Critical Thinking 7

10/30/2022

MIS581-1

Dr. Jamia Mills

Michael Dendinger

# Abstract

When trying to determine the likelihood of seeing birds for the activity of bird watching, it can be difficult to know where to go to best find birds. This research endeavored to find the best location to find where those who were interested in bird watching could find the most birds. The hypothesis of this study was to see if one was more likely to see birds in national parks as opposed to locations with more human intervention. It would seem logical that the birds would congregate in areas that have been protected for animal and natural ecosystems that have not been tainted by commercial or residential areas.

This research used the data collected by Cornell University in partnership with the Audubon Society to see where individuals had seen birds. The data was a collection surveys from individuals who reported seeing birds on given dates, what type of bird it was, and how long the bird was in that location. This data was then overlaid with map data of where national parks are located. The bird sighting data was then analyzed using hot spot analysis to find where birds are most often to reside.

The research found that most of the bird sightings that were reported were significantly skewed toward heavily populated city areas. The assumption for this is that there was just simply a greater number of individuals reporting seeing birds in these locations. This makes it so that the birds may not be more plentiful in those locations, just those who report seeing them.

# Introduction

Birding and bird watching has seen a boom in activity since the pandemic and has been reaching new levels of interest over the last decade. With the media attention from the 2011 film The Big Year, starring Steve Martin, Owen Wilson, and Jack Black, birding was brought into a new light and sparked the interest of thousands of new birders. In 2020 some birding businesses reported a growth of almost 80% in sales of bird related products (Dahanesha, 2020). With the growth in birding in our backyards, individuals are now able to leave their homes and continue their birding adventures in the great outdoors. It can be tricky entering the moderately competitive world of birding. It is sometimes tricky knowing where different birds prefer to fly through during migratory seasons and what types of environments are best suited to them. In my proposal, I will create a data set that will connect bird migratory patterns with the locations of different national parks so that individuals will be able to visit our nation’s great treasures as well as learn more about our ornithological diversity.

## The Problem

In the years leading up to the pandemic, national parks in the United States had an average of 323 million visitors a year (National Parks Service, 2022). In 2021, the numbers were down to 297 million visitors. While the decline in attendance was impacted by the pandemic, there is also a concern that it will be hard to get individuals to come to national parks as the experience of attendees seems to be changing in a perceived negative way (National Geographic, 2010). One complaint is that many of the individuals coming to the parks are looking for higher “excitement” types of activities like snowmobiling or four wheeling through these locations. Others describe the issue of attendance as being from individuals who are more affluent and have more money and time to visit these parks (Rott, 2016). All in all, the issue is one of perceived value of national parks. What are they for? Who are they for? Why should we care? The NPS Mission is stated as: “[To] preserve unimpaired the natural and cultural resources and values of the national park system for the enjoyment, education, and inspiration of this and future generations” (Department of the Interior, 2012).

As aligned with the NPS mission statement, the goal of the following research is to create a tool for individuals to better participate in enjoyment and education within the national parks as related to the activity of bird watching. While it may not be the most lively of activities, the requirements for bird watching are patience and a love for learning. One big barrier is the time it takes for many individuals to learn about where they can go to view some of our nation’s beautiful birds and when they will be in local areas. With data from the NPS, EPA, and the Cornell eBird data sets, we will be able to make a system for individuals to learn about birds within national parks near them so that they may learn more about their regional ecology and biology.

Working the NPS into the ever-growing activity of bird watching in a way that encourages individuals new to the activity will allow for a greater exposure of the beauty that is offered at the national parks. While the National Audubon Society and the National Parks have worked together in the past for conservation, this will be a new step for the two to share data across one another to encourage individuals to visit the parks.

# Questions

When it comes to migratory bird data, one can ask any multitudes of questions. This research will focus on the migratory patterns of birds as related to national parks. This means that the questions within this research will focus on how to optimize bird viewing through the locations where there are national parks within the United States.

