Spatio - Temporal Analysis of Bird Migration

Simranjeet Bilkhu (400611361)

Abstract

Environmentalists and scientists are agreed that the rapid growth of human activities has lead and will continue to lead to environmental deterioration [30]. Consequently, bird migration dates have shifted [11] making it ever more challenging to create forecasts on when and where birds are migrating without incorporating environmental factors. Models that exist to incorporate ecological factors such as *Bird Cast* [34] only serve to provide short-term forecasts that are updated every 6 hours. To address the issue of creating long term forecasts, we propose two methods that aim to provide long term spatio-temporal forecasts of bird counts in North America. We conclude that while including climate factors as well as spatio-temporal correlations were important in predicting bird counts, limitations in the data make it hard to create reliable long term forecasts using these methods.

Introduction

The Importance of studying migratory patterns

The importance of studying migratory patterns in birds should not be overlooked. Migratory birds may connect ecosystems across continents, and their seasonal movements can highlight the health of different habitats [39]. Moreover, changes in migration timing and routes may serve as early indicators for environmental shifts [3]. Work has also been done in predicting West Nile virus transmission in North America using mixed models and the *eBird data set* [19].

Previous Work

Previous work has been done in involving climate patterns to predict bird migration, namely *Bird Cast* [34]. *Bird Cast* is the result of interdisciplinary efforts by the *Cornell Lab*, *Colorado State University*, and the *University of Massachusetts* to develop detection and prediction tools for bird migration using radar detection and climate data. However, the predictions available from *Bird Cast* are only short term, nocturnal migration forcasts that are updated every 6 hours.

Problem Statement

In order to create long term forecasts that predict migratory patterns months or years into the future, we propose two methods that aim to provide long term spatio-temporal forecasts of bird counts in North America. The first approach uses a classical penalized poisson regression model, while the second uses poisson regression with additive *Gaussian Random Fields* to account for the spatio-temporal structure involved in the data set. We will start the discussion by discussing the source of our data, followed by an exploratory analysis of each data set. We will then present our methods, followed by our results and then a discussion of our limitations and conclusions.

Dataset Sourcing

The venerable *eBird* data set is available on the *Global Biodiversity Information Facility* (GBIF) [7]. The *eBird* data set is part of a citizen science project which allows users from all over the world to submit their own bird sightings, in combination the bird sightings that have been submitted based on historical data. While the data set houses observations form all over the world and involve sightings as recent as 2024, we restrict our analysis to monthly bird sightings, from early 2001 to the end of 2005 for computational reasons. To further ease the computational burden, we restrict our attention to the problem of predicting and analyzing only 2 species of birds; *Falco peregrinus* and *Geococcyx californianus.* We then aggregated bird counts in 60×60 km grids.

To include spatially and temporally correlated climate data, we looked to the NASA Earth Data Search data library. The Earth Data Search data library houses earth science data collected by NASA. We specifically used three separate data sets from this library. The first being satellite data from the *Terra Moderate Resolution Imaging Spectroradiometer*, which provides land surface temperature data for both day and night [35]. The second being another satellite gathered data set from the same tool which provides snow coverage percentages [12]. Lastly, we used precipitation data gathered by the *Precipitation Processing System* [17], which contains the average liquid precipitation (mm) of given a month and location.

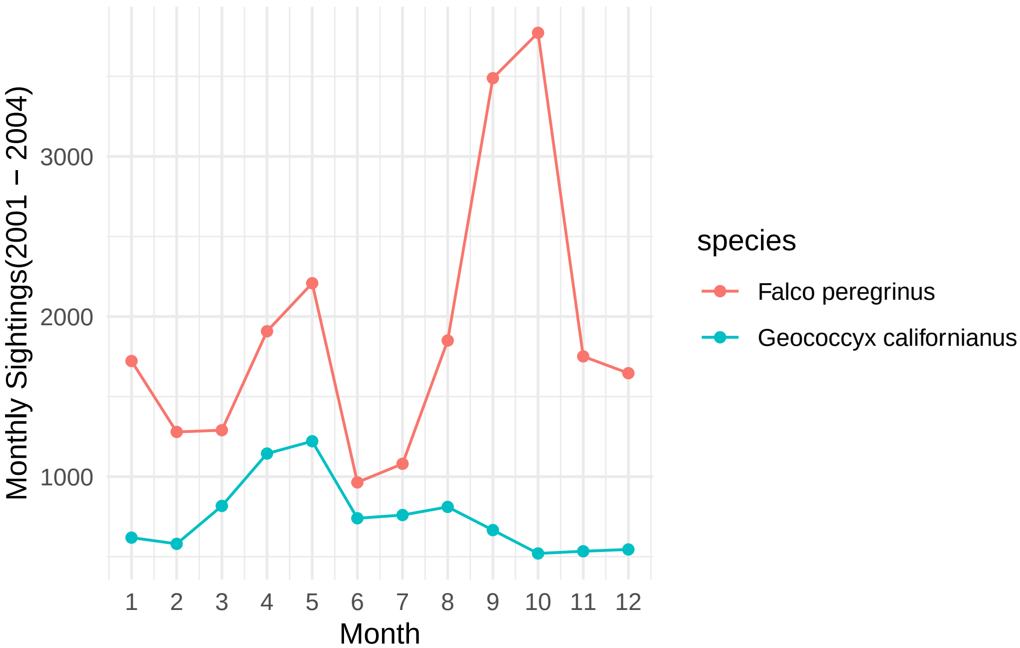
After processing the data, we end up with a data set of 13642 observations with 9 variables. With these data sets, we aim to predict, for a given month, year, longitude and latitude the number of birds by species using day time temperatures, night time temperatures, snow coverage and amounts of precipitation.

Exploration

We now begin our exploration of the data by first looking at the bird occurrences and how they vary by time, followed by how they vary by space. We then look at the precipitation data and how it varies by time.

Bird Occurrences and Time

The first task is to explore how do species occurrences vary by time. It's well known that birds often migrate based on a seasonal patterns, part of this behavior may be captured in the following time series.

