**Optimal Birdwatching in Massachusetts: Species Frequency Forecast &**

**Birding Activity Analysis**

**Authors:** Michael Aldoroty, Neil Ghosh, Nhat Pham, Hajera Siddiqui

**Github:** <https://github.com/nhathpham/Bird-Watching-Recommendation-System>

1. **Summary**

Bird watching is a valuable activity for enthusiasts, conservationists, and researchers, linking the joys of observation with the needs of environmental conservation. However, planning effective bird-watching excursions, especially for beginners, can be challenging. Current platforms, like eBird.com, while invaluable for their extensive historical data and explorative features, can be challenging to navigate and lack predictive capabilities. Our project addresses this by providing an interactive dashboard forecasting weekly bird species detection rates in Massachusetts (MA) counties, coupled with recommendations for optimal bird-watching spots and time of day for each species.

For the second phase of the project, our focus is threefold. Firstly, we aim to enhance the model's accuracy and scalability. Prophet and Greykite were our chosen forecasting algorithms due to their proven performance and scalability in the first phase. We further fine-tuned both models using parameter tuning via grid search, cross-validation, and parallel processing. After evaluating their performance and processing speed, we selected Prophet as the more efficient method to extend our forecasts to all bird species. We also established a semi-automated pipeline that covers the entire process from eBird data collection to dashboard updates, facilitating regular model refreshes with new data.

Second, incorporating weather and pollution metrics as regressors into our time series model for predicting bird detection rates was explored as an option, as these environmental factors intricately influence bird movement and migration patterns. By exploring the potential correlations between regressors and detection rates across various counties, our model gains the ability to provide more accurate and nuanced predictions.

Third, we expanded the recommendations to birders through identifying underbirded areas in MA that would benefit from more data collection. We evaluated birding activities and species diversity across the state, and used a scoring system to classify areas into five groups. We also accounted for seasonal variations, recognizing the shifting popularity of areas throughout the year. The findings are displayed via interactive leaflet maps. The maps indicate patterns and variations in birding activities across MA, effectively pinpointing underbirded locations that offer new opportunities for birdwatchers and are beneficial for conservation efforts.

We continue to use the eBird dataset from the Cornell Lab of Ornithology, focusing on data specific to MA from 2013 to 2022. This decision was influenced by our survey findings, which highlighted a local birdwatching preference among MA birders, and by resource limitations. The dataset is structured around checklists, each representing a bird-watching trip, with features like location, date, time, distance traveled, and participant count. Each row in the dataset corresponds to an observation from these checklists, detailing information about individual species. Data cleaning and filtering were performed using the 'auk' package in R, designed for eBird data. For a detailed explanation of the eBird dataset and our processing methods, please refer to our Phase 1 final report in Appendix (App.) 1.

We acquired supplemental data to support the environmental regressor implementation and the birding activity analysis. For the former, we sourced monthly temperature and precipitation data for MA counties from 2013 to 2022 from the National Centers for Environmental Information (NCEI). For the latter, we incorporated data on uncommon bird species using the American Birding Association’s 2023 species checklist and the Massachusetts Avian Records Committee’s birds-in-review list.

1. **Methods** 
   1. **Forecast models**

We built on the success of Prophet and Greykite from Phase 1, with a focus on enhancing both their speed and accuracy. Our approach involves fine-tuning the hyperparameters that influence model components such as trend, seasonality, and autoregression. Each model forecasts the detection rate for a specific bird species in a county. Due to the varying detection rate patterns across different bird-county combinations, we employ grid search and cross-validation to select the most effective hyperparameters for each model, using the Root Mean Square Error (RMSE) as our selection criterion. Given the scale of our analysis, we adopt parallel processing to improve computational efficiency. This is feasible since each bird-county model functions independently. During initial testing on 50 species, fine-tuned Prophet and Greykite showed comparable accuracy, but Prophet was three times faster, as detailed in App. 2. Hence, we prioritized the optimization of Prophet for scaling the model to all species.

* + 1. **Greykite**

The fine-tuned Greykite model incorporates piecewise linear growth and offers the option of adding trend changepoints to specify changes in growth rates. In handling changepoints, we allow three alternatives: no changepoints, automatic changepoint detection, and a customized detection. The customized method is designed to effectively manage significant yearly seasonality that isn't mitigated through aggregation, allowing for a clear distinction between long-term trends and seasonal patterns. Within this method, we apply a limit of five potential changepoints based on empirical testing. Additionally, we do not allow changepoints in the last 15% of the data period. Ridge regression is used for model fitting to mitigate the risk of overfitting. For model evaluation, we conducted a 2-fold cross-validation, utilizing RMSE and MAE metrics for accuracy assessment. Finally, parallel processing was implemented across eight cores.

* + 1. **Prophet**

Prophet is an open-source time-series forecasting tool developed by Facebook, that utilizes a Bayesian framework. The model represents time-series data by trend, seasonality, and noise elements. The trend determines the direction, seasonality determines patterns, and anything outside of these parameters is described with the noise element. Additionally, Prophet excels at handling missing data and adapting to shifts in trends (Khare, 2023).

To align with Prophet's requirements, we streamlined the dataset by removing unnecessary columns and renaming the date and target columns, along with adjusting the data types. We employed logistic regression in the model, setting the carrying capacity to one and the floor to zero. This approach intended to stabilize forecasted values as they approached these limits over time. However, in cases where forecasted values fell below zero, we adjusted forecasts to zero to maintain logical consistency.

For model optimization, we focused on tuning two parameters through cross-validation: the changepoint prior scale, which adjusts trend flexibility, and the seasonality prior scale. Testing revealed that the seasonality prior scale had a minimal impact, leading us to concentrate solely on fine-tuning the changepoint prior scale. We implemented grid search to identify the parameter setting that minimized RMSE, treating each bird and county independently. To efficiently execute this process, we developed a function that tests all potential parameters for a given bird-county pair. We leveraged Python’s multiprocessing library, specifically the “Pool” object, to facilitate parallel testing of these combinations, enhancing the speed and efficiency of our parameter tuning efforts.

