IYSE6420: BirdCLEF Birdcall Distribution Maps

Anthony Miyaguchi <acmiyaguchi@gatech.edu>

December 5, 2022

Abstract

We utilize geospatial features of bird call metadata from the BirdCLEF 2022 competition to derive a distribution map for a subset of species in the Western United States. We build several Bayesian models incorporating species frequency information and remote sensing data from USGS and NASA in a regular lattice built from area-of-interest (AOI) geometries. We use hierarchical modeling to incorporate information across a subset of species and a conditional autoregressive (CAR) distribution for spatial random effects and generate several species distribution maps. All code and data are available on request and may be available publicly on GitHub/GCP, pending permission from course instructors [1].

1 Introduction

We are interested in whether we can produce and analyze species distribution maps using Bayesian methods. The BirdCLEF Challenge is a yearly competition held as part of the Conference and Labs of the Evaluation Forum. The purpose of the challenge is to classify bird call segments and species from soundscapes captured from audio recording devices deployed in the fields. The competition hosts provide a training dataset derived from xeno-canto, a crowd-sourced platform for sharing user-generated recordings of bird calls worldwide. We build species distribution maps using discrete summaries of birdcall recording observations and remote sensing data in a lattice. We fit a Poisson generalized linear model (GLM) to the data and incorporate spatial random effects to introduce smoothing.

2 Dataset

2.1 BirdCLEF 2022 Training Metadata

The BirdCLEF 2022 competition provides over 14,800 recordings from 152 species [7]. We are interested in the spatial features from the recording metadata, which are the latitude and longitude of the recording. The competition draws recordings from the xeno-canto library, but it does not reflect the entirety of the library. Each species is capped at 500 recordings in the dataset, with the frequency of recordings generally correlated with their rarity in nature and population density, among other factors.

2.2 Google Earth Engine

Google Earth Engine is a cloud-based platform for processing geospatial data [5]. We use remote sensing data from the United States Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA) to obtain elevation, temperature, land cover classification, and population density data from various sources hosted on Earth Engine.

The NASA Shuttle Radar Topography Mission dataset (`USGS/SRTMGL1_003`) [1] provides elevation data. The MODIS/Terra Land Surface Temperature/Emissivity dataset (`MODIS/006/MOD11A1`) [2] provides temperature data. The MODIS/Terra+Aqua Land Cover Type dataset (`MODIS/006/MCD12Q1`) [6] provides land cover classification statistics. The Gridded Population of the World, Version 4 dataset (`CIESIN/GPWv411/GPW_Population_Density`) [4] provides population density data.

We generate a regular lattice of polygons from area-of-interest (AOI) geometries that derive a dataset for a region at a fraction of a degree resolution. We chose California and the Western United States as our AOI due to familiarity with the region and the diversity of geography and climate. The dataset is generated over each polygon in the lattice, computing an aggregate statistic from each data source.

[1] GitHub source `https://github.com/acmiyaguchi/iyse4620-birdcall-distributions`.

3 Data and Analysis

3.1 Conditional Autoregressive Distribution

The CAR distribution is a particular form of the multivariate normal with a covariance matrix that captures the adjacency structure of the neighborhood graph. PyMC provides this distribution out-of-the-box, which we utilize for spatial random effects [3]. The following equation gives the likelihood of a conditional autoregression (CAR) distribution:

$$f(x|W,\alpha,\tau) = \frac{|T|^{1/2}}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}(x-\mu)'T^{-1}(x-\mu)\right\} \tag{1}$$

where

$$\begin{aligned} T &= (\tau D(I-\alpha W))^{-1} \\ D &= diag(\sum_i W_{ij}) \end{aligned} \tag{2}$$

We set the following notation for use in our model definitions:

$$\phi_i|\mu_i,\tau_i,\alpha,W \sim CAR(\mu_i,\tau_i,\alpha,W) \tag{3}$$

We build a simple model to observe the effects on a simple intercept and spatial random effects model.

$$\begin{aligned} Y_i &\sim Poisson(\theta) \\ \log(\theta) &= \beta_0 + \phi_i \\ \phi_i &\sim CAR(0,10^{-3},\alpha,W) \end{aligned} \tag{4}$$

We vary values of α in figure 1 using equation 3. We observe smoothing over our map based on each cell's immediate neighbor. Note that spatial random effects quickly drop off with a Manhattan distance of 1 because the autoregressive effects are only affected by the immediate neighbors.