1. What percentage of migratory birds in the United States can be seen by just going to national parks?
2. When is the best time of year to see the widest variety of birds at a national park by different ecological system?
3. Which national park has the highest diversity of migratory birds?

All of these questions relate back to the core issue discussed in the initial problem finding mission of this course. As the National Parks System is trying to rebuild after several years of COVID, the parks need activities that will bring people into their parks. With these questions, one will be able to see which park has the birds they would most like to see as well as how they can see them during the year.

# Hypothesis

For this research there will only be one hypothesis based on the research that will also answer the above questions.

## Ha

One can view a higher rate of diverse birds through the national parks than by traveling to non-federal lands.

## Ho

There are no particular locations that are more advantageous to seeing a diverse number of birds on a statistically significant level.

# Literature Review

Mapping seabird sensitivity to offshore wind farms

The first article to be used as a reference for this paper is about how seabirds are sensitive to wind farms. This article is rather self-explanatory as it investigates the areas around offshore windfarms and how bird migratory patterns are impacted by them. This article will support the thesis of my research paper by discussing how birds interact with areas where there is significant amount of unmonitored manmade objects away from residential or commercial areas. As most national parks are explicitly protected from having any large structures like this on them it will be simple to make a comparison to how the migratory birds change patterns to move to locations where there is less human intervention in nature. This paper discusses how seabirds are impact by these structures specifically, but it may be also applicable to different types of birds that are reflected in my research.

## Influence of individual biological traits on GPS fix-loss errors in wild bird tracking

This paper discusses the importance of different kinds of tracking methods for migratory birds beyond GPS location errors. As it is fairly obvious to most individuals, just using survey data on where individuals have viewed certain birds in this research seems limited, the need to show flaws with individual GPS locations as well. This research mostly focuses on raptor migratory patterns as they were the birds large enough to have the type of GPS trackers necessary for this kind of research. This also means that the birds studied in this paper mostly focus on birds that spend a majority of their time actually flying rather than on the ground. This makes some of the research important to my research as raptors may not be as likely to be catalogued in the eBird data set since it can be harder to identify a bird in flight as compared to a stationary bird. So, having the information as related to how using individual reports can be beneficial adds credence to the research of my paper and validity to the process for measurement.

## Mapping Migratory Wading Bird Feeding Habitats using Satellite Imagery and Field Data, Eighty-Mile Beach, Western Australia

This paper is similar to the research done as related to GPS coordinates. The larger piece of information in this research is how the artificial intelligence of image analysis in collaboration with individual reports should go hand-in-hand to provide a more complete picture of how different species of birds migrate to different locations. While it would be a pleasure to include image analysis of birds into my research, the tools required for this are beyond my abilities for this research paper. However, the discussions of how to integrate field data reports into the study of migratory birds is very important. This paper does focus specifically on a singular beach area. So, if my research where to be narrowed down to a specific park or given location, this type of research would be very simple to apply. There are some nuances like sand being a dramatic color difference to most birds studied in the images, but none-the-less, the lessons from this research will be valuable when applying the counting methods to my research paper.

## Different location sampling frequencies by satellite tags yield different estimates of migration performance: pooling data requires a common protocol

This last paper to be used as a reference is how sampling frequencies are important with how data is collected. As the eBird data being used is a constantly filling database of individuals reporting their sightings of birds, it is important to remember the importance of sampling. Since all these sightings are important to calculate, the research needs to be as objective as possible to remove potential outliers of research. The ability to tag certain sightings within a GIS software will allow for a greater sense of where certain areas may have a higher rate of reported bird sightings. As low population areas may not have individuals reporting bird sightings, the importance of tagging where the sightings *do* exist is that much more important within a geographic sense.

# Methodology

This research will use quantitative methods to identify the locations of the birds and the highest likelihood of seeing them in different areas. The data as related to this project will be a mixed methods approach. It is defined as mixed methods due to the fact that it uses sightings form individuals that could inherently subjective as it does not appear that anyone is 100% perfect at identifying birds. The quantitative data for this is the locations of the parks as well as ecosystem maps that will be used to perform the analysis for this project.