After the parallelized testing process, the results of each parameter were divided into a separate data frame, for each of which we calculated the standard deviations and averages of the RMSE. Finally, the four best performing (lowest standard deviation and average) parameters were selected by only keeping the best performing cross validation parameters.

The model was run using the best changepoint prior parameter found by cross validation to make future predictions for the next two years. The resulting future dataframe and the historical data were then collected within the pool function to be combined into one dataframe that could be used for the Tableau web visualization application. Additionally, the RMSE and MAE values from the cross validation for each bird and county were saved to a CSV file.

* 1. **Environmental Regressors**

In the context of predicting bird species’ detection rates, incorporating weather and pollution metrics is crucial for enhancing model accuracy. Birds' migratory and movement behaviors are intricately tied to environmental conditions, and incorporating relevant regressors allows the model to capture these relationships. Weather variables like temperature, precipitation, and wind speed can influence bird migration patterns, breeding seasons, and overall activity. For instance, certain species may be less active or exhibit different behaviors during rainy or stormy weather conditions, directly affecting their detection rates.

Similarly, pollution factors can also play a significant role in understanding bird behaviors and detection rates. Birds are sensitive to changes in their environment, and pollution levels can impact their habitats, food sources, and overall well-being. High pollution levels might lead to alterations in migration routes or changes in foraging behavior, influencing the likelihood of bird detections in a particular area. By incorporating these environmental regressors into the time series model, we can empower it to discern and leverage the correlations between weather, pollution, and bird movement, ultimately refining the accuracy of our predictions and providing a more comprehensive understanding of the factors influencing detection rates across different regions.

The two environmental regressors we used were average temperature and precipitation. These metrics were sourced from the NCEI (NCEI, 2023), whose goal is to provide environmental data, products, and services covering the depths of the ocean to the surface of the sun to drive resilience, prosperity, and equity for current and future generations. The temperature and precipitation data from NCEI was the best publicly available data we could find since it was at the county level, which matched our dataset. The main issue with the data was that it was aggregated on a monthly basis, while our eBird data was on a weekly basis.

| **County** | **Common Name** | **No Regressor Model RMSE** | **Regressor Model RMSE** |
| --- | --- | --- | --- |
| Barnstable | Northern Cardinal | 0.0419 | 0.0422 |
| Barnstable | Tufted Titmouse | 0.0428 | 0.0443 |
| Essex | Northern Cardinal | 0.0435 | 0.0442 |
| Essex | Tufted Titmouse | 0.0419 | 0.0449 |

*Table 1. RMSE for models with vs without regressors*

As seen from the table above, the inclusion of temperature and precipitation as regressors in the Prophet model resulted in a slight increase in RMSE. This could likely be attributed to the mismatch in data aggregation levels: the NCEI dataset is monthly, while eBird data is weekly, causing identical regressor values for four data entries within the same month. Because the constant regressor values did not enhance the model, we decided to exclude them from the final model iteration. In the future, to include regressors that potentially improve model performance, a dataset aggregated on a weekly basis is needed.

* 1. **Semi-automated pipeline**

To complete the project, the model was packaged to connect all the data cleaning files with the model. This allows the application to be easily maintained, and give the users up to date and accurate predictions. Initially, data collection involves manually downloading the new eBird dataset each month in a CSV format. The manually pulled data is then input into the data cleaning file in R which creates the new\_ebd.csv file, connected by using the subprocess module. Once that file is created, the python data cleaning file uses the allbirds\_detection.feather file, along with the new\_ebd.csv file to further filter the data and obtain weekly detection rates. Following successful data cleaning, the Python script handling the Prophet model receives the final allbirds\_detection.csv file to generate forecasts for the next two years. Results, both testing and forecasting, are saved to their respective CSV files, the latter being used for Tableau updates. Updating the Tableau dashboard requires a manual step. After each update the updated dashboard is manually published or republished to Tableau Public.

* 1. **Birding activity analysis**

This analysis aims to compare birding activities and species diversity across MA using the 2022 eBird data and additional sources. We initially explored two methods for defining comparison areas. The first method, using over 3,500 eBird hotspots, was ultimately deemed unsuitable due to varying hotspot sizes, lack of hotspot-specific area data, and potential exclusion of valuable data from user-input locations. We instead adopted a second approach, dividing the state into 8km x 8km grid cells, each treated as a separate area (App. 3). This concept was inspired by a Cornell Lab project using grid maps to show bird population changes (Cornell, 2022), and refined through testing and consultation with an expert birder. While being a simplified approach, it provides a more equitable and comprehensive analysis framework.

To assess birding activities within these grid areas, we identified key metrics and processed the data accordingly. Based on feedback received from birdwatching communities, along with data constraints and research objectives, we selected four metrics: total number of checklists, total number of unique observers, median number of species per checklist, and number of uncommon species. To identify uncommon species, we used the American Birding Association’s checklist, labeling species with rarity codes of 2 or higher as uncommon (Floyd, 2023). We cross-referenced this with the Massachusetts Avian Records Committee's unofficial list (Trimble, n.d) and our project's past detection rate data, ultimately compiling a list of 156 uncommon species. After defining key metrics, we aggregated eBird data at the individual checklist level, assigned each checklist to a corresponding grid cell using its location coordinates, and then calculated the selected metrics for every cell on a monthly basis.

To categorize each grid area by its species diversity and popularity to birdwatchers, we devised a scoring system using the selected metrics. Scores were calculated as: Popularity (P score) = 0.6 \* number of rare birds + 0.4 \* median number of species, and Diversity (D score) = 0.6 \* unique observers + 0.4 \*number of checklists. The chosen weights reflect the feedback from birdwatching communities which emphasized the presence of interesting species and number of observers over checklist count. We then classified areas into five categories: ‘Underbirded’ (low P, high D), ‘Heavily Birded’ (high P, low D), ‘Popular’ (high P and D), ‘Low Interest’ (low P and D), and ‘Not Enough Information’ (fewer than 3 checklists). ‘Low’ and ‘high’ were defined relative to the median score for each metric across all areas.