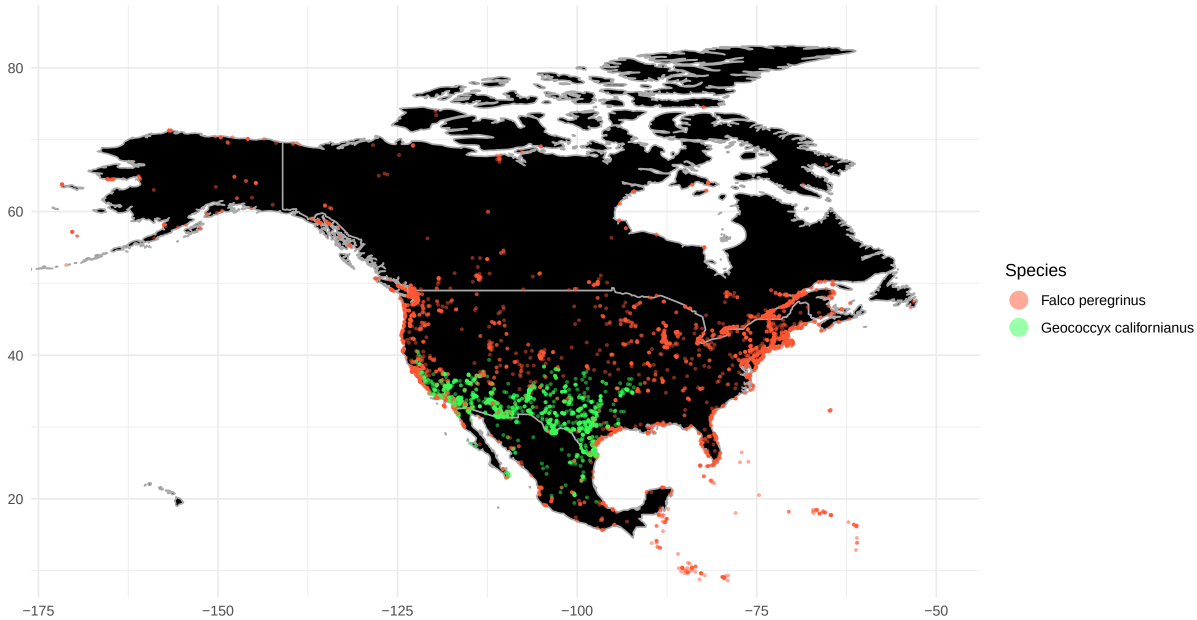


It appears that some bird species may prefer to migrate earlier than others. For example, The highest number of occurrences for *Falco peregrinus* appear to be in the months of September and October. Whereas it's more common to see *Geococcux californianus* in the months of April and May than other months.

Bird Occurrences and Space

Now to explore how do species occurrences vary by space.

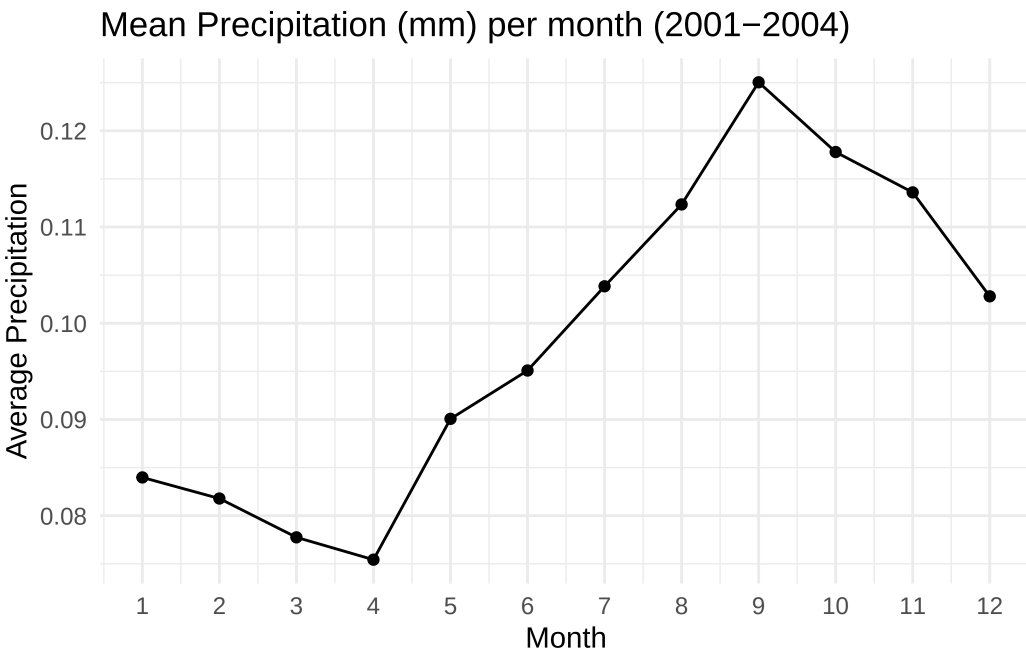
Map of Bird Spottings in North America (2004)



It appears that *Falco peregrinus* appear to stay mostly on the west and east costs of the *United States of America*, whereas *Geococcyx californianus* prefer to stay around the southern *United States* and *Mexico*.

Precipitation by Time

We now explore how precipitation varies by time, keeping in mind that precipitation is likely to be heavily influenced by the seasons.



It seems that we see the least amount of precipitation in the months of January to April, and the most in September.

The results of the exploratory analysis of snow coverage, and temperatures yield intuitive common knowledge results. Thus we conclude our exploratory analysis and we now discuss our proposed methods.

Methods

This section will outline the two methods used to create spatio-temporal predictions of bird counts. We present our methods in the following order: first we present and formally define the outcome, then describing each individual models and their pros and cons. Finally, we will describe the training and validation scheme used to train the models.

Response

The response for both models was the total counts of birds observed in a given month aggregated over a $\sim 60 \times 60$ km grid over north america. Thus, each observation of the response is indexed by both time and space. The temporal resolution is monthly starting from March 2001 and ending in December 2005, the spatial resolution is approximately $1200km^2$.

Method 1

The first approach to creating spatio-temporal predictions of bird counts was to use a ridge penalized generalized linear model using `glmnet`. The outcome was modeled as a *poisson* distribution with a *log* link function. For a discussion of generalized linear models and their applications, refer to McCullagh et. al. [22]. The predictors along with their corresponding information are included below.