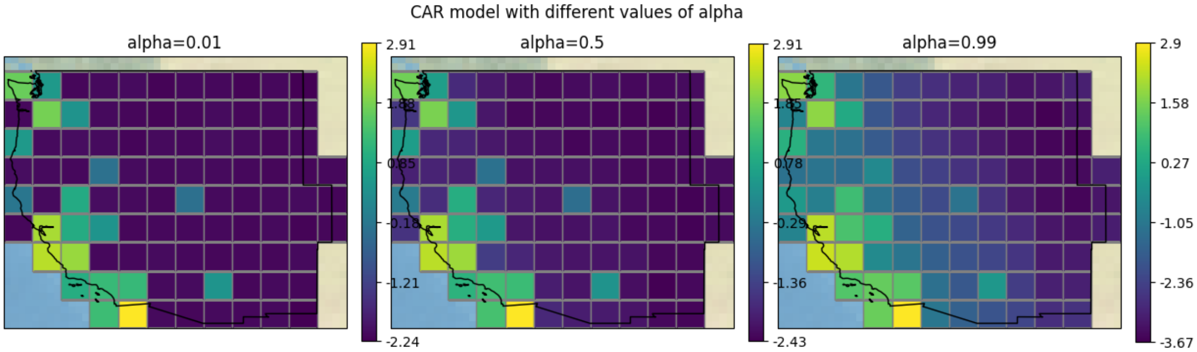

Figure 1: CAR model of the California Quail with varying alpha values, where we set missing values to 0. Higher alpha values mean more information sharing across neighbors.

3.2 Imputing Missing Values

In our dataset, we need to choose how to deal with the sparsity of observations. Our lattice is often empty, depending on the bird species' rarity and the lattice grid's resolution. We would like to include all the cells in our model to obtain the posterior predictive for cells that do not have observations. We can either let PyMC impute values for these cells implicitly using a missing completely at random (MCAR) assumption or fill in the missing values with zeros to account for no observations. We can consider this a censored problem since the dataset is the sampled accumulation of birdcall recordings since the start of the xeno-canto effort; we effectively right censor new observations. However, we simplify our modeling by not including a censoring assumption.

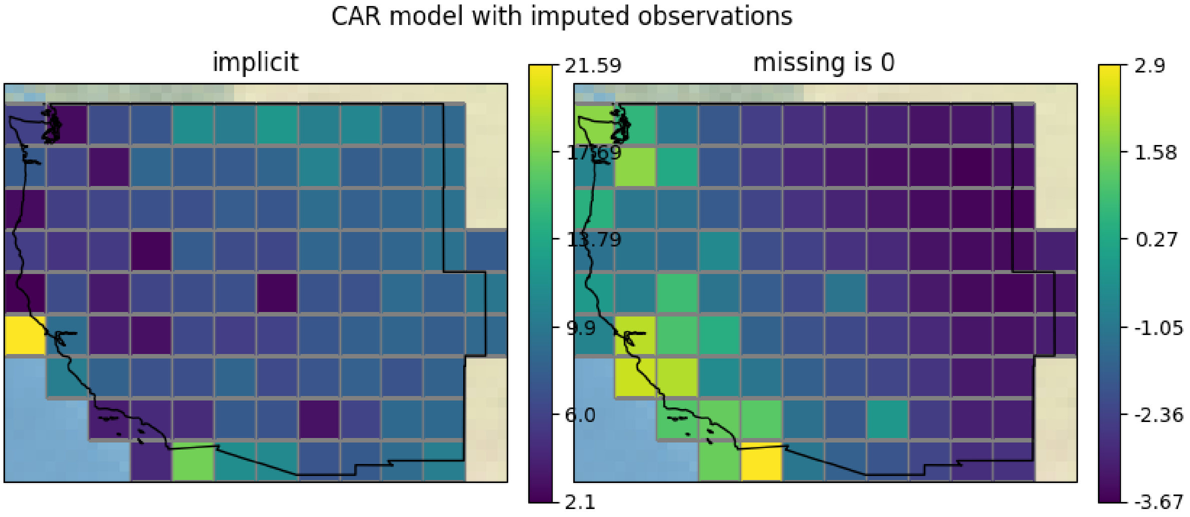

Figure 2: CAR model of the California Quail (calqua), using implicit imputation from the data (missing at random) and filling missing data with 0.

We find in figure 2 that imputing missing entries using zeros results in smoothing of the original data points, whereas the implicit imputation replaces empty entries with the posterior derived from the data. We opt to impute missing observations with zeroes for further models.

3.3 Stochastic Search Variable Selection

There are a total of 27 features that are available for modeling. Many of these features, such as land cover classification, may have a significant degree of skewness affecting the fitted model. We use a stochastic search variable selection (SSVS) to inform our choice of feature transformation.

