland on relative abundance is less extreme, and therefore features less of a decrease in relative abundance for high proportions of wetland terrain.

The effect displays for the urban covariate are in the right column of Figure 2. In the GAM fit, relative abundance has a dip at around 50 percent urban terrain, with a steeper rise on the right side of the dip than on the left side. Meanwhile, the dip in the HGAM effect display is less pronounced and located at approximately 25 percent instead of 50 percent. Additionally, the values of relative abundance are less extreme overall in the HGAM effect display. The white ibis is known to have increased its presence in urban areas in response to Florida's urbanization [27]. It is therefore reasonable to think that high percentages of urban terrain correspond with the highest values of relative abundance. This is reflected in both effect displays. With that being said, we found the HGAM effect display to be a better representation of the relationship between the proportion of urban terrain and relative abundance as it depicts a more steady increase. The pronounced peaks and dip of the GAM effect display imply that areas with low urbanization are more conducive to the white ibis than areas that are moderately urbanized, which we lack a clear explanation for in the context of the white ibis.

For all three of the above covariates, in the GAM fit the effect of that covariate on the response was determined to be statistically significant, which indicates that each respective relationship was not observed simply due to chance. In the HGAM fit, where we adjusted for spatial dependence, only the covariate for the time a checklist was started was found to be statistically significant; the environmental covariates for wetland and urban were not deemed so. This aligns with the results of other research, which have found that adjusting for spatial dependence has a tendency to impact coefficient estimates and the significances of environmental covariates in particular [15] [31]. Additionally, the pattern of environmental covariates decreasing in statistical significance with the incorporation of spatial autocorrelation, and doing so more than effort covariates, generally persisted for the other species in our study.

Incorporating spatial autocorrelation after having already accounted for the environmental and effort covariates present in our model allows us to control for spatial factors such as climate information, species dispersal patterns, and colonial breeding habits that we cannot easily include as explicit covariates. Therefore, the coefficient estimates and inferences drawn from a model that accounts for spatial dependence (in our case, an HGAM) are generally more reliable than those from the corresponding non-spatial model (a GAM). In particular, our spatially explicit models can protect us from overstating the relationships between certain environmental covariates and relative abundance.

As an alternative method of assessing the GAM and HGAM fits, we made prediction plots for the white ibis species across our study area; see Figure 3. To make these plots, we obtained the true values of environmental covariates across BCR 31 for 2016. We then set standard values for our effort covariates. In particular, we set checklist start time to noon, duration to one hour, distance to one kilometer, and number of observers to one. We then generated predictions using our model fits. This is best practice when making predictions across a geographic region with models that use both effort and environmental covariates, as it implies the assumption of spatially homogeneous observer effort [13].

The prediction plot for the GAM fit has areas of extreme low and extreme high relative abundance scattered across the study area. The prediction plot for the HGAM fit is more stable with