Predictor	Description	Transformations
Snow Coverage	Standardized percentage of land in a $\sim 60 \times 60$ km grid covered in snow [12].	A natural cubic spline was applied to the predictor with only boundary knots.
Precipitation	Standardized average rainfall in a given month over a 60×60 km grid [17]	A natural cubic spline was applied to the covariate with 8 knots
Day Time Surface Temperature	Standardized average daytime surface temperature in a given month over a 60×60 km grid [35]	A natural cubic spline was applied to the covariate with 8 knots
Night Time Surface Temperature	Standardized average night time surface temperature in a given month over a 60×60 km grid [35]	A natural cubic spline was applied to the covariate with 8 knots
Time	Taken to be months	Each month was taken to be a factor to incorporate any potential seasonal pattern. Included via dummy encoding
Longitude and Latitude	Taken to be the center of the 60×60 km grid	None
Species	Either *Falco peregrinus* or *Geococcyx californianus*	Dummy encoding

Pros and Cons of Method 1

Method 1 benefits from it's simplicity - ridge penalized poisson regression was easy to implement using the `glmnet` package, and was not computationally expensive. On the other hand, this method does not take into account the spatio-temporal correlation structure that may be present in the data due to it's spatial and temporal components.

Method 2

The second approach to creating spatio-temporal forecasts of bird counts was to use a mixed model with spatio-temporal random effects using the `sdmTMB` package. We provide a brief review of mixed models before introducing the specifics of our model.

A review of mixed models

Mixed models are a powerful tool that allow the user to incorporate so called *random effects* that allow the user to incorporate a correlation structure to better represent their data. We breifly introduce an example from Jiang et.al [18]. Consider a study with multiple subjects, and multiple observations per subject collected over time. It's reasonable to assume, and perhaps even necessary to consider observations within each subject to be correlated. We can formulate a mixed model for this example as follows: lets assume that there are n individuals with observations collected at time points $t_1, ..., t_J$. Then a linear mixed model may be expressed as:

$$y_{ij} = X_i^T \beta + \alpha_j + \epsilon_{ij}$$

Where $X_i^T \beta$ are the fixed effects familiar to us in classical regression contexts, α_j are i.i.d mean zero Gaussian distributed random variables with covariance parameter τ and ϵ_{ij} are i.i.d mean zero Gaussian distributed random variables with covariance parameter σ^2. It follows that the correlation for observations within the same individual is given by $\frac{\tau}{\sigma^2 + \tau}$. Thus, the *random effects* α_j allow us to model the correlation structure between observations within the same individual. More complex correlation structures can be modeled by assuming a different covariance structure to the random effects, which will be demonstrated in the following section.

Adding Spatio Temporal Random Effects

We induce a spatio-temporal correlation structure to the mixed model in the exact same way outlined above. That is, we assume that observations within the same grid cell are correlated, and that observations within the same *month* and *year* are correlated. To introduce the spatio-temporal correlation structure, we use random effects to model the correlation between observations within the same grid cell and time point. We then use an *autoregressive* correlation structure to model the correlation between observations within the same month and year. The model is expressed as follows: Assume that we have n grid cells, and m time points. Then the model is expressed as:

$$y_{st} = X_{st}^T \beta + \alpha_s + \gamma_t + \epsilon_{st}$$

Where s represents a point in space, and t represents a point in time. $X_{st}^T\beta$ are the fixed effects, α_s are the spatial random effects, γ_t are the temporal random effects, and ϵ_{st} is the spatio temporal random effect. The spatial random effects are modeled using a *Matérn* kernel which assumes, that observations that are closer in space are more correlated than observations that are further apart. A further discussion of the *Matérn* kernel can be found in *Anderson et. al.* [1]. The temporal random effects can be modeled with several different correlation structures, such as an *autoregressive* correlation structure. We also can introduce a *spatio-temporal* correlation structure which assumes that observations that are closer in space and time are more correlated than observations that are further apart. The details of the correlation structure used in this model are provided in the next section.

Predictors, and Correlation Structure

The predictors used in this model are the same as the predictors used in the first model. Due to the flexibility of the `sdmTMB` package, we were able to include more flexible basis transformations for the predictors outlined in the following table:

Predictor	Description	Transformations
Snow Coverage	Standardized percentage of land in a $\sim 60 \times 60$ km grid covered in snow [12].	Thin plate regression spline
Precipitation	Standardized average rainfall in a given month over a 60×60 km grid [17]	Thin plate regression spline
Day Time Surface Temperature	Standardized average daytime surface temperature in a given month over a 60×60 km grid [35]	Thin plate regression spline
Night Time Surface Temperature	Standardized average night time surface temperature in a given month over a 60×60 km grid [35]	Thin plate regression spline
Time	Taken to be months	Cyclic cubic spline with 6 knots

We assume that the spatio-temporal random effects follow an AR(1) process so that observations that are closer in time and space are more correlated than observations that are further apart. We also used a $Matrn$ covariance function to model the spatial correlation structure.

The distribution of the response was chosen to be a poisson distribution, with a log link funciton.

Pros and Cons of Method 2

Method 2 allows us to capture the spatio-temporal correlation structure present in the data, which may lead to more accurate and stable predictions. However, this method is computationally expensive, and it's computationally unfeasible to include a ridge penalty in the model.

Training and Validation

Due to the spatio-temporal dependence possibly present in the data, we cannot simply use regular k-fold cross validation to train the models. Instead, we the `blockCV` to create spatial blocks of the data, on top of that we block by each year. We saved the block from 2005 for comparing each models, and the blocks from 2001-2004 are used to train the models, using cross validation to find the optimal ridge parameter for method 1.

Results

We present two methods of comparison for this analysis. First we compare the test errors achieved, and then compare effect sizes of the covariates included in each model.