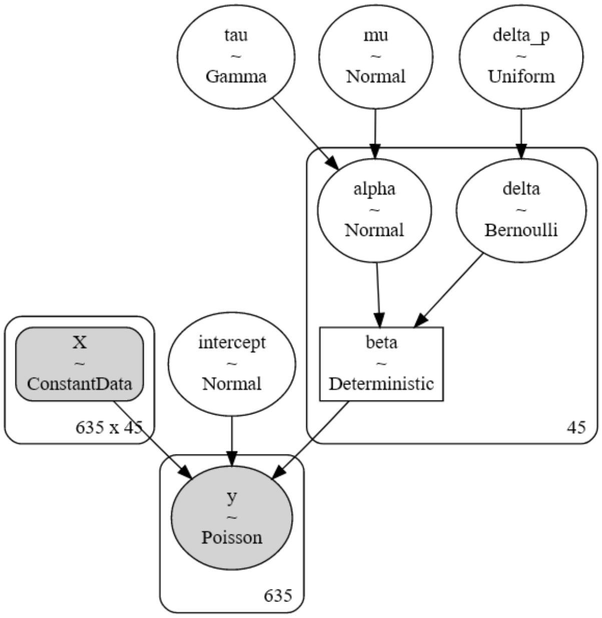

Figure 3: Graphical SSVS model procedure.

$$
\begin{aligned}
Y &\sim Poisson(\theta) \\
\log(\theta) &= \beta_0 + X\beta \\
\beta &= \delta\alpha \\
\alpha_i &\sim N(\mu, 1/\tau) \\
\mu &\sim N(0, 1000), \tau \sim Ga(0.1, 0.1) \\
\delta_i &\sim Be(p) \\
p &\sim Unif(0, 1)
\end{aligned}
\tag{5}
$$

We fit a Poisson GLM to the Western US dataset where β terms are a deterministic function of δ representing the binary likelihood that the model includes a feature and α representing the strength of the covariate in the regression, as per equation 5. We standardize all features to be zero mean and unit variance, which helps prevent model divergence. In our first model, we include all raw (or non-transformed) features and concatenate them with the log-scaled population density features and land cover classification. We do not scale elevation and temperature because these features are not heavy-tailed.

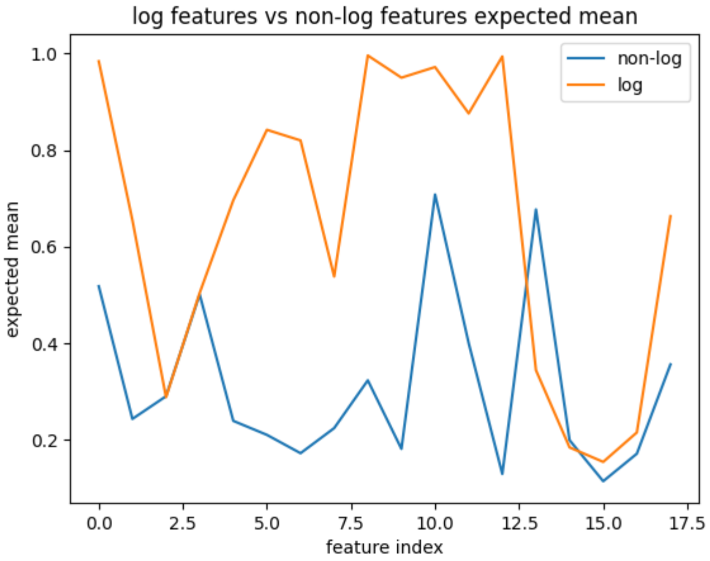

Figure 4: A comparison of non-transformed and transformed features.

	feature	probability
0	000010110000000000010010000100011001111100000	0.000450
1	100010110000000000000000000100010101111100001	0.000425
2	100010110000000000000000000100011101111100001	0.000350
3	000010110000000000000010000100011101111100001	0.000325
4	000011010000000000000000000110011001111100000	0.000250

Figure 5: Top models with Poisson GLM with non-transformed and log-scaled covariates. Each string represents a binary mask for whether a model includes a feature. The first 27 positions are non-transformed, while the last 18 are log-transformed.

We observe in figure 4 that mean values of δ are more prominent and, therefore, more likely to be included in a model. We count the instances of each possible model and find the most likely models supported by the data. In table 5, we find that the SSVS procedure favors log-scaled land cover classification features over the non-transformed features, corroborating the comparison of δ means in figure 4.