Once the data has been cleaned, the process will then use a strictly qualitative methodology to perform the analysis required to identify the areas where the birds reside most often as related to national parks.

# Research Methods

For this research, three datasets will be used to discover whether the hypothesis is true. The tree data sets have been described above. The process to perform this research will have to be very delicate as the eBird data set that contains the migratory bird sighting records is quite large and will require finesse in the analysis to make sure that all data is correctly aggregated.

## GIS Location Aggregation

What is immediately necessary for the aggregation of this data is to consolidate the large eBird data set down into a more simplified dataset so that it can be consumed by traditional data analytic methods. The data set will be broken down using the Python library Pandas to break down the large data set into smaller chunks where we can rationally perform our analytics on a looping function. This will be to take the roughly 800 million results and lay them over a map of all counties within the United States. Then a method referred to as “Point in Polygon” will be used to see how many birds and how many types of birds exist within each county’s border. The same action will be performed with the polygon data as related to the national parks.

The research will be using counties as the core geographic polygon in this research as the average size of county in the United States is 2,915 square kilometers and the average national park is 3,350 square kilometers (National Parks Service, 2018). This will allow for a better comparison between the counties and the national parks as the areas are similar enough and have a fairly similar range of different sizes that will make the aggregations within a reasonable comparison for analysis. They even have a similar distribution of sizes, with the smallest county in the United States as 0.03 square kilometers and the largest county at 51,000 square kilometers (U.S. Census Bureau, 2022). The smallest national park is 0.04 square kilometers, and the largest national park is 35,000 square kilometers (National Parks Service, 2018). With these numbers, we can use similar sized polygons on average across the United States to find rational comparisons in land size when discussing the likelihood of a bird appearing in different land areas.

## Polygon Overlap Analysis

One area of concern with the analysis described above is that there are counties that reside within national parks. To help eliminate this issue, once the counties have their counts recognized, the overlapping areas of counties and national parks will be determined. From this point the counties that have some overlap will be cut to only represent the land that is outside of a national park and the percentage of remaining land will be divided by the total number of birds seen in the county. So, if a county is 50% covered by a national park and 100 birds were initially sighted in that count, it will be divided by .5, resulting in 50 bird sightings in that given county. This will not be performed in the opposite direction. As there will be counts for sightings within national parks in a more accurate measure, there will not be a need for such analysis to be performed.

# K-Means Clustering

Once the counties have their counts of sightings and diversity aggregated, they will be clustered together using K-Means clustering. This will show which counties have similar rates of bird sightings and diversity outside of national parks. This will provide a group of counties that have higher rates of bird sightings as well as higher levels of bird diversity. Through this we will be able to compare those counties with the national parks. This process is reliable as it has been used to perform analysis on other types of patterns like weather forecasting (Xu, 2021). Since the birds move in patterns as related to the ecological systems, the use of K-Means clustering to find similar bird movement patterns will be useful and accurate.

# Limitations

As one can assume, there are quite a few limitations within this project. The data is one of the largest limitations as well as the computing process for this data. Both aspects do have processes that can be implemented to reduce the limitations of the project, however. Hopefully they will shed enough light on how this project plans to mitigate those issues.

## Data Limitations

As stated earlier, the data for this project as related to the sighting of the migratory birds comes down to individuals reporting the locations of migratory birds. In so many words, this means that someone has to have seen the bird for it to have existed there. This means that if a bird is in a location where no one can see it, it will not be counted (Wade, 2008). It also does note account for individual birds in an area being counted twice. As many birds travel throughout the day and night to hunt, it is very possible that individuals may double count certain birds as they are performing their daily functions and could throw off the data.