Test Error

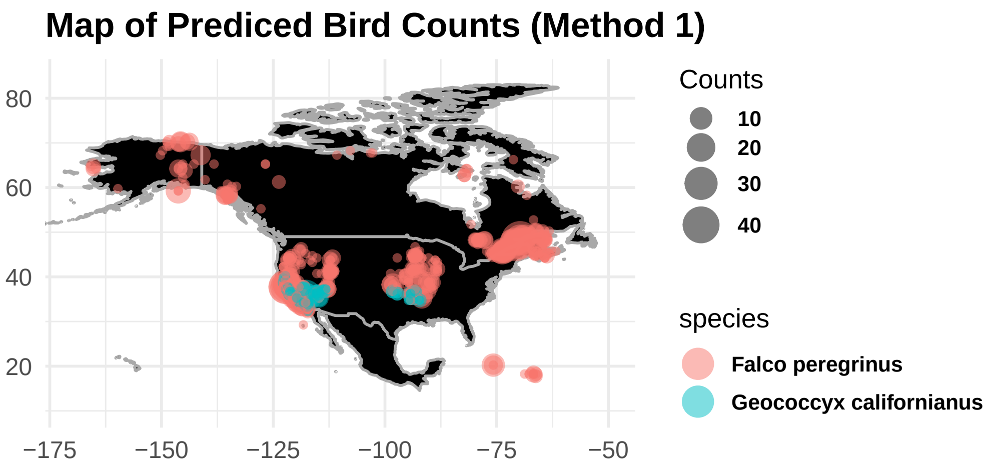
The models achieved the following test error:

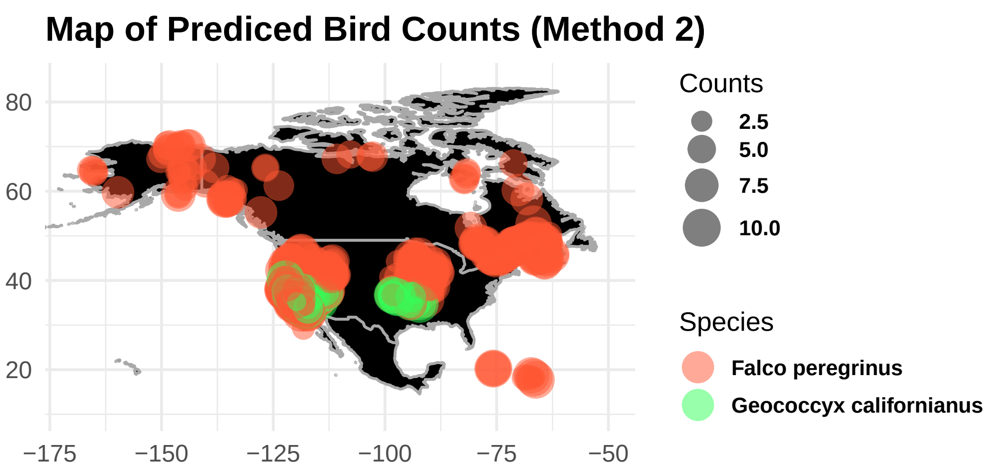
	Test Error
Method 1	1377.37
Method 2	30.88

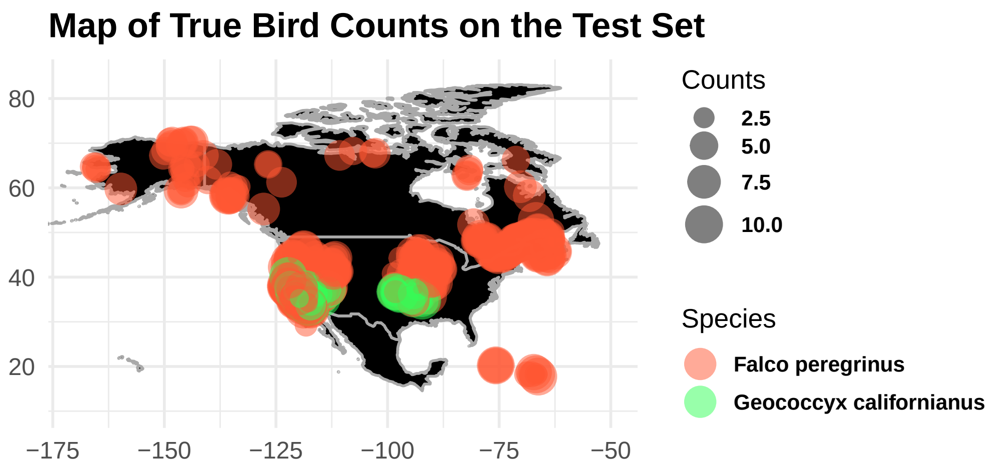
We see that method 2 achieved a much lower test error than method 1, whereas method 1 appears to be giving rather unstable predictions. The instability of the predictions of method 1 may be due to the spatial blocks from the training set being far apart from the test set since latitude and longitude were used directly for prediction in the first model.

Visual Comparison

We now provide a visual comparison of the models to provide some insight on how realistic the predictions are from each model







We see that the predictions from method 2 be in the same range of the true bird counts, whereas the predictions from method 1 are not in the same range of the true bird counts and in fact we appear to have overestimated the number of birds.