	mean	sd	hdi_3%	hdi_97%	feature_name
delta[0]	0.993	0.084	1.0	1.0	population_density
delta[1]	0.674	0.469	0.0	1.0	elevation_p5
delta[2]	0.601	0.490	0.0	1.0	elevation_p50
delta[3]	0.619	0.486	0.0	1.0	elevation_p95
delta[4]	1.000	0.000	1.0	1.0	LST_Day_1km_p5
delta[5]	0.857	0.350	0.0	1.0	LST_Day_1km_p50
delta[6]	0.733	0.442	0.0	1.0	LST_Day_1km_p95
delta[7]	1.000	0.000	1.0	1.0	LST_Night_1km_p5
delta[8]	0.651	0.477	0.0	1.0	LST_Night_1km_p50
delta[9]	0.588	0.492	0.0	1.0	LST_Night_1km_p95
delta[10]	0.991	0.094	1.0	1.0	land_cover_01
delta[11]	0.478	0.500	0.0	1.0	land_cover_02
delta[12]	0.778	0.416	0.0	1.0	land_cover_03
delta[13]	0.975	0.158	1.0	1.0	land_cover_04
delta[14]	0.983	0.130	1.0	1.0	land_cover_05
delta[15]	0.932	0.253	0.0	1.0	land_cover_06
delta[16]	0.879	0.326	0.0	1.0	land_cover_07
delta[17]	1.000	0.000	1.0	1.0	land_cover_08
delta[18]	0.991	0.095	1.0	1.0	land_cover_09
delta[19]	0.892	0.311	0.0	1.0	land_cover_10
delta[20]	1.000	0.007	1.0	1.0	land_cover_11
delta[21]	1.000	0.000	1.0	1.0	land_cover_12
delta[22]	0.708	0.455	0.0	1.0	land_cover_13
delta[23]	0.388	0.487	0.0	1.0	land_cover_14
delta[24]	0.480	0.500	0.0	1.0	land_cover_15
delta[25]	0.422	0.494	0.0	1.0	land_cover_16
delta[26]	0.972	0.165	1.0	1.0	land_cover_17

Figure 6: Probability that a model includes a feature via SSVS. We note that population density and land cover classification features are log-scaled.

	feature	probability
0	111111111111111111111111111	0.050600
1	111111111111111111111110111	0.013237
2	111111111111111111111111011	0.012725
3	111111111111111111111111101	0.009838
4	111111111110111111111111111	0.008563

Figure 7: Top models with Poisson GLM with non-transformed elevation and temperature features, and log-scaled population density and land cover classification features.

We fit a new model that drops non-transformed features for population density and land cover classification. We provide the full summary of δ posteriors in table 6. We also provide the most likely models in table 7. The most likely model includes all features in this set, so we use these features going forward.

3.4 Model Selection

We try out a variety of Poisson Generalized Linear Models (GLMs) for developing our distribution map. We chose the Poisson distribution because it best represents events distributed over geographical areas and time. These models have three main components: an intercept term, a feature term, and a random spatial effects term. We use hierarchical and multi-level modeling techniques to pool and vary information across observations, species, and geographical grid indices.

In addition to the models in equation 3 and equation 5, we include variants that allow for varied intercepts and slopes for each species. We are limited to a single instance of the CAR distribution per model, leading to certain limitations in the final distribution maps. We define our final model in equation 6.

$$
\begin{aligned}
Y_{ij} &\sim Poisson(\theta_{ij}) \\
\log(\theta_{ij}) &= \beta_{j0} + \beta_j X + \phi_i \\
\beta_{jk} &\sim N(\bar{\mu_\beta}, \bar{\sigma_\beta}) \\
\bar{\mu_\beta} &\sim N(0, 1.5), \bar{\sigma_\beta} \sim Exp(1) \\
\phi_i &\sim CAR(0, \bar{\sigma_\phi}, \alpha, W) \\
\alpha &\sim Beta(5, 1), \bar{\sigma_\phi} \sim Unif(0, 20)
\end{aligned}
\tag{6}
$$

where i indexes a lattice cell, j indexes bird species, and k indexes dataset features.

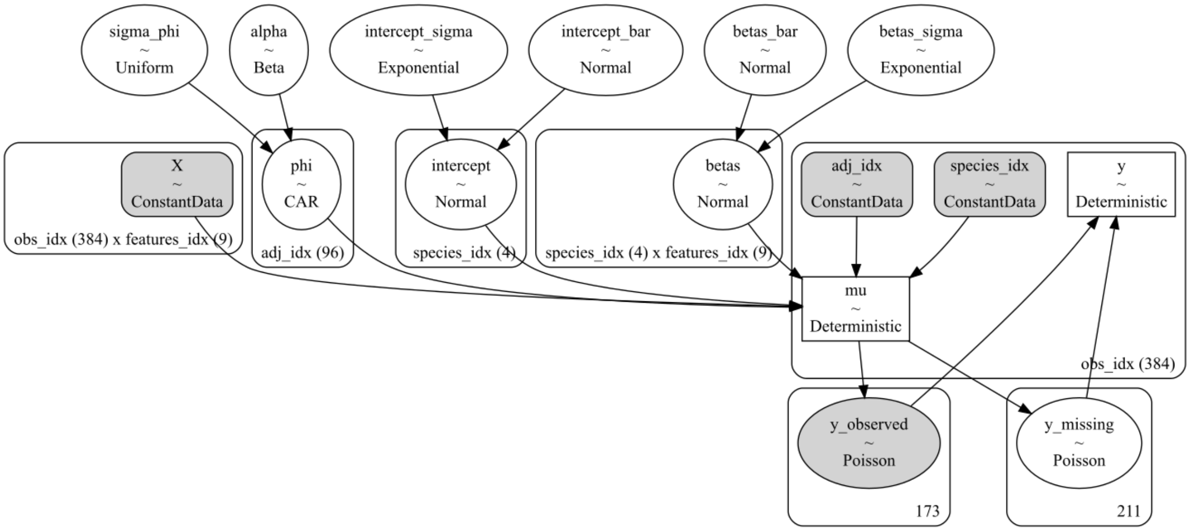

Figure 8: Varying intercept, varying covariate, CAR model

We compare our models using the widely applicable information criterion (WAIC) on a deviance scale in table 9. We choose the model with the lowest deviance to build species distribution maps. According to the ranking, the best model is the varying intercept, slope, and CAR model.