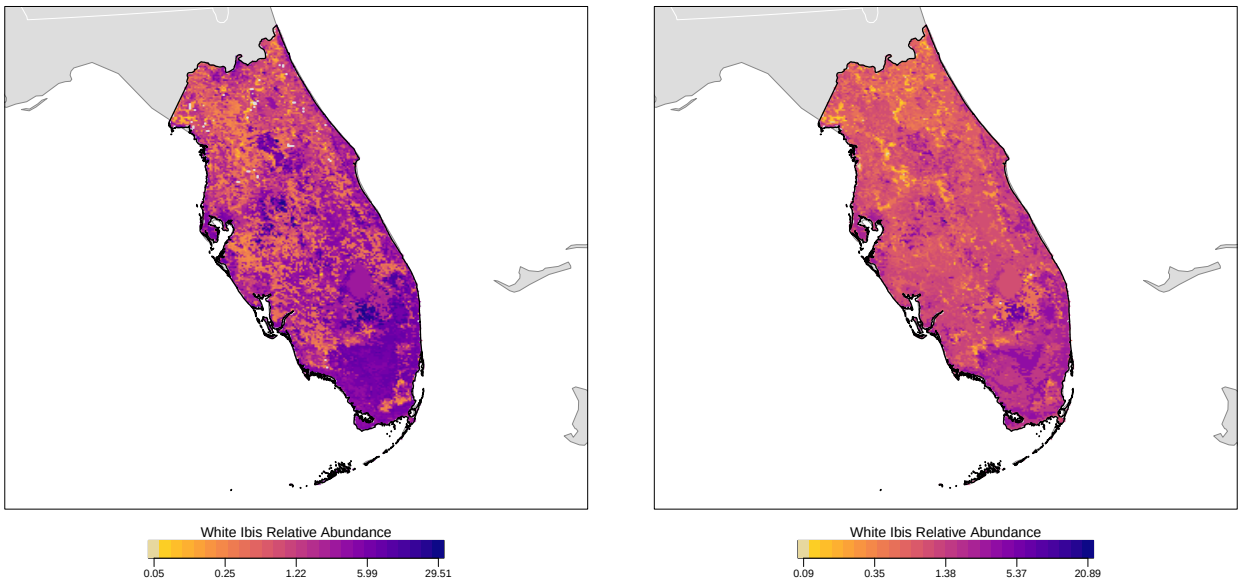

Figure 3: Left: Prediction plot map for quasi-Poisson GAM. Right: Prediction plot map for quasi-Poisson HGAM. The scale for relative abundance changes slightly between maps.

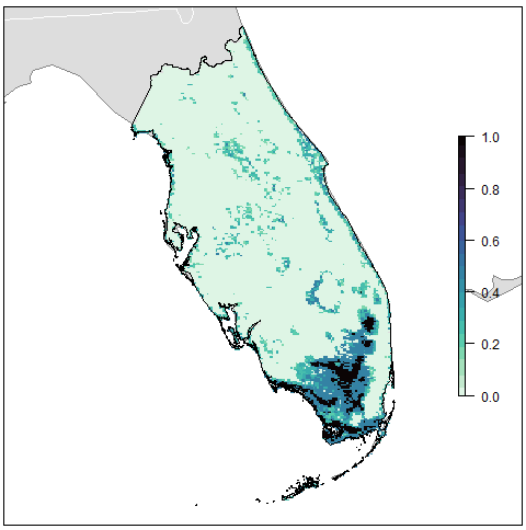

Figure 4: The proportion of wetland land cover across BCR 31 in 2016.

regards to its predictions across the study area. Considering that the white ibis population is known to be spread across peninsular Florida, this appears to be more sensible. For both model fits, the area roughly corresponding to the Everglades has generally higher values of predicted relative abundance. This corresponds with our domain knowledge and is a sign that our model fits have been at least partially successful.

From the prediction plot, we observed that the HGAM fit identified areas in BCR 31 with very high proportions of wetland as corresponding to high values of relative abundance. The dark streaks in the bottom of the HGAM plot in Figure 3 correspond to areas with high proportions of wetland in Figure 4. This indicates that the wetland covariate has practical significance, which is interesting as this covariate was not found to be statistically significant in the HGAM fit. Of further note is the fact that this result contrasts with what we observed in the effect display for wetland in the HGAM fit (see Figure 2), where there was less dependence of relative abundance on the proportion of wetland terrain. This could potentially indicate that the HGAM fit adapts differently to actual combinations of environmental covariates present in BCR 31.

4 Discussion

4.1 Discussion of Results

We found that the quasi-Poisson distribution has superior predictive performance over the negative binomial distribution and the zero-inflated Poisson distribution in the context of both GLM fits and GAM fits. We then observed that the quasi-Poisson GAM had stronger predictive performance than the quasi-Poisson GLM. To make both of these conclusions, we used MAD values generated by predicting for our 2017 test data.