Next, we created interactive choropleth maps within a Jupyter Notebook environment. We used Python’s geopandas package for handling spatial data, ipyleaflet for integrating interactive Leaflet maps, and ipywidgets for creating interactive user interfaces. Key features of the maps include a month slider for a smooth transition between viewing maps for different months, and tooltips with detailed metrics and hotspots information upon user interaction. To make the maps publicly accessible, we transformed the notebook outputs into a standalone application using the Voila package. Then, we employed Binder, a free service that builds a Docker image from our GitHub repository to create a live environment for hosting the notebook.

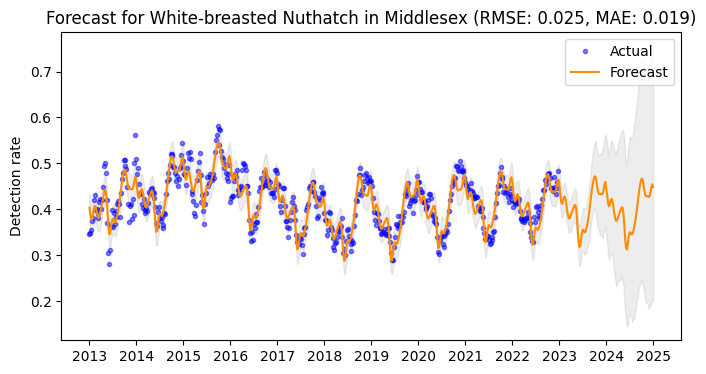
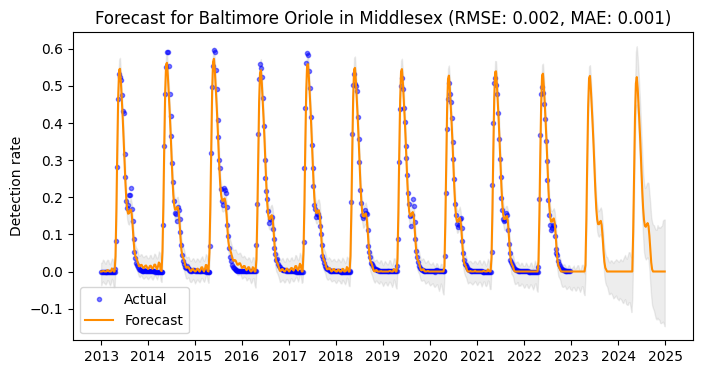
1. **Final Results**
   1. **Final Forecast Models**

The final Prophet model was run on 465 birds across all 14 MA counties, resulting in 6510 models. The process took 1 hour and 30 minutes, averaging approximately 11.6 seconds per species. This marked a significant improvement in efficiency compared to a prior version, which averaged 27.6 seconds per species. The performance of the models was satisfactory, with a median RMSE of 0.007 and median MAE of 0.005 (Table 2). As anticipated, model efficacy varied across species and counties due to differing behavioral patterns and habitat preferences.

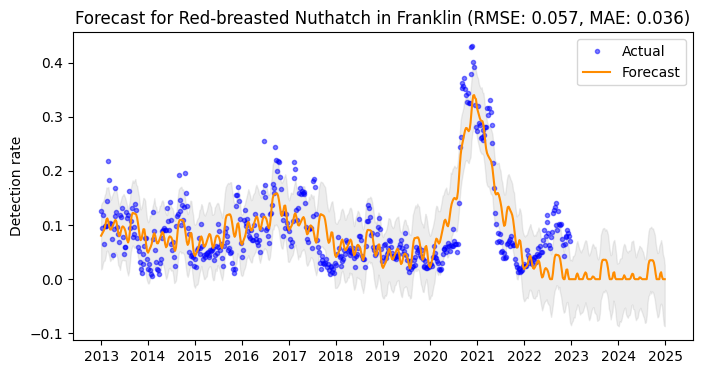
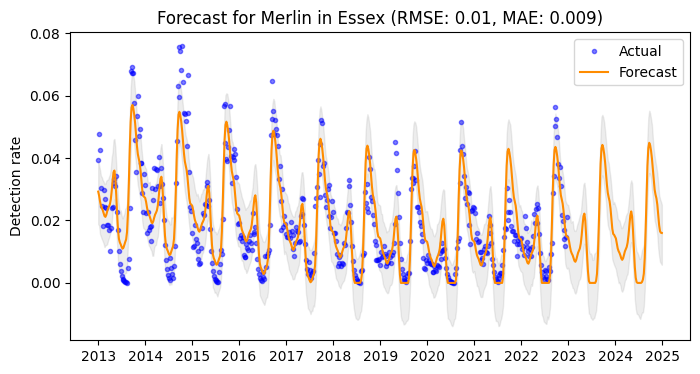
| **Metric** | **Min** | **Max** | **Median** |
| --- | --- | --- | --- |
| RMSE | 2.27e-9 | 0.197 | 0.007 |
| MAE | 1.36e-9 | 0.170 | 0.005 |

*Table. 2. Range of RMSE and MAE for all models*

The model demonstrated high predictive accuracy for migratory birds, such as warbler species (Yellow, Pine, Blackburnian Warbler etc.), winter ducks (Hooded Merganser, Ring-necked Duck, Common Eider, etc.), thrushes, orioles, and grosbeaks. This is likely attributed to their pronounced seasonal patterns. Models for resident, year-round birds like Blue Jay, Black-capped Chickadee, and White-throated Sparrow have higher RMSEs and wider confidence intervals, possibly due to their widespread presence and the resulting variable detection patterns, underreporting, and complex behaviors influenced by environmental and human factors. Overall, the model accurately captured trends and seasonality in detection rates, particularly when changes occurred gradually, as seen with the notable increase for the Red-breasted Nuthatch in Franklin in 2020-2021 (Fig. 4). However, it struggles with abrupt, short-term fluctuations, like the sharp and small rise in 2022 in the same figure. This is typical for generalized models, which may not predict atypical events well without prior similar trends. The model also performed well with species that are more difficult to detect, such as falcons and owls. Note that uncommon species with consistently low detection rates were excluded from the analysis, despite showing good model scores. Figures 1-3 below illustrate the model’s performance for a migratory bird (Baltimore Oriole), a resident bird (White-breasted Nuthatch), and an elusive bird (Merlin).