model_name	rank	elpd_waic	p_waic	elpd_diff	se	dse
varying_intercept_varying_slope_car_model	0	967.05	108.07	0.00	29.64	0.00
pooled_intercept_varying_slope_car_model	1	1237.93	198.19	270.88	54.20	44.13
varying_intercept_pooled_slope_car_model	2	1291.85	160.71	324.80	129.09	118.68
varying_intercept_car_model	3	1296.86	163.01	329.81	129.58	119.23
varying_intercept_varying_slope_model	4	1748.39	307.34	781.34	155.76	149.82
pooled_intercept_varying_slope_model	5	1986.66	376.32	1019.61	182.03	175.32
varying_intercept_pooled_slope_model	6	2094.71	292.58	1127.66	270.45	261.44
pooled_intercept_car_model	7	3384.61	634.20	2417.56	430.55	418.98
pooled_intercept_pooled_slope_model	8	3896.91	631.80	2929.86	536.62	526.20
varying_intercept_model	9	4021.92	115.95	3054.87	799.54	792.66

Figure 9: Comparison of models using WAIC on a deviance scale. Lower is better. A model component varies if there is an instantiation of a random variable for each species. A component pools if there is a shared variable across all species.

3.5 Posterior Predictive Distribution Maps

We fit our best model to the data and find the posterior predictive for all cells in the Western US (2-degree resolution) dataset. We prepare the dataset by grouping birdcall recording locations into their bounding cells and primary species labels. To avoid excess computation and sparsity, we choose a subset of bird species and set a default category of ”other”. All features are standardized to zero mean and unit variance.

We model using two variations of the prepared dataset. In figure 12, we choose the top 3 species and implicitly impute the rest of the missing cells from the posterior derived from the data. We note that this results in a total of $211/384$ or 55% of entries that need to be imputed. Sampling from the posterior takes 7 minutes and 40 seconds with 16 cores for 56000 samples. The second model uses the top 15 species to generate the dataset, which has greater sparsity with $1178/1536$ or 77% of entries that need to be imputed. With greater sparsity comes higher computational costs, with the larger dataset and model taking 37 minutes and 44 seconds to sample. We also fit models where we impute missing values with 0 ahead of time. In this case, the model takes 4 minutes and 50 seconds and 8 minutes and 2 seconds, respectively.

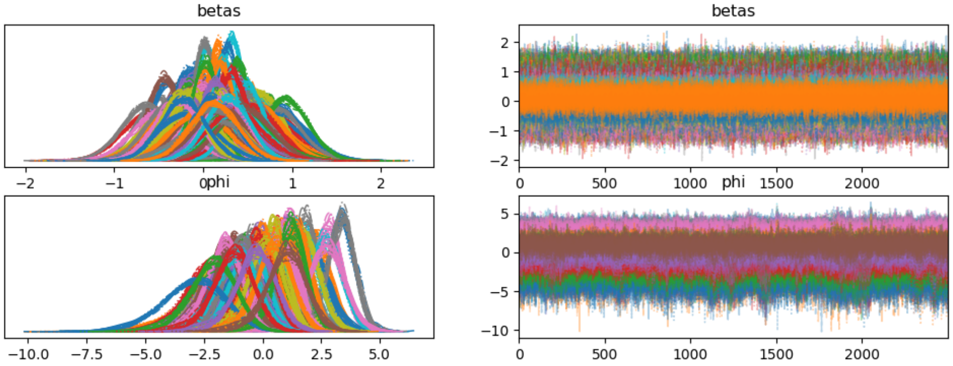

Figure 10: Trace plot of slopes β and spatial random effects ϕ.

index	mean	sd	hdi_3%	hdi_97%
betas[other, landcover_evergreen_broadleaf_forest]	0.371	0.156	0.083	0.668
betas[other, landcover_mixed_forests]	-0.391	0.195	-0.756	-0.026
betas[other, landcover_grasslands]	-0.440	0.186	-0.797	-0.092
betas[other, landcover_urban_and_built-up]	0.664	0.228	0.239	1.094
betas[other, landcover_cropland/natural_vegetat...	0.330	0.125	0.088	0.559
betas[other, landcover_permanent_snow_and_ice]	0.274	0.132	0.029	0.526
betas[other, landcover_water_bodies]	0.385	0.159	0.079	0.674

Figure 11: A subset of random slopes from the other bucket of the top 15 species that are significant given a 94% HDI.