Effect Size

We present the estimated effect sizes for the second method

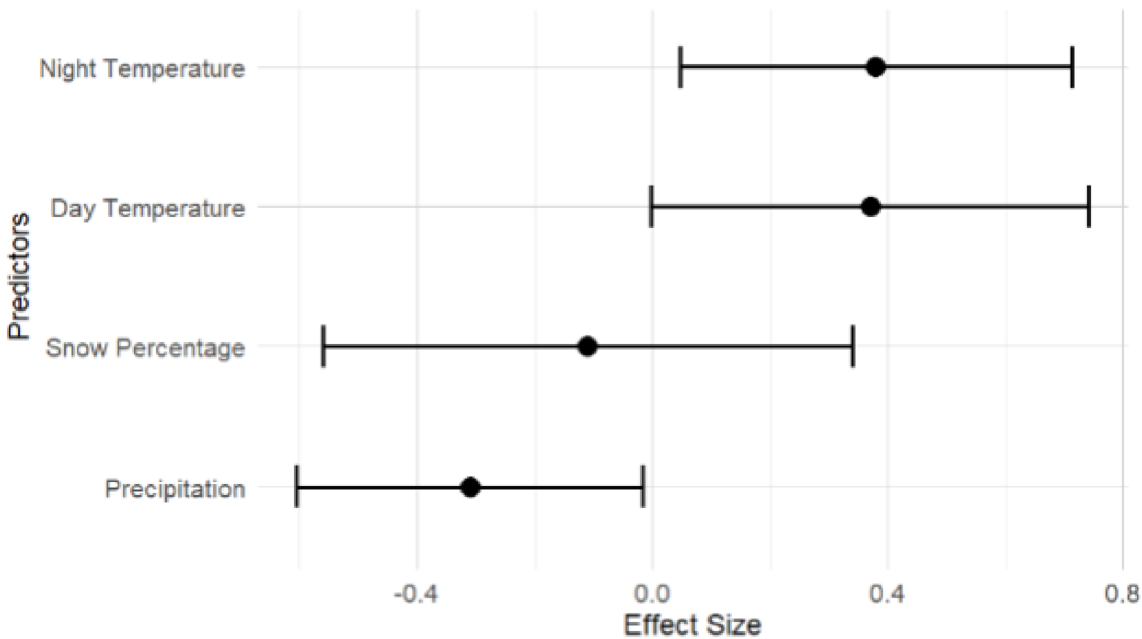


Figure 1: Figure: Effect Sizes for Method 2

We see that night and day temperatures had the largest effect for the number of birds spotted, and that higher temperatures were associated with more birds. As one might expect, precipitation had a negative relationship with the number of bird occurrences. Surprisingly, snow percentage didn't have a substantial effect with the number of bird occurrences.

Limitations

There were several limitations involved in building the second model. The first limitation was due to computational cost. To address the computational problems, we limited the dataset to only include observations where at least 1 bird was observed. However, we would like to not only create predictions based on counts, but also forecasts on where birds will be, hence a larger data set would be required and big data tools would be needed. Secondly, there was the nature of the missingness of observations of the *eBird* data set. Observations are only detected once a birdwatcher actually submits the observation, however there are more bird watchers in some areas than others, thus there are certain areas where the number of bird occurrences are more under reported than others. Furthermore, it stands to reason that people are less likely to be outdoors when the weather is colder, and whenever there's rain or precipitation, thus there are certain climate conditions where the number of bird occurrences are more under reported than others. Finally, the last limitation was data availability, NASA precipitation data was not available in more recent years, which limited the study to the years that were chosen for the analysis.

Conclusions

To conclusion, we have presented two methods for creating spatio-temporal predictions of bird counts in North America. The first method was a simple ridge penalized poisson regression model, and the second was a mixed model with spatio-temporal random effects. We found that the second model achieved a much lower test error than the first model, and that the second model was able to capture the spatio-temporal correlation structure present in the data. However, the second model was very computationally expensive. Despite this fact, we were able to conclude that night and day temperatures had the largest effect on the number of birds spotted out of the covariates included in the model, and that higher temperatures were associated with more birds. We also concluded that precipitation had a negative relationship with the number of bird occurrences. Surprisingly, snow percentage didn't have a substantial effect with the number of bird occurrences. We conclude that while including climate factors as well as spatio-temporal correlations were important in predicting bird counts, limitations in the data make it hard to create reliable long term forecasts using these methods.

Further Discussion

Further work could be done to improve the second model. Primarily, we could attempt to create forecasts for the probability of birds being in a certain location at a certain time by using a logistic model. This would require a substantially larger data set to be used, and would require big data tools to conduct the analysis. Furthermore, we could attempt to preprocess the data using spatial clustering methods based on climate features to create some more interpretable results. Finally, we could attempt to try to address the missingness of the data by including other sources of bird occurence detection, such as radar data.

References

[1] Anderson, S C, Ward, E J, English, P A, Barnett, L A K and Thorson, J T (2024). sdmTMB: An r package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields. *bioRxiv*. **2022.03.24.485545**

[2] Barrett, T, Dowle, M, Srinivasan, A, Gorecki, J, Chirico, M, Hocking, T and Schwendinger, B (2024). *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table

[3] Dale, V R, Bolton, M, Dornelas, M, Magurran, A E, Dennis, R, Broad, R, Riddiford, N J, Harvey, P V, Riddington, R, Shaw, D N, Parnaby, D and Reid, J M (2024). Among-species variation in six decades of changing migration timings explained through ecology, life-history and local migratory abundance. *Global Change Biology*. **30** e17400. https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.17400

[4] Fischer, B, Smith, M and Pau, G (2024). *Rhdf5: R Interface to HDF5*. https://bioconductor.org/packages/rhdf5

[5] Friedman, J, Hastie, T and Tibshirani, R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*. NIH Public Access. **33** 1

[6] Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Pedro, A, Sciaini, Marco, Scherer and Cédric (2024). *viridis(Lite) - Colorblind-Friendly Color Maps for r*. https://sjmgarnier.github.io/viridis/

[7] GBIF.Org User (2024). Occurrence Download. The Global Biodiversity Information Facility. https://www.gbif.org/occurrence/download/0013132-241024112534372

[8] GBIF.Org User (2024). Occurrence Download. The Global Biodiversity Information Facility. https://www.gbif.org/occurrence/download/0014753-241024112534372

[9] GBIF.Org User (2024). Occurrence Download. The Global Biodiversity Information Facility. https://www.gbif.org/occurrence/download/0012878-241024112534372

[10] Gordo, O (2007). Why are bird migration dates shifting? A review of weather and climate effects on avian migratory phenology. *Climate Research*. **35** 37–58. https://www.mendeley.com/catalogue/054d20ba-5a3c-324f-98a8-27681c141ca0/

[11] Gordo, O (2007). Why are bird migration dates shifting? A review of weather and climate effects on avian migratory phenology. *Climate Research*. **35** 37–58. https://www.int-res.com/abstracts/cr/v35/n1-2/p37-58/

[12] Hall, D K and Riggs, G A (2015). MODIS/terra snow cover monthly L3 global 0.05 deg CMG. *(No Title)*. NASA National Snow; Ice Data Center Distributed Active Archive Center

[13] Hijmans, R J (2024). *Terra: Spatial Data Analysis*. https://CRAN.R-project.org/package=terra