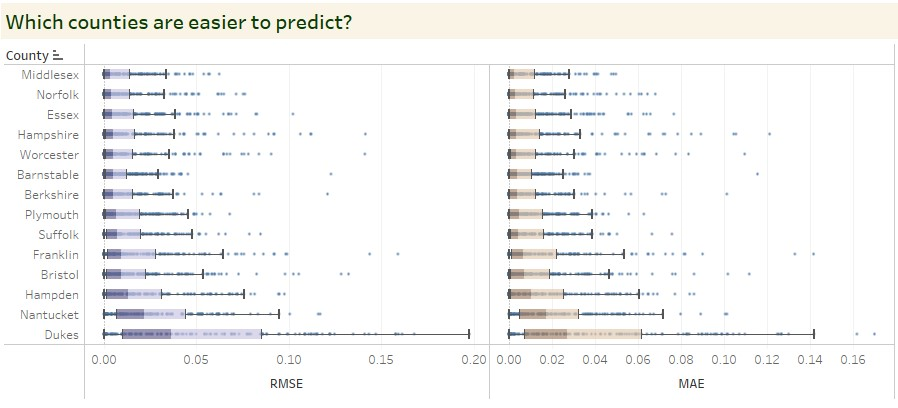


*Fig 1. Forecast for a migratory species Fig 2. Forecast for a year-round species*



*Fig 3. Forecast for an elusive species Fig 4. Forecast with changes in trend*

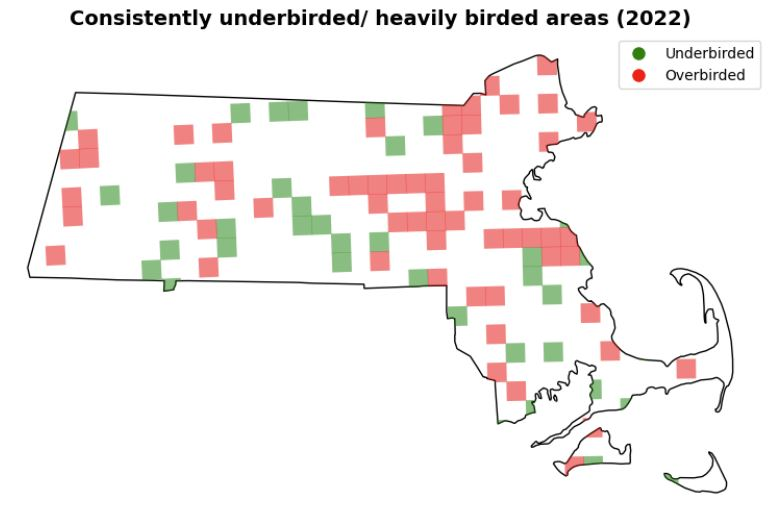
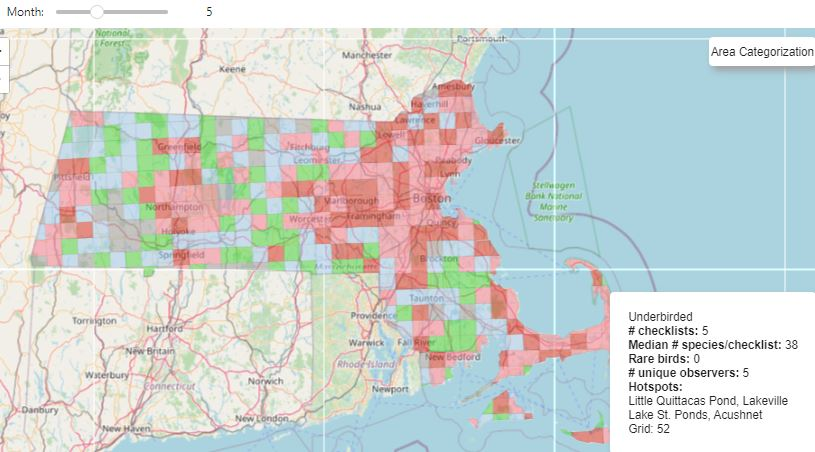
Model performance varied across different counties, as can be seen in Fig. 5. Densely populated counties like Middlesex, Norfolk, and Essex show lower median RMSE and MAE values with less variability, which may be influenced by the extensive conservation efforts and the high frequency of birdwatching activities in these regions. Barnstable county, with its coastal and marshland habitats, along with Hampshire and Berkshire, characterized by their natural landscapes and lower population densities, also exhibited strong model performance. In contrast, the islands of Nantucket and Dukes, which have limited data and unique bird populations, along with Hampden County's mixed urban-rural setting, introduced a greater degree of variability in bird detection rates.