We end up with an extensive list of significant covariate slopes that do not intersect with the origin with a 94% credible interval. We provide a subset of the significant features relative to the ”other” bucket in figure 11. We see that an urban land cover classification positively affects an observation. The urban classification relationship aligns with the distributional assumption that recordings are more likely if they are more accessible to people. We see other intuitive relationships, such as grasslands having a negative effect on birds due to the lack of nesting areas.

In our three species example in figure 12, we observe the effects of our choice of hierarchical modeling. We note the high-frequency cells (yellow) in the northern and southernmost points of the birdcall recording plot. The prediction plot shows that the model shares information across species. The likeliest source of this comes from the CAR distribution, which is instantiated once across all species. Instead of other information, the random effect of that particular geographical cell is more prominent.

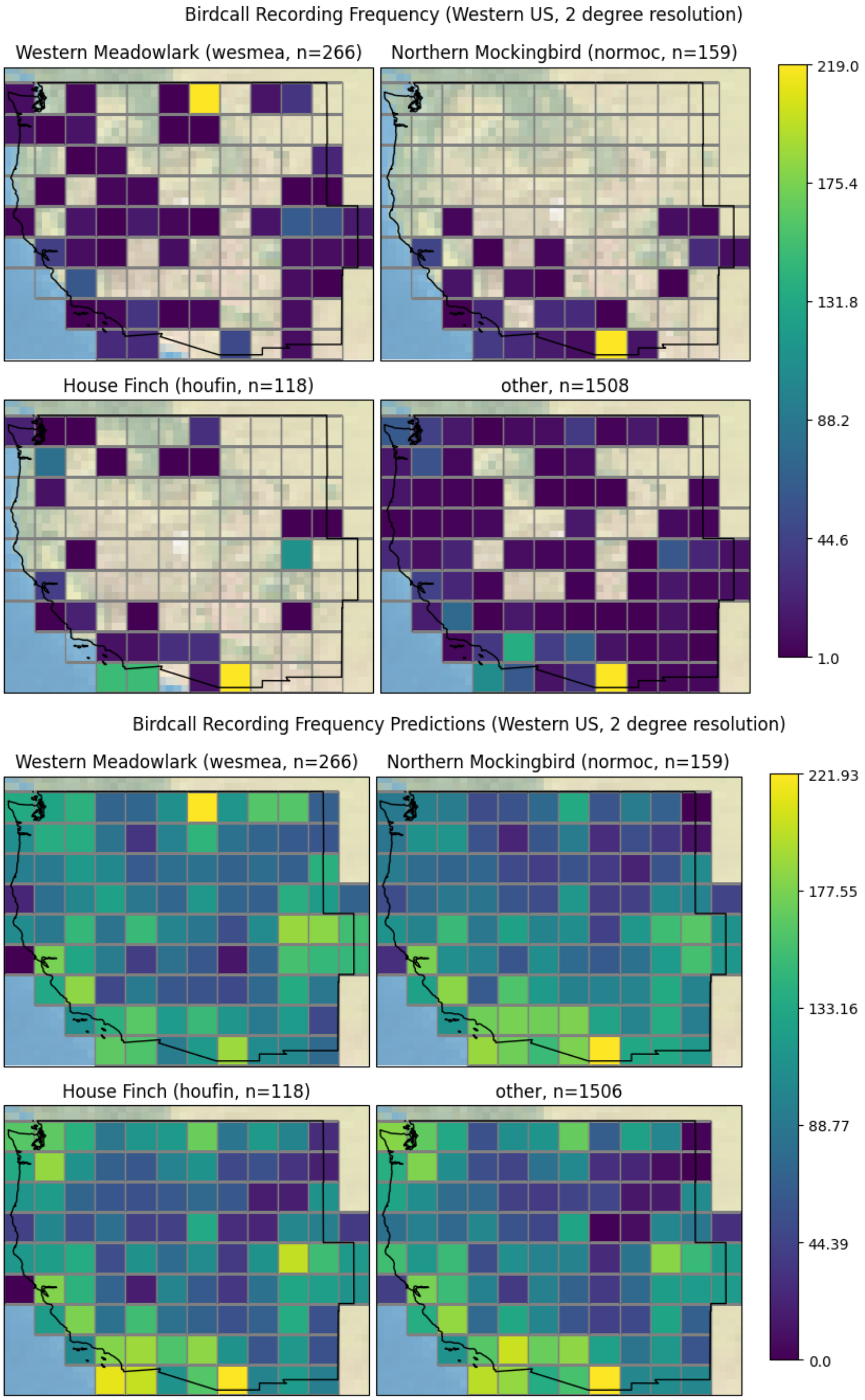
Birdcall Recording Frequency (Western US, 2 degree resolution)
Western Meadowlark (wesmea, n=266)
Northern Mockingbird (normoc, n=159)
House Finch (houfin, n=118)
other, n=1508
219.0
175.4
131.8
88.2
44.6
1.0
Birdcall Recording Frequency Predictions (Western US, 2 degree resolution)
Western Meadowlark (wesmea, n=266)
Northern Mockingbird (normoc, n=159)
House Finch (houfin, n=118)
other, n=1506
221.93
177.55
133.16
88.77
44.39
0.0

Figure 12: Frequency maps of the 15 most frequent species.