[14] Huffman, G J, Stocker, E F, Bolvin, D T, Nelkin, E J and Tan, J (2024). GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V07. Research Data Archive at the National Center for Atmospheric Research, Computational; Information Systems Laboratory. https://rda.ucar.edu/datasets/dsd736000/

[15] Huffman, G J, Stocker, E F, Bolvin, D T, Nelkin, E J and Tan, J (2024). GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V07. Research Data Archive at the National Center for Atmospheric Research, Computational; Information Systems Laboratory. https://rda.ucar.edu/datasets/dsd736000/

[16] Huffman, G, Stocker, E, Bolvin, D, Nelkin, E and Tan, J (2024). GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V07. UCAR/NCAR - Research Data Archive. https://rda.ucar.edu

[17] Huffman, G J, Bolvin, D T, Braithwaite, D, Hsu, K, Joyce, R, Xie, P and Yoo, S-H (2015). NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). *Algorithm theoretical basis document (ATBD) version*. **4** 2020–05

[18] Jiang, J and Nguyen, T (2021). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer New York, New York, NY. https://link.springer.com/10.1007/978-1-0716-1282-8

[19] Kain, M P and Bolker, B M (2019). Predicting west nile virus transmission in north american bird communities using phylogenetic mixed effects models and eBird citizen science data. *Parasites & vectors*. Springer. **12** 1–22

[20] Leibfried, F, Dutordoir, V, John, S T and Durrande, N (2022). A Tutorial on Sparse Gaussian Processes and Variational Inference. arXiv. http://arxiv.org/abs/2012.13962

[21] magazine, J E G / M S C / W and Magazine, W Guía de aves en Waste magazine. *Waste Magazine*. https://wastemagazine.es

[22] McCullagh, P (2019). *Generalized Linear Models*. Routledge

[23] Miller-Rushing, A J, Lloyd-Evans, T L, Primack, R B and Satzinger, P (2008). Bird migration times, climate change, and changing population sizes. *Global Change Biology*. **14** 1959–72. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2008.01619.x

[24] Müller, K (2020). *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here

[25] NASA/METI/AIST/Japan Spacesystems And U.S./Japan ASTER Science Team (2001). ASTER Level 1A Data Set - Reconstructed, unprocessed instrument data. NASA EOSDIS Land Processes Distributed Active Archive Center. https://lpdaac.usgs.gov/products/ast_l1av003/

[26] NASA/METI/AIST/Japan Spacesystems And U.S./Japan ASTER Science Team (2019). ASTER Global Digital Elevation Model V003. NASA EOSDIS Land Processes Distributed Active Archive Center. https://lpdaac.usgs.gov/products/astgtmv003/

[27] NASA/METI/AIST/Japan Spacesystems And U.S./Japan ASTER Science Team (2019). ASTER Global Water Bodies Database V001. NASA EOSDIS Land Processes Distributed Active Archive Center. https://lpdaac.usgs.gov/products/astwbdv001/

[28] Richard A. Becker, O S code by, Ray Brownrigg. Enhancements by Thomas P Minka, A R Wilks R version by and Deckmyn., A (2023). *Maps: Draw Geographical Maps*. https://CRAN.R-project.org/package=maps

[29] Satterthwaite, D (2009). The implications of population growth and urbanization for climate change. *Environment and Urbanization*. **21** 545–67. https://doi.org/10.1177/0956247809344361

[30] Satterthwaite, D (2009). The implications of population growth and urbanization for climate change. *Environment and Urbanization*. **21** 545–67. https://journals.sagepub.com/doi/10.1177/0956247809344361

[31] Shmulik (2024). Understanding Bird Migrations: Patterns and Paths. *Birds Dive Center*. https://birdsdivecenter.com/understanding-bird-migrations-patterns-and-paths/

[32] Somveille, M, Manica, A, Butchart, S H M and Rodrigues, A S L (2013). Mapping Global Diversity Patterns for Migratory Birds. *PLOS ONE.* **8** e70907. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0070907

[33] Valavi, R, Elith, J, Lahoz-Monfort, J J and Guillera-Arroita, G (2019). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution.* **10** 225–32

[34] Van Doren, B M and Horton, K G (2018). A continental system for forecasting bird migration. *Science.* **361** 1115–8. https://www.science.org/doi/10.1126/science.aat7526

[35] Wan, Z, Hook, S and Hulley, G (2021). MODIS/terra land surface temperature/emissivity monthly L3 global 0.05 deg CMG V061. *(No Title).* NASA EOSDIS Land Processes Distributed Active Archive Center

[36] Wickham, H (2016). *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org

[37] Wickham, H, Averick, M, Bryan, J, Chang, W, McGowan, L D, François, R, Grolemund, G, Hayes, A, Henry, L, Hester, J, Kuhn, M, Pedersen, T L, Miller, E, Bache, S M, Müller, K, Ooms, J, Robinson, D, Seidel, D P, Spinu, V, Takahashi, K, Vaughan, D, Wilke, C, Woo, K and Yutani, H (2019). Welcome to the tidyverse. *Journal of Open Source Software.* **4** 1686

[38] Wood, S N (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* Oxford University Press. **73** 3–36

[39] Zhang, W, Wei, J and Xu, Y (2023). Prioritizing global conservation of migratory birds over their migration network. *One Earth.* **6** 1340–9. https://www.sciencedirect.com/science/article/pii/S2590332223003962

[40] Spatial and Spatiotemporal SPDE-Based GLMMs with TMB. https://pbs-assess.github.io/sdmTMB/#overview

Supplementary Material

Libraries

```
library(rhdf5); library(ggplot2); library(here); library(tidyverse);
 ↪  library(terra); library(data.table);library(viridis); library(maps);
 ↪  library(mgcv); library(blockCV); library(sdmTMB); library(splines);
 ↪  library(glmnet); library(sf)
```