We then incorporated any remaining spatial autocorrelation into our modeling using a quasi-Poisson HGAM with spatially correlated random effects. CAR $\hat{\rho}$ values for our eight HGAM fits indicated approximate spatial autocorrelation of around 0.2 for our data. We therefore observed a moderate level of spatial dependence in our relative abundance data after having already accounted for our environmental and effort covariates. This supports our decision to incorporate spatial dependence into the modeling task. Continuing to use non-spatial models for relative abundance in the context of eBird data will result in biased and untrustworthy model fits. Additionally, our MAD values showed that HGAMs perform better than GAMs when predicting counts for new data.

Using effect displays, we observed that the GAM fit for the white ibis had more extreme and nonlinear relationships for environmental covariates as compared to the corresponding relationships of the HGAM fit. In other words, the HGAM fit had adjusted coefficient estimates such that the environmental variables had less pronounced effects on relative abundance. This result matches that of previous studies [14] [15] [31]. Our prediction plots showed that the HGAM fit was able to predict more smoothly across the study area. Interestingly, we also observed using the prediction plots that the HGAM fit identified the impact of very high wetland proportions on relative abundance in the context of real geographic locations. Together, these results suggest that incorporating spatial dependence into the modeling task provides a more nuanced and reliable understanding of relative abundance for a bird species across a given geographic region.

Our study provides a thorough approach for modeling the relative abundance of bird species by making the best use of information available in eBird citizen science data. With the incorpora-

tion of spatial dependence, we obtained a more accurate understanding of species' reliance on environmental covariates. Our modeling approach can be used to identify suitable areas for bird species which, in turn, could be used as recommendations for conservation efforts. Our modeling technique also has use for climate change planning, as it can be applied to assess species' relative abundance in areas with changing environmental characteristics. Through incorporating spatial dependence, such conclusions will be more realistic and reliable [14]. As bird populations decline, our study presents an opportunity to understand both the areas currently most relevant to their abundance and sites that may gain or lose suitability with potential shifting environmental characteristics.

4.2 Areas of Future Research

There remain areas of future research relating to this study.

One matter is the issue of X counts, which we had to remove in order to model relative abundance. Our modeling would have been stronger if we had some way to fill in predicted values for these X counts. We know that all X counts are nonzero values. This is therefore lost information. Filtering out a specific type of observation from our data harms the modeling process.

Another issue is the non-convergence of the roseate spoonbill and tricolored heron species. For both of these species, the quasi-Poisson HGAM failed to fit. Overdispersed species data with excess zeros are supposed to be suited to a quasi-Poisson HGLM with spatially correlated random effects [28]. Even though convergence issues are not unheard of for hierarchical models, investigation into this topic is required [20].

Furthermore, we are in need of better methods for measuring the statistical significance of spatial dependence in our research context. eBird data is not suitable for traditional tests for spatial dependence such as Moran's I, which require that each location has exactly one observation. Additionally, we were unable to use a likelihood ratio test to measure the statistical significance of spatial random effects in a quasi-Poisson HGAM. This is because likelihood statistics are not available for the quasi-Poisson family in the hglm package. Better procedures should be developed for this context.

Finally, future research could look into methods for creating confidence bands for HGAM fits and their predictions. Our effect displays would have been more informative with confidence bands.

References

[1] Bird Conservation Regions. https://nabci-us.org/resources/bird-conservation-regions/. US NABCI. Accessed: 2022-04-05.

[2] Birding as your 'primary purpose' and complete checklists. https://support.ebird.org/en/support/solutions/articles/48000967748-birding-as-your-primary-purpose-and-complete-checklists. eBird. Accessed: 2022-04-05.

[3] The eBird review process. https://support.ebird.org/en/support/solutions/articles/48000795278-the-ebird-review-process. eBird. Accessed: 2022-04-05.

[4] Guide to eBird protocols. https://support.ebird.org/en/support/solutions/articles/48000950859-guide-to-ebird-protocols. eBird. Accessed: 2022-04-05.

[5] Adde, A., C. Casabona i Amat, M. J. Mazerolle, M. Darveau, S. G. Cumming, and R. B. O'Hara (2021). Integrated modeling of waterfowl distribution in western canada using aerial survey and citizen science (ebird) data. *Ecosphere 12*(10), e03790.