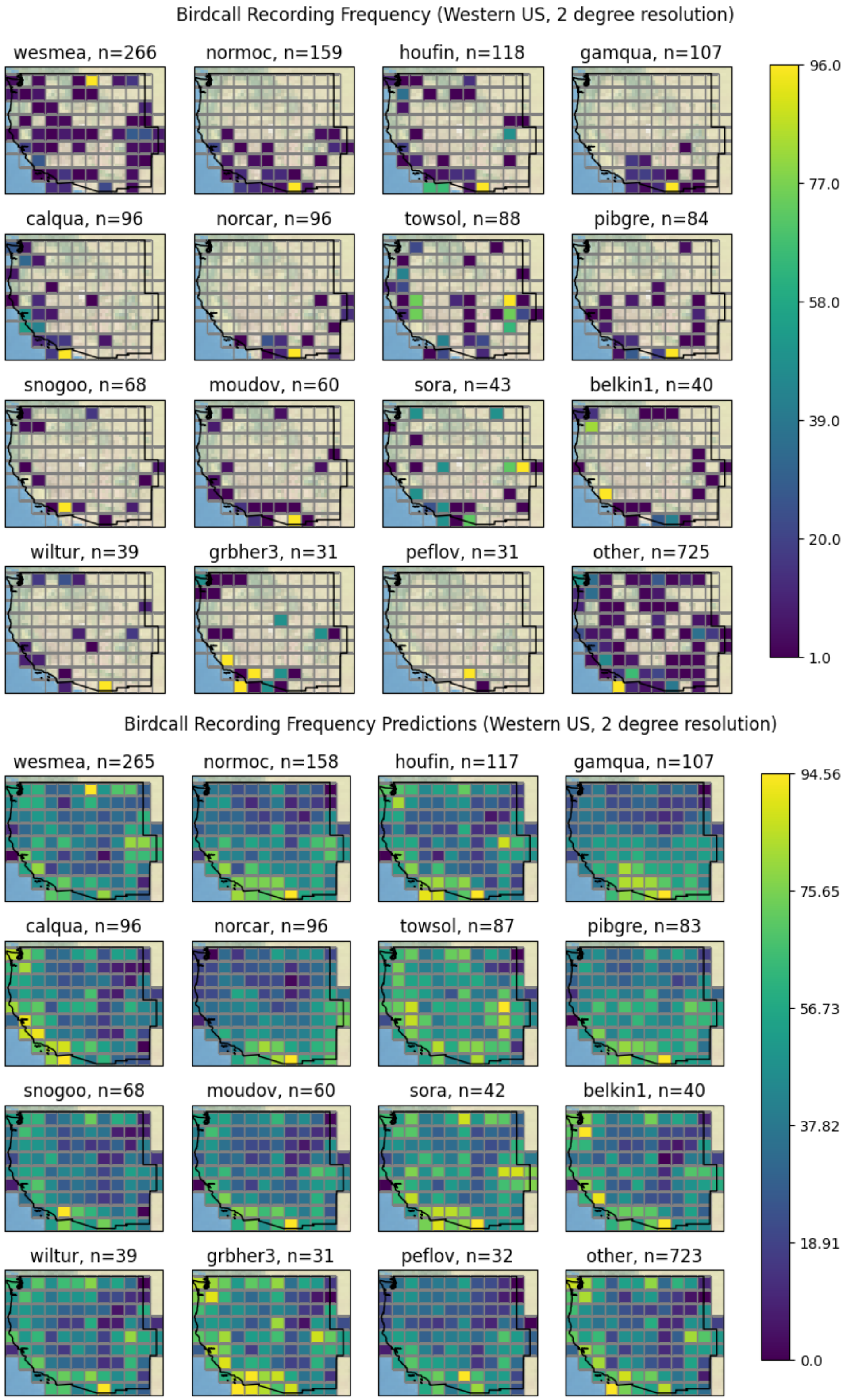

Figure 13: Frequency maps of the 15 most frequent species.