*Fig. 5. RMSE and MAE for all models by county*

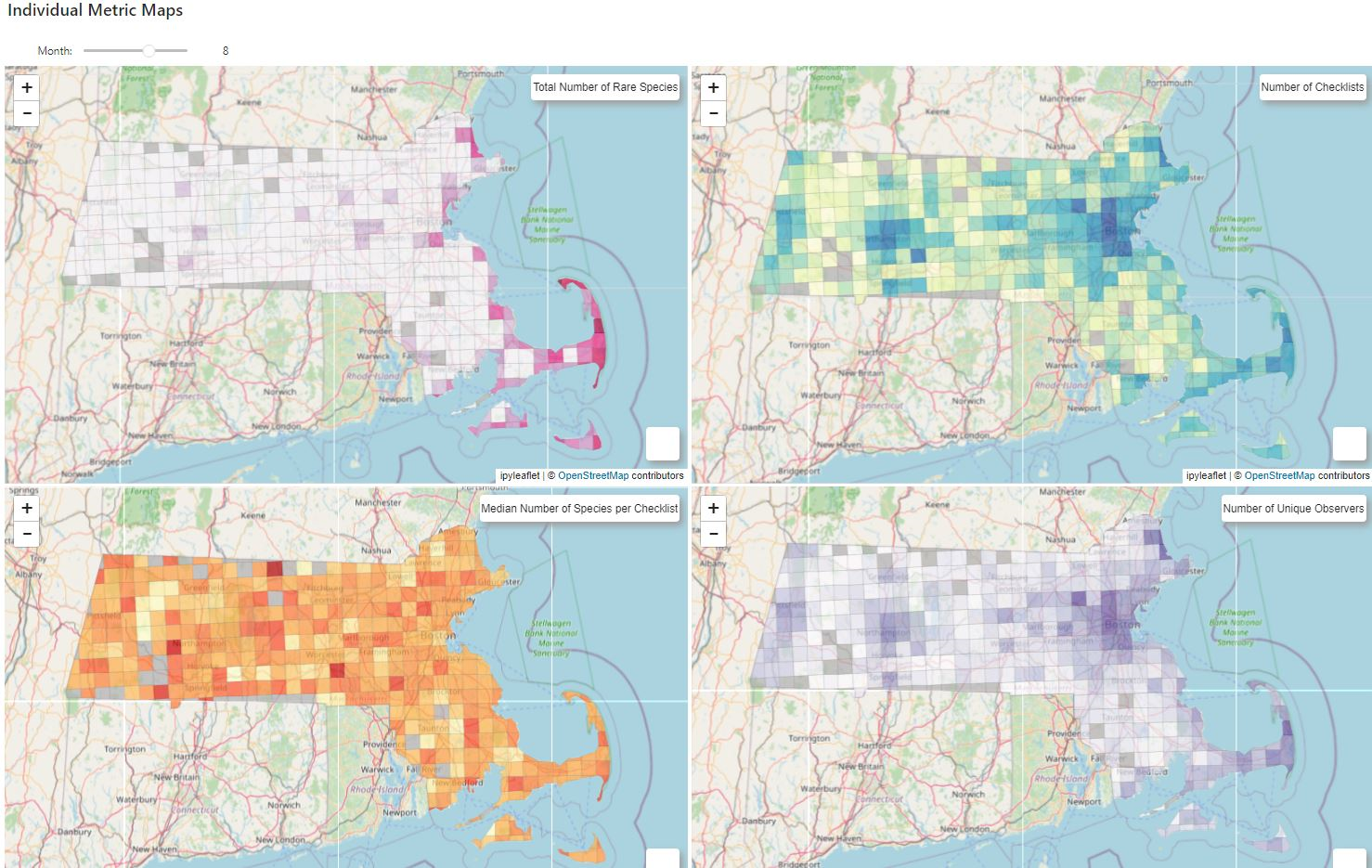
Finally, we updated the Tableau dashboard to display the forecast model results and descriptive analyses on timing and hotspot suggestions for all species. A user guide is included, and the interactive dashboard is accessible at <https://tabsoft.co/3uFSm5S>.

* 1. **Birding Activity Analysis**



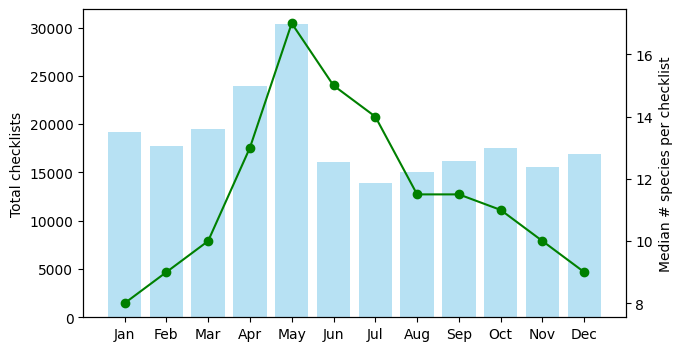
*Fig 6. Category map for May 2022 Fig 7. Frequent under/heavily birded areas*

Five interactive maps, each with 12 layers for 12 months, allowed us to visualize and categorize areas by birding activity. The maps are accessible at <https://bit.ly/3R7xhZH>, with a pdf view included in App. 4. The category map shows a concentration of popular and heavily birded areas in coastal and urban regions, as well as the Connecticut River Valley and the Berkshires during the summer (Fig. 6). Factors like accessibility and diverse habitats likely make these areas favorites among birdwatchers. Figure 7 identifies spots consistently categorized as underbirded or heavily birded for at least 6 months a year. The frequently popular spots suggest established birding communities and the need for careful management to mitigate potential habitat stress. Contrastingly, underbirded areas were more dispersed, revealing lesser-known yet promising birding locations in Southern Worcester during spring and fall, and Plymouth county in winter and summer. Comparing the area categorization maps with MA’s population density and topography maps (App. 5-6) reinforced our hypothesis: popular birdwatching areas are often in densely populated, lower-elevation regions, enhancing accessibility.

****

*Fig 8. Individual metric map for August 2022*

The maps displaying individual metrics uncovered additional insights (Fig. 8). The distribution of checklist submissions and unique observer counts remained predominantly high in urban and coastal areas throughout the year, aligning with the trends of popular and overbirded sites. However, species diversity, indicated by the median number of species per checklist, varied geographically and temporally. Interestingly, while coastal areas did not consistently report high species diversity, they maintained regular birdwatching activity, particularly during winter. Spearman correlation analysis (App. 7) showed a weak association between the number of observers (a measure of site popularity) and species diversity (coefficients from 0.08 to 0.28), but a stronger one with rare species sightings (coefficients from 0.34 to 0.61). This suggests that rare species more strongly influence birding activity than species diversity. Figure 9 displays monthly birding activity in MA, with the total checklists peaking in May during spring migration, when the median number of species observed per checklist is highest. Notably, while the number of species observed fluctuates, birdwatching activities are relatively stable throughout the year.