Data Cleaning and Processing

Here is the code that we used to read the raster data and read into into several dataframes.

```
### Helper Functions ###
process_hdf5_files <- function(hdf5_dir, variable_name, crop_extent,
 ↪  output_file, aggregation = NULL) {
  hdf5_files <- list.files(hdf5_dir, full.names = TRUE)

  # Initialize data table
  src <- rast(hdf5_files[[1]])
  variable_data <- src[[variable_name]]
  variable_data <- crop(variable_data, ext(crop_extent))
  data_table <- as.data.table(variable_data, xy = TRUE)

  data_table <- data_table |> mutate(
    index = 1:nrow(data_table),
    month = rep(1, nrow(data_table)),
    year = rep(2000, nrow(data_table))
  )

  row_size <- nrow(data_table)

  # Process subsequent files
  for (i in 2:length(hdf5_files)) {
    src <- rast(hdf5_files[[i]])
    variable_data <- src[[variable_name]]
    variable_data <- crop(variable_data, ext(crop_extent))
    df <- as.data.table(variable_data, xy = TRUE)
    df <- df |> mutate(
      index = row_size + 1:nrow(df),
      month = rep(i %% 12, nrow(df)),
      year = rep(2000 + (floor(i / 12)), nrow(df))
```

```
    )
    row_size <- row_size + nrow(df)

    if (!is.null(aggregation)) {
      df <- df |> group_by(x, y, month, year) |> summarise(
        mean_value = mean(get(variable_name), na.rm = TRUE)
      )
    }

    data_table <- merge(data_table, df, by = intersect(names(data_table),
 ↪  names(df)), all = TRUE)
    print(paste("Processed file:", i, "/", length(hdf5_files)))
  }

  fwrite(data_table, output_file)
  return(data_table)
}

### Processing Temperature Day Data ###
hdf5_dir <- here::here('Temperature_Day')
temperature_day_data <- process_hdf5_files(
  hdf5_dir = hdf5_dir,
  variable_name = "LST_Day_CMG",
  crop_extent = c(-170, -50, 10, 85),
  output_file = here('clndat', 'temperature_data_day.csv')
)

### Processing Temperature Night Data ###
temperature_night_data <- process_hdf5_files(
  hdf5_dir = hdf5_dir,
  variable_name = "LST_Night_CMG",
  crop_extent = c(-170, -50, 10, 85),
  output_file = here('clndat', 'temperature_data_night.csv')
)

### Processing Snow Cover Data ###
hdf5_dir <- here::here('Snow Cover')
snow_data <- process_hdf5_files(
  hdf5_dir = hdf5_dir,
  variable_name = "Snow_Cover_Monthly_CMG",
  crop_extent = c(-170, -50, 10, 85),
  output_file = here('clndat', 'snow_cover_data.csv'),
```

```r
  aggregation = TRUE
)

# Filter Snow Cover Data
filtered_snow_data <- snow_data |> mutate(
  cloudy = ifelse(Snow_Cover_Monthly_CMG == 250, 1, 0)
) |> filter(
  Snow_Cover_Monthly_CMG != 254,
  Snow_Cover_Monthly_CMG != 253,
  Snow_Cover_Monthly_CMG != 211
)

fwrite(filtered_snow_data, here('clndat', 'filtered_snow_data.csv'))

### Processing Precipitation Data ###
hdf5_dir <- here::here('Precipitation')
precipitation_data <- process_hdf5_files(
  hdf5_dir = hdf5_dir,
  variable_name = "precipitation",
  crop_extent = c(900, 2000, 2000, 3500),
  output_file = here('clndat', 'precipitation_data.csv')
)
```

Here, we joined all the data sets together to create a single data set for analysis.

```r
setwd(here("clndat"))

bird_dat = fread("bird.csv")
night_temp = fread("Night_temp_906.csv")
day_temp = fread("day_temp_906.csv")
precip = fread("precipitation_summarized_906.csv")
snow = fread("snow_data_906.csv")

# Handle month 0

snow$year_month <- sapply(snow$date, function(date) {
  parts <- unlist(strsplit(date, "-"))
  year <- as.integer(parts[1])
  month <- as.integer(parts[2])

  if (month == 0) {
```

```r
    year <- year - 1
    month <- 12
  }

  return(paste0(year, "-", month))
})

# Convert to Date object (assuming the first day of the month)
snow$time <- as.Date(paste0(snow$year, '-', snow$month, "-01"), format =
↪   "%Y-%m-%d")
# Shift dates by months
snow$time <- snow$time %m+% months(14)  # Shift forward by 3 months
bird_dat$time <- as.Date(paste0(bird_dat$year, "-", bird_dat$month, "-01"),
↪   format = "%Y-%m-%d")
night_temp$time <- as.Date(paste0(night_temp$year, "-", night_temp$month,
↪   "-01"), format = "%Y-%m-%d")
day_temp$time <- as.Date(paste0(day_temp$year, "-", day_temp$month, "-01"),
↪   format = "%Y-%m-%d")
precip$time <- as.Date(paste0(precip$year, "-", precip$month, "-01"), format
↪   = "%Y-%m-%d")

precip = precip |>
  rename(
    x = x_bin_center,
    y = y_bin_center
  )

snow = snow |> select(-year_month, -year, -month)
precip = precip |> select(-year, -month)
night_temp = night_temp |> select(-year, -month)

day_temp = day_temp |> select(-year, -month)

bird_dat = bird_dat |> select(-year, -month)

bird_dat = bird_dat |>
  left_join(night_temp, by = c("x", "y", "time")) |>
  left_join(day_temp, by = c("x", "y", "time")) |>
  left_join(precip, by = c("x", "y", "time")) |>
  left_join(snow, by = c("x", "y", "time"))
```

```
bird_dat = drop_na(bird_dat)
bird_dat = bird_dat |>
  select(
    x, y, species, n, time, mean_Night_temp, mean_day_temp,
↪   mean_precipitation, mean_snow
  )

fwrite(bird_dat, "ANALYSIS_DATA.csv")
```