[6] Alam, M., L. Ronnegard, and X. Shen (2015). Fitting conditional and simultaneous autoregressive spatial models in hglm. *The R Journal 7*(2), 5–18.

[7] Alden, P. et al. (1998). *National Audubon Society Field Guide to Florida*. Alfred A. Knopf.

[8] Amatulli, G., S. Domisch, M.-N. Tuanmu, B. Parmentier, A. Ranipeta, J. Malczyk, and W. Jetz (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data 5*. Article number: 180040.

[9] Axelson, G. (2019). Vanishing: More than 1 in 4 birds has disappeared in the last 50 years. https://www.allaboutbirds.org/news/vanishing-1-in-4-birds-gone/?__hstc=161696355.417a30ab61029526cb676ebc0e35dcc9.1649175443327.1651544200843.1654530629024.4&__hssc=161696355.1.1654530629024&__hsfp=1353873515&_gl=1*1bvk7u7*_ga*MjgwMjQ5MjU2LjE2NDkxNzU0NDM.*_ga_QR4NVXZ8BM*MTY1NDUzMDYyNy40LjAuMTY1NDUzMDYyNy42MA..#_ga=2.196850159.235101567.1654530628-280249256.1649175443. All About Birds. Accessed: 2022-06-07.

[10] Bird Studies Canada and NABCI (2014). Bird Conservation Regions. https://www.birdscanada.org/bird-science/nabci-bird-conservation-regions/. Bird Studies Canada. Accessed: 2022-04-17.

[11] Bivand, R. S., E. J. Pebesma, V. Gómez-Rubio, and E. J. Pebesma (2008). *Applied spatial data analysis with R*. Springer.

[12] Boman, E. M., M. De Graaf, A. S. Kough, A. Izioka-Kuramae, A. F. Zuur, A. Smaal, and L. Nagelkerke (2021). Spatial dependency in abundance of queen conch, aliger gigas, in the caribbean, indicates the importance of surveying deep-water distributions. *Diversity and Distributions 27*(11), 2157–2169.

[13] Bonnet-Lebrun, A.-S., A. Karamanlidis, M. de Gabriel Hernando, I. Renner, and O. Gimenez (2020). Identifying priority conservation areas for a recovering brown bear population in greece using citizen science data. *Animal Conservation 23*(1), 83–93.

[14] Crase, B., A. Liedloff, P. A. Vesk, Y. Fukuda, and B. A. Wintle (2014). Incorporating spa-

tial autocorrelation into species distribution models alters forecasts of climate-mediated range shifts. *Global Change Biology 20*(8), 2566–2579.

[15] Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global ecology and biogeography 16*(2), 129–138.

[16] Fitzpatrick, J. W. and P. P. Marra (2019). The crisis for birds is a crisis for us all. `https://www.nytimes.com/2019/09/19/opinion/crisis-birds-north-america.html`. New York Times. Accessed: 2022-06-07.

[17] Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

[18] Fox, J. and S. Weisberg (2019). *An R Companion to Applied Regression* (Third ed.). Thousand Oaks CA: Sage.

[19] Friedl, M. and D. Sulla-Menashe (2015). MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006. `https://doi.org/10.5067/MODIS/MCD12Q1.006`. NASA EOSDIS Land Processes DAAC.

[20] Hertzog, L. R., C. Frank, S. Klimek, N. Röder, H. G. Böhner, and J. Kamp (2021). Model-based integration of citizen science data from disparate sources increases the precision of bird population trends. *Diversity and Distributions 27*(6), 1106–1119.

[21] Hyndman, R. J. and G. Athanasopoulos (2018). *Forecasting: Principles and Practice* (2 ed.). OTexts. Accessed: 2022-04-17.

[22] Jackman, S. (2020). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory*. Sydney, New South Wales, Australia: United States Studies Centre, University of Sydney. R package version 1.5.5.