****

*Fig 9. Monthly total checklist & Median number of species per checklist*

1. **Discussion**

The project presents two user-friendly tools to enhance birdwatching experiences and inform habitat management and conservation efforts. The interactive Tableau dashboard provides accurate bird detection forecasts, enabling birdwatchers, from beginners to experts, to effectively plan their outings with detailed statistical insights. This tool can also aid habitat managers in monitoring bird populations and focusing conservation efforts on identified hotspots. The interactive choropleth maps provide a broader view of birding activities in MA, highlighting underbirded and heavily birded areas. This helps distribute birdwatching activities more evenly, reducing the environmental impact on heavily visited areas. The temporal and geographical analysis provided by these tools allows for in-depth research into bird migration patterns, species diversity, and the effects of urbanization on bird populations.

Building on this foundation, we are engaging with different birdwatching groups to collect user feedback, ensuring the tools are fine-tuned to meet their diverse needs. To enhance and expand the project, several opportunities are being explored. First, to scale the forecast model across states or countries, we can streamline it by clustering similar bird species and regions. Second, given the potential impact of environmental changes on bird detection rates, adding weather regressors remains a promising avenue to improve model performance. Future efforts will focus on sourcing and integrating granular weather data that aligns with the existing time series data.

Furthermore, by incorporating detailed data on various habitat types (forests, wetlands, urban areas etc.) and their conditions (vegetation density, land use etc.) the project can offer a more detailed perspective on bird populations. This enhancement not only promises more accurate predictions of species' detection rates but also enriches our understanding of birding activities in diverse habitats. Shifting from a grid-based layout to a focus on specific underbirded and overbirded hotspots, and adding hotspots accessibility information, will make the interactive maps more practical for birdwatchers and helpful for habitat managers in directing conservation efforts.

Currently limited by resources in data storage and hosting, we aim to improve the existing pipeline’s accessibility and automation. By leveraging cloud-based platforms like AWS or Google Cloud for efficient data hosting, and creating a Flask application to centrally host the dashboard and interactive maps, we can streamline the user experience and expand our project’s scope and reach.

1. **Statement of Contribution**

All members contributed equally to the project’s design and exploratory data analysis. Hajera and Michael took the lead in developing, refining, and finalizing the Prophet forecast model, effectively scaling it to all bird species. Neil spearheaded the integration of the weather regressors into the final Prophet model. Nhat was in charge of the birding activity analysis and incorporated the final forecast and descriptive results into the Tableau dashboard. Michael and Nhat collaborated on the model pipeline. Collective efforts were made in testing the products and compiling the final report.

1. **References**

Cornell Lab of Ornithology. (n.d.). *Trends maps*. eBird Science. Retrieved November 19, 2023, from https://science.ebird.org/en/status-and-trends/trends-maps

Cornell Lab of Ornithology. (2023). *Use eBird data and tools*. eBird Science. Retrieved November 3, 2023, from https://science.ebird.org/en/use-ebird-data

Floyd, T., Minardi, R., & Retter, M. L. (2023). *ABA Checklist*. American Birding Association. Retrieved December 6, 2023, from https://www.aba.org/aba-checklist/

*Greykite - A flexible, intuitive and fast forecasting library*. (n.d.). LinkedIn Open Source. Retrieved October 16, 2023, from https://linkedin.github.io/greykite/

Khare, P. (n.d.). *Understanding FB Prophet: A Time Series Forecasting Algorithm*. Medium. Retrieved November 3, 2023, from https://medium.com/illumination/understanding-fb-prophet-a-time-series-forecasting-algorithm-c998bc52ca10

NCEI. (2023). *County Time Series | Climate at a Glance*. National Centers for Environmental Information. Retrieved December 6, 2023, from https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/time-series

Trimble, J., & Schain, R. (n.d.). *Review List – Massachusetts Avian Records Committee*. Massachusetts Avian Records Committee. Retrieved December 6, 2023, from https://maavianrecords.com/review-list/