Model Training

```
# Set the seed
set.seed(1)

# Load the data
bird_dat <- fread('clndat/ANALYSIS_DATA.csv')

# Filter and process the data
bird_dat <- bird_dat |>
  group_by(species) |>
  summarise(n = n()) |>
  filter(species %in% c('Falco peregrinus', 'Geococcyx californianus'))

# Save the data from 2006 for model checking
format(bird_dat$time, '%Y') %>% unique()

bird_dat <- bird_dat |>
  mutate(species = factor(species))

sf_data <- st_as_sf(bird_dat, coords = c('x', 'y'), remove = FALSE)
spatial_block <- cv_spatial(x = sf_data, k = 3, column = 'species')
sf_data$spatial_block <- spatial_block$folds_ids
bird_dat <- as.data.table(sf_data)

# Split the data into TRAIN and TEST sets
TRAIN <- bird_dat |>
  filter(format(time, '%Y') != '2005') |>
  filter(spatial_block != 3)

TEST <- bird_dat |>
  filter(format(time, '%Y') == '2005') |>
```

```r
  filter(spatial_block == 3)

# Temporal and spatio-temporal block assignments
TRAIN <- TRAIN |>
  mutate(
    temporal_block = as.numeric(factor(format(time, '%Y'))),
    spatio_temporal_block = as.numeric(factor(paste(spatial_block,
     ↪  temporal_block, sep = "_")))
  ) |>
  select(-geometry, -spatial_block, -temporal_block)

# Standardize TRAIN and TEST data
standardize <- function(df) {
  df |>
    mutate(
      snow = mean_snow / sqrt(var(mean_snow)),
      day = mean_day_temp / sqrt(var(mean_day_temp)),
      night = mean_Night_temp / sqrt(var(mean_Night_temp)),
      precip = mean_precipitation / sqrt(var(mean_precipitation)),
      time_num = as.numeric(factor(time)),
      month = as.numeric(factor(format(time, '%m')))
    )
}

TRAIN <- standardize(TRAIN)
TEST <- standardize(TEST)

# Build model matrix
X <- model.matrix(~ x + y + factor(month) + species + ns(snow, 2) + ns(day,
 ↪  10) + ns(night, 10) + ns(precip, 10), data = TRAIN)

# Fit a GLMNET model
cv_glmnet <- cv.glmnet(x = X, y = TRAIN$n, family = 'poisson', alpha = 0,
 ↪  foldid = TRAIN$spatio_temporal_block)
LAMBDA_MIN <- cv_glmnet$lambda.min
final_model <- glmnet(x = X, y = TRAIN$n, family = 'poisson', alpha = 0,
 ↪  lambda = LAMBDA_MIN)

# Predictions for GLMNET
prediddictions_cv_glmnet <- predict(final_model, newx = model.matrix(~ x + y +
 ↪  factor(month) + species + ns(snow, 2) + ns(day, 10) + ns(night, 10) +
 ↪  ns(precip, 10), data = TEST), type = 'response')
```

```
# Fit an sdmTMB model
mesh <- make_mesh(TRAIN, xy_cols = c("x", "y"), cutoff = 5)
gp_mod <- sdmTMB(
  formula = n ~ 1 + s(month, bs = 'cc', k = 6) +
    s(snow) + s(day) + s(night) + s(precip),
  data = TRAIN,
  mesh = mesh,
  time = 'time_num',
  spatiotemporal = 'ar1',
  family = poisson(link = 'log')
)

# Predictions for sdmTMB
predictions_gp <- predict(gp_mod, newdata = TEST, type = 'response')

# Calculate Mean Squared Error (MSE)
mse_gp <- mean((TEST$n - predictions_gp$est)^2)
mse_glmnet <- mean((TEST$n - prediddictions_cv_glmnet)^2)
```

Exploratory Analysis and Visualization for the Report

Exploratory Analysis

```
# Importing and Cleaning Data ---------------------------------------------
here::here('bird_dat/bird_data.csv')

data <- fread(here::here('bird_dat/bird_data.csv') , select = c("species",
 ↪  "decimalLatitude",
                                                    'decimalLongitude',
                                                    'month', 'year'), quote =
                                                     ↪  "")

# View the first few rows


data = data |>
  filter(2007>year)

data = data |>
  filter(year>1999)
```

```r
data = data |>
  mutate(species = factor(species))

data |>
  group_by(month, species) |>
  summarise(
  n = n()
) |>
  ggplot(mapping = aes(x = month, y = n, color = species)) +
  geom_line() +
  geom_point() +
  theme_minimal() +
  scale_x_continuous(breaks = c(1,2,3,4,5,6,7,8,9,10,11,12))

data = data |> rename(
  lat = decimalLatitude,
  lon = decimalLongitude
)

data = data |>
  filter(lat > 10 & lat < 85) |>
  filter(lon > -170 & lon < -50)

# Data Visualization ----------------------------------------------
colors = c("#FF5733", "#33FF57", "#3357FF", "#FF33A1", "#33FFF7",
           "#FFC300", "#DAF7A6", "#FF851B", "#B10DC9", "#FFDC00")
# How do bird occurences vary in space? -----------------------------------


# Plot the world map with data points
north_america_map <- map_data("world") |>
  dplyr::filter(region %in% c("USA", "Canada", "Mexico"))

data_2004 = data |> filter(year == 2004)

ggplot() +
  geom_polygon(data = north_america_map, aes(x = long, y = lat, group =
   ↪  group), fill = "black", color = 'darkgrey') +
  geom_point(data = data_2004, aes(x = lon, y = lat, color = species), size =
   ↪  0.01, alpha = 0.5) +
  coord_fixed(xlim = c(-170, -50), ylim = c(10, 85)) +
```

```r
  theme_minimal() +
  scale_color_manual(values = colors, name = "Species") +
  guides(color = guide_legend(override.aes = list(size = 5, fill = colors,
  ↪  alpha = 0.5))) +
  labs(title = "Map of Bird Spottings in North America (2004)", x= '', y='')


data_2016 = data |> filter(year == 2006)
ggplot() +
  geom_polygon(data = north_america_map, aes(x = long, y = lat, group =
  ↪  group), fill = "black", color = 'darkgrey') +
  geom_point(data = data_2016, aes(x = lon, y = lat, color = species), size =
  ↪  0.01, alpha = 0.01) +
  coord_fixed(xlim = c(-170, -50), ylim = c(10, 85)) +
  theme_minimal() +
  scale_color_manual(values = colors, name = "Species") +
  guides(color = guide_legend(override.aes = list(size = 5, fill = colors,
  ↪  alpha = 0.5))) +
  labs(title = "Map of Bird Spottings in North America (2016)", x= '', y='')
```