## Data Limitations – Remediation

To deal with these issues, the analysis will make sure to use sampling methods described in migratory bird research studies from other sources to complete the analysis that is required for this project. Most of these practices will use sample methods to form an expected value for the sightings and also remove sightings that occur of the same type of bird within too short of a period of time. There will also be an aspect of this research that will strictly look at the diversity of species sighted and perform analysis that way. With this piece in place, even if one bird is counted one thousand times, it will only count as one species sighted in that specific location as related to the research.

This process will work in the reverse as related to not sighting birds in certain areas where people are not located. The sampling method that will be used can be applied to neighboring geographic areas to determine the likelihood of the bird being in that given location during a period of time to determine whether or not a sighting in that area is possible (Tanferna, 2012). This may not be an ideal choice for a quantitative analytical process, but with the wealth of data that is available to us it will be a useful approach.

## Data Processing Limitations

The data set for the bird sightings is 292 GB in size and contains over 700 million rows of data. This means that this project is using big data. Big data is a problem for this kind of research as there will not be use of cloud or enterprise level computing systems. For a local system running on a solid-state drive and 32 GBs of memory, the process of simply looping through all of the rows of data took over 10 hours. This means that performing complicated calculations in an iterative or batched process will take extensive time.

## Data Processing Limitations – Remediation

To adapt to the size of the data, several processes will be adapted to fit the needs of the given process. As described above with the sampling method being used as well as the implementation of the clustering analysis, the quantity of data will be able to decline in size as it is related to the project.

The data cleaning process will use several Python libraries and extensions to make this task more efficient and accurate. Specifically, the use of the Dask library in Python that allows for large, rectangular data sets to be analyzed quickly and then produce and output in a manageable format will be used during the initial stages of the data cleaning process. Once the data has been aggregated into the correct shape and format, Python’s PySpark will be used to perform some of the higher level analytical processes required for this project. PySpark is a library specifically designed to take advantage of all of the processing capabilities of a standalone computer system and will be able to perform many calculations in parallel so that the process may be performed at a higher rate (Shanahan, 2017).

# Ethical Considerations

One thing to be aware of with this analysis is the potential for individuals to find endangered birds for poaching of harmful practices. This is already taken care of through the eBird data set as they do not share migratory data on those species to protect them. This does mean that it may be harder for individuals to sight rare birds, but if it is done within the framework to protect endangered species, it is worth it.

It is also important to note that this form of geographic comparisons could have the potential to encourage trespassers to encroach on territory that they do not have permission to enter. The aggregation of the bird data to the county level will help eliminate the temptation of such individuals to go where they are not wanted or welcome.

# Analysis

To perform the analysis for this research, the data was collected into functional chunks that of bird sighting data across the United States. The data consisted of over 700 million bird sightings covering the last 50 years of bird sightings. As this research is mostly focusing on how birders can currently view birds, the data was brought down to the last five years of available data (2016-2021). This brought our data count down to 12 million recorded bird sightings. Using the Dask process listed above, the bird sightings were output into chunks of data that were related to specific counties and territories for the analysis.

Once this process had been performed, we needed to identify the difference between standard counties and national parks. To do this, we created a GIS shapefile that removed the counties that overlapped with national parks and then replaced the empty space with the polygon and data that represented the national parks. The map that came out of this process looked like this:

Map

Description automatically generated

ArcGIS Pro Polygon Map (2022)

In this map, one can see the national parks as the territories highlighted with a dark black border. From here, the analysis to find how many sightings occurred within each polygon was performed. The output for identifying polygons that had the highest rate of bird sightings was this:

Map

Description automatically generated

ArcGIS Pro Polygon Choropleth Map (2022)

From general inference, one does not necessarily notice any areas with high trending bird sightings as related to the location of national parks. However, it can sometimes be difficult to make this type of analysis just by looking at an image like above. With this in mind, the next step of the research used the Neural Network approach to identify hotspots of bird sightings.