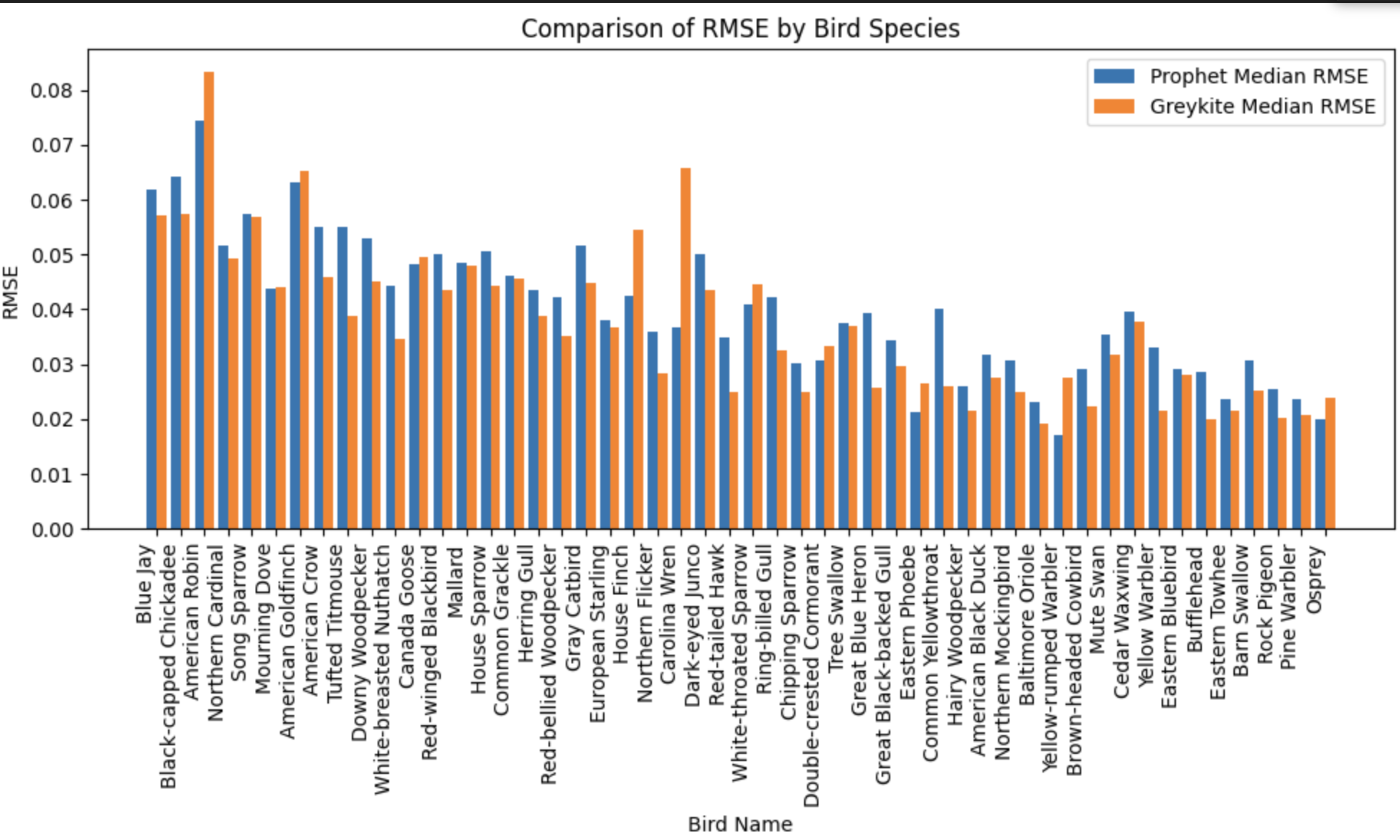
*12.2 Prophet model | Forecasting: Principles and Practice (3rd ed)*. (n.d.). OTexts. Retrieved October 16, 2023, from https://otexts.com/fpp3/prophet.html

1. **Appendix**

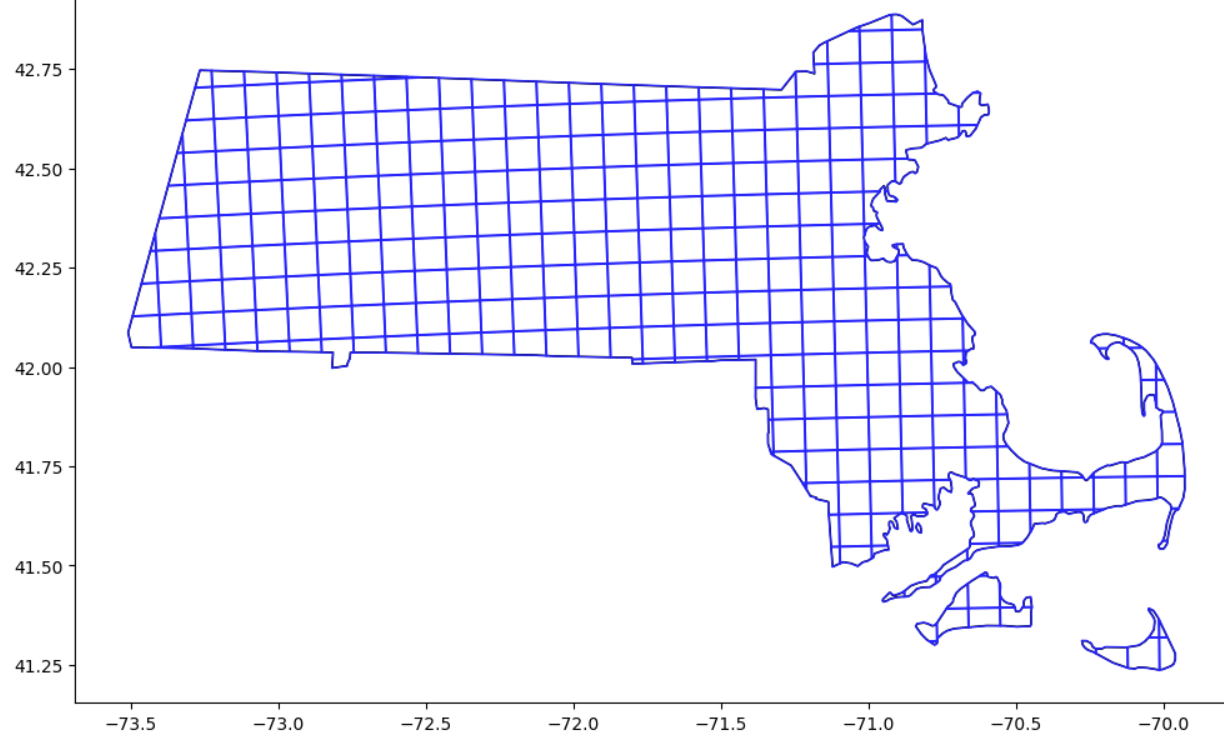
**App. 1**: [Final Report - Phase 1](https://docs.google.com/document/d/1RzNR8ONx-41TrBU4t1UDzu0jS8FRH-hdJwYDk8962AA/edit?usp=sharing)

**App. 2**: Results for fined-tuned Greykite and Prophet during initial testings on 50 species

|  | **Prophet** | **Greykite** |
| --- | --- | --- |
| Runtime (50 species) | 23.5 minutes | 60 minutes |