[23] James, G., D. Witten, T. Hastie, and R. Tibshirani (2021). *An Introduction to Statistical Learning with Applications in R* (2 ed.). Springer.

[24] Johnston, A., W. Hochachka, M. Strimas-Mackey, V. Ruiz-Gutierrez, O. Robinson, E. Miller, T. Auer, S. Kelling, and D. Fink (2020a). Analytical guidelines to increase the value of citizen science data: using eBird data to estimate species occurrence. *Diversity and Distributions*, 1265–1277.

[25] Johnston, A., W. Hochachka, M. Strimas-Mackey, V. Ruiz-Gutierrez, O. Robinson, E. Miller, T. Auer, S. Kelling, and D. Fink (2020b). *Best Practices for Using eBird Data* (1 ed.). Cornell Lab of Ornithology.

[26] Kent, A. M., C. Faulhaber, and C. Watson (2017). Peninsular Florida Bird Conservation Region (BCR 31) plan. `https://www.acjv.org/documents/BCR_31_final.pdf`. Florida Fish and Wildlife Conservation Commission and Atlantic Coast Joint Venture. Accessed: 2022-04-17.

[27] Kidd-Weaver, A., J. Hepinstall-Cymerman, C. N. Welch, M. H. Murray, H. C. Adams, T. J. Ellison, M. J. Yabsley, and S. M. Hernandez (2020). The movements of a recently urbanized wading bird reveal changes in season timing and length related to resource use. *PloS one 15*(3), e0230158.

[28] Lee, Y., M. M. Alam, M. Noh, L. Rönnegård, and A. Skarin (2016). Spatial modeling of data with excessive zeros applied to reindeer pellet-group counts. *Ecology and evolution 6*(19),

7047–7056.

[29] Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal 10*(1), 439–446.

[30] R Core Team (2021). R: A language and environment for statistical computing. `https://www.R-project.org/`. R Foundation for Statistical Computing.

[31] Record, S., M. C. Fitzpatrick, A. O. Finley, S. Veloz, and A. M. Ellison (2013). Should species distribution models account for spatial autocorrelation? a test of model projections across eight millennia of climate change. *Global Ecology and Biogeography 22*(6), 760–771.

[32] Ronnegard, L., X. Shen, and M. Alam (2010). hglm: A package for fitting hierarchical generalized linear models. *The R Journal 2*(2), 20–28.

[33] Rönnegård, L., M. Alam, and X. Shen. Fitting spatial models in the R package: hglm. `https://www.diva-portal.org/smash/get/diva2:685966/FULLTEXT02`. Dalarna University. Accessed: 2022-04-17.

[34] Rönnegård, L., M. Alam, and X. Shen. The hglm package (version 2.0). Accessed: 2022-04-17.

[35] Sibley, D. A. (2017). *The Sibley Field Guide to Birds of Eastern North America* (2 ed.). Alfred A. Knopf.

[36] Stan Development Team. *Stan User's Guide*. Version 2.29.

[37] Steen, V. A., C. S. Elphick, and M. W. Tingley (2019). An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity and Distributions 25*(12), 1857–1869.

[38] Strimas-Mackey, M., E. Miller, and W. Hochachka (2018). *auk: eBird Data Extraction and Processing with AWK*. R package version 0.3.0.

[39] Sulla-Menashe, D. and M. A. Friedl (2018). User guide to Collection 6 MODIS Land Cover (MCD12Q1 and MCD12C1) Product. `https://lpdaac.usgs.gov/documents/101/MCD12_User_Guide_V6.pdf`. Accessed: 2022-04-17.

[40] Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4 ed.). New York: Springer. ISBN 0-387-95457-0.

[41] Wood, S., N., Pya, and B. S"afken (2016). Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association 111*, 1548–1575.

[42] Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B) 65*(1), 95–114.

[43] Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association 99*(467), 673–686.

[44] Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B) 73*(1), 3–36.

[45] Zeileis, A., C. Kleiber, and S. Jackman (2008). Regression models for count data in R. *Journal of Statistical Software 27*(8).