For this, we needed a slightly different shape of our data. To perform this, we loaded the 12 million bird location records to see where hotspots existed within the given areas of the map. This process was run in the ArcGIS Pro tool for identifying hot spots. The output of this map appeared as below:

Map

Description automatically generated

ArcGIS Heatmap with National Parks Polygons (2022)

The red squares represent the hotspots designated by the Neural Network to identify areas where one is most capable to report the sighting of birds. The green polygons represent all the national parks within the United States.

## Conclusion

According to the analysis, there is no proof to show that national parks a better location to view birds. This analysis does not seem to have survived the issues brought up in the limitations section of the research. As stated in that section, this data is about where bird sightings are reported. This means that there was inevitably going to be a skew to areas where people who are already passionate about birding would sight birds. This does not represent a negative correlation between sighting birds and being in a national park.

Another thing to be aware of is the diversity of birds sighted in this data. Of all of the sightings, 6% of all sightings were of three birds: The American Robin, the Morning Dove, and the American Crow. These three birds are extremely common around the United States and would not require a visit to a national park to see in any manner. The next step of data such as this would be to identify which birds are rarest in given ecosystems and then identifying where they are most commonly seen. This may result in places more associated with national/state parks. Which brings the research to the next point.

As national parks are not the only places away from civilization for people, they are not going to always be accessible to all individuals. It would be advantageous to incorporate state or even local parks into the research as well. This would make the process of finding rare birds in local areas more accessible as well.

Finally, one piece of data that would be useful would be tracking data from bird researchers. This would be fairly tricky to acquire as the labor to do this form of research is extensive and mostly focuses on singular breeds of bird. To look at this data across all birds would be extensive and unrealistic.

# Bibliography

Dahanesha, N. (2020, 8 6). *Birdwatching Is a Bright Spot in a Pandemic-Stricken Economy*. Retrieved from The National Audobon Society: https://www.audubon.org/news/birdwatching-bright-spot-pandemic-stricken-economy

Department of the Interior. (2012). *Contingency Plan.*

National Geographic. (2010, 5 26). *Top 10 Issues Facing National Parks*. Retrieved from National Geographic: https://www.nationalgeographic.com/travel/article/national-parks-issues

National Parks Service. (2018). *The National Parks Index: 2012-2016.* Washington DC.

National Parks Service. (2022). *Visitation Numbers.* USNPS.

National Parks Services. (n.d.). *Theodore Roosevelt Quotes*. Retrieved from NPS: https://www.nps.gov/thro/learn/historyculture/theodore-roosevelt-quotes.htm

Oracle. (2022). *What is Big Data*. Retrieved from Oracle Big Data: https://www.oracle.com/big-data/what-is-big-data/

Rott, N. (2016, 3 9). *Don't Care About National Parks? The Park Service Needs You To*. Retrieved from National Public Radio: https://www.npr.org/2016/03/09/463851006/dont-care-about-national-parks-the-park-service-needs-you-to

Shanahan, J. (2017). Large Scale Distributed Data Science from scratch using Apache Spark 2.0. *Proceedings of the 26th International Conference on World Wide Web Companion*, (pp. 955-957). San Francisco.

Sullivan, B. C. (2009). eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation*.

Tanferna, A. (2012). Different location sampling frequencies by satellite tags yield different estimates of migration performance: pooling data requires a common protocol. *PloS one [PLoS One]*, e49659.

U.S. Census Bureau. (2022). *Census QuickFacts.* Washington DC.

Wade, S. (2008). Mapping Migratory Wading Bird Feeding Habitats using Satellite Imagery and Field Data, Eighty-Mile Beach, Western Australia. *Journal of Coastal Research*, 683-770.

Xu, F. (2021). A Unifying Framework for Analysis of Spatial-Temporal Event Sequence Similarity and Its Applications. *ISPRS International Journal of Geo-Information*, 594.

Zaki, A. (2022). An object-based image analysis in QGIS for image classification and assessment of coastal spatial planning. *The Egyptian Journal of Remote Sensing and Space Sciences* , 349-359.