We make a few species-specific observations in the top 15 species in figure 13. First, we take a look at Gambel's Quail (gamqua) and the California Quail (calqua), which have distinctive geographical distributions localized to the US Southwest and California, respectively, as per distribution maps on eBird [9] [8]. The raw recording frequencies generally fall within this distribution. We imputed all missing values in our initial implementation as missing completely at random (MCAR). We noted that the posterior predictive for these two quail species did not accurately reflect the actual distribution of birdcall recording frequencies. Instead of treating missing observations as areas with a low likelihood of predictive values, it instead tries to impute them using a reasonable value from the posterior of observed data. The bottom of 13 includes the final model where we set missing values to 0. The model reasonably captures these quail species' spatial frequency to a reasonable degree. However, we note a strong bias toward observations across all species.

When we look at the distribution of our values of ϕ in figure 10, we see many values are more significant than 1. Because the model shares spatial random effects across species, we end up with predictive values that extend further than what the actual natural distribution might be. However, we are limited in PyMC because we cannot create a multi-dimensional CAR variable with one variable per species. If we could model random geospatial effects on a per-species basis using a shared hyper-prior on alpha, our models may fit our data better.

4 Conclusion

We made design choices during this analysis, including data collection resolution (i.e., how large are grid cells for Google Earth Engine), dealing with missing observations, feature selection, and modeling structure. We found that hierarchical models, while powerful, can introduce complicating factors that reduce the interpretability of the final species distribution map. We also have to remember the distributional semantics of the birdcall recording metadata. The set of people who use and upload to xeno-canto are more likely to visit and record birds in easily accessible regions of the world. In addition, the set of recordings included in the BirdCLEF competition is a much smaller subset than what is available from the source site.

Given more time, we want to explore the effect of the underlying dataset resolution on modeling and choose a more extensive geography that spans several continents. We also want to explore hexagonal lattices, which we can implement with trivial modifications to our source. Finally, it would be interesting to use data directly from xeno-canto instead of using the BirdCLEF competition training subset. The source dataset would contain far more data points and might make up for structural choices that we impose on our model.

On the modeling side, it would be interesting to create an instance of a CAR distribution per species. This limitation causes unwanted shrinkage and forces us to build a model per species without access to multi-level modeling techniques. Further research is needed to find multivariate distributions that accept a kernel, allowing for a more nuanced smoothness parameter across a neighborhood of cells. For example, instead of a binary adjacency matrix, we might define a weighted adjacency matrix such that influence drops at an exponential rate up to a Manhattan distance of 3 from the originating cell.

Through this project, we can build a species distribution map of geospatial metadata attached to birdcall recordings using Bayesian modeling methodologies. Using Google Earth Engine, we can incorporate information from various public sources by discretizing our data into regular lattices. Though there are computational and modeling choices that make geospatial modeling challenging, it is a reasonable way to generate species distribution maps.

References

[1] N. JPL, *NASA Shuttle Radar Topography Mission Global 1 arc second*, Type: dataset, 2013. DOI: 10.5067/MEASURES/SRTM/SRTMGL1.003. [Online]. Available: https://lpdaac.usgs.gov/products/srtmgl1v003/ (visited on 12/01/2022).

[2] Z. Wan, S. Hook, and G. Hulley, *MOD11A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006*, Type: dataset, 2015. DOI: 10.5067/MODIS/MOD11A1.006. [Online]. Available: https://lpdaac.usgs.gov/products/mod11a1v006/ (visited on 12/01/2022).

[3] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in Python using PyMC3," en, *PeerJ Computer Science*, vol. 2, e55, Apr. 2016, Publisher: PeerJ Inc., ISSN: 2376-5992. DOI: 10.7717/peerj-cs.55. [Online]. Available: https://peerj.com/articles/cs-55 (visited on 12/01/2022).

[4] C. F. I. E. S. I. N.-C.-C. University, *Gridded Population of the World, Version 4 (GPWv4): Population Density*, Type: dataset, 2016. DOI: 10.7927/H4NP22DQ. [Online]. Available: http://beta.sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density (visited on 12/01/2022).

[5] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," en, *Remote Sensing of Environment*, Big Remotely Sensed Data: tools, applications and experiences, vol. 202, pp. 18–27, Dec. 2017, ISSN: 0034-4257. DOI: `10.1016/j.rse.2017.06.031`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0034425717302900` (visited on 12/01/2022).

[6] M. Friedl and D. Sulla-Menashe, *MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006*, Type: dataset, 2019. DOI: `10.5067/MODIS/MCD12Q1.006`. [Online]. Available: `https://lpdaac.usgs.gov/products/mcd12q1v006/` (visited on 12/01/2022).

[7] S. Kahl, A. Navine, T. Denton, *et al.*, "Overview of birdclef 2022: Endangered bird species recognition in soundscape recordings," in *Proceedings of the Working Notes of CLEF 2022-Conference and Labs of the Evaluation Forum*, 2022.

[8] *California Quail - eBird*, en. [Online]. Available: `https://ebird.org/species/calqua` (visited on 12/04/2022).

[9] *Gambel's Quail - eBird*, en. [Online]. Available: `https://ebird.org/species/gamqua` (visited on 12/04/2022).