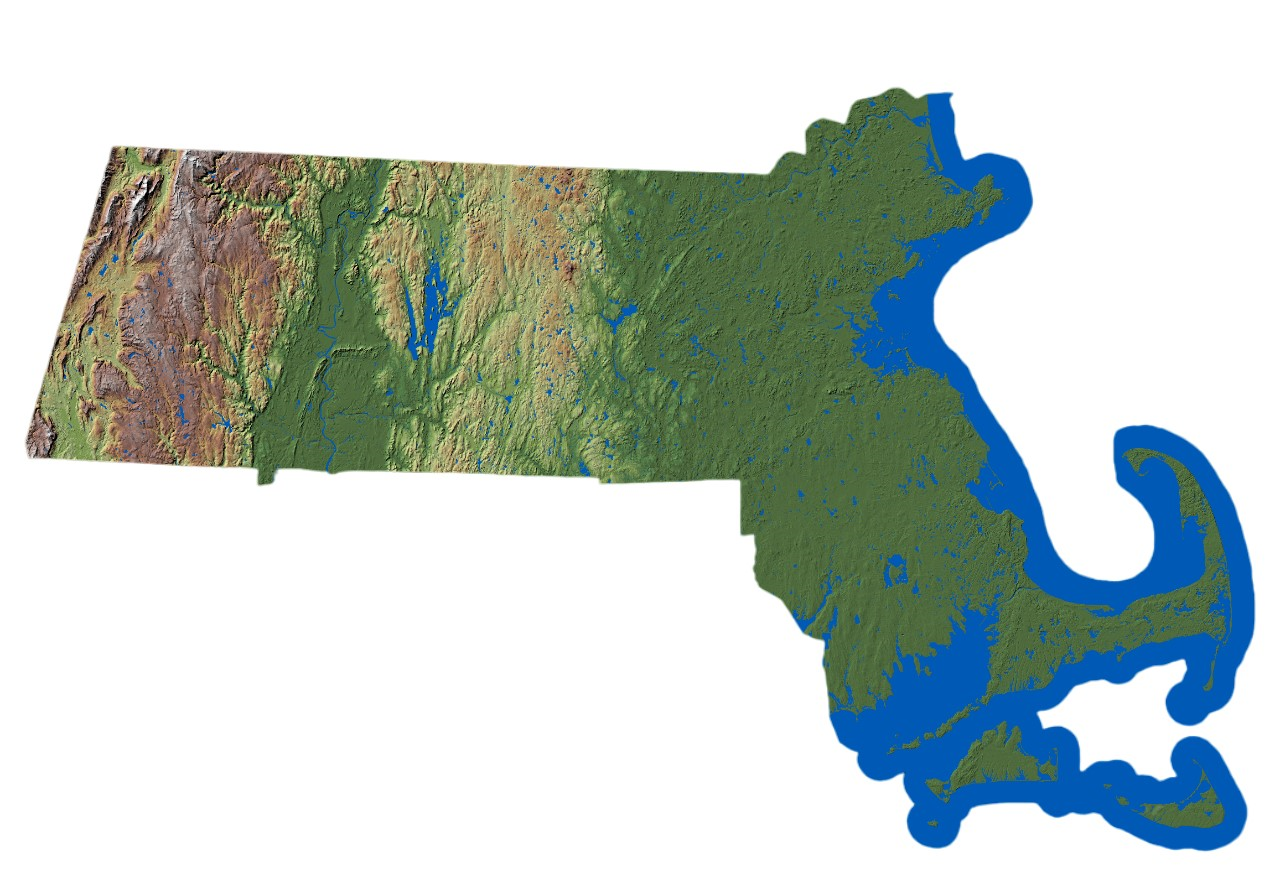


**App. 3**: MA map with 8km x 8km grid cells

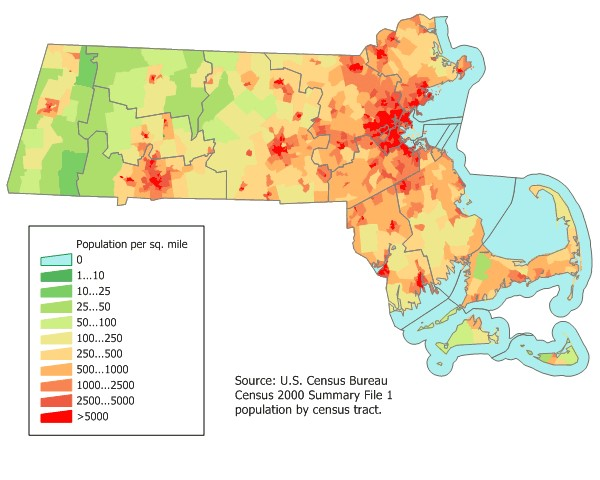


**App. 4**: [activityMaps.pdf](https://drive.google.com/file/d/1QqrkRx-laVf_OlGAKCKftw8pFZDUudef/view?usp=sharing) <https://bit.ly/3NbXNzM>

**App. 5**: National Elevation Dataset Shaded Relief of MA, US Geological Survey (https://eros.usgs.gov/media-gallery/shaded-relief/massachusetts)



**App. 6**: MA Population Density Map, U.S. Census Bureau



**App. 7**: Monthly Correlations: Unique Observers with Uncommon Species and Median Species Numbers

| **Month** | **Corr\_Rarebirds** | **Corr\_Numspecies** |
| --- | --- | --- |
| 1 | 0.506 | 0.228 |
| 2 | 0.462 | 0.206 |
| 3 | 0.481 | 0.273 |
| 4 | 0.37 | 0.238 |
| 5 | 0.36 | 0.223 |
| 6 | 0.345 | 0.132 |
| 7 | 0.435 | 0.084 |
| 8 | 0.428 | 0.233 |
| 9 | 0.545 | 0.172 |
| 10 | 0.5 | 0.282 |
| 11 | 0.479 | 0.283 |
| 12 | 0.608 | 0.236 |