Bias-free notion of bird rarity using statistical clustering. NURA Fall 2021

Jimson Huang[1], Dr. Paul Hurtado[2]

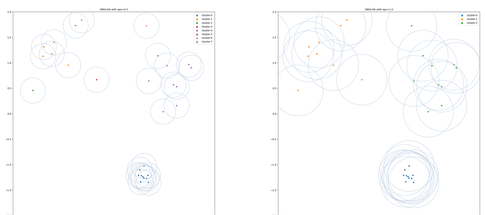
[1]Mathematics (Statistics Focus) - University of Nevada, Reno (UNR); [2]Faculty, Department of Mathematics Statistics - University of Nevada, Reno (UNR)

Abstract

eBird is a large-scale citizen science project for monitoring bird populations globally using observer-reported checklists of bird observations. One source of bias in such data is the over-reporting of regionally rare birds, which attract a lot of attention from observers. Bird watchers have a different notion of rarity from other biological definitions, which reflects the likelihood of encountering that species at a specific site or in a given region. We limit our scope to the state of Nevada, but the technique being developed should be applicable to any region with eBird data. This project helps correct for over-reporting bias when analyzing eBird data by using a DBSCAN (Density-Based Spatial Clustering of Applications with Noise) statistical clustering algorithm to cluster bird observations by location and time. The two parameters of this clustering technique are the spatial radius of a cluster and the temporal size of each cluster. A goal of this project is to figure out a way to find these two parameters for each rare bird species observed in a way that best reduces the impact of observation biases, resulting in a rarity ranking of each species that more accurately aligns with the opinions of birding experts. In a later stage of the project, birding experts will be consulted for their opinion on the rarity of various species, to better assess the outcome of the generated ranks. Preliminary results indicate that changing the parameters of the statistical clustering do significantly affect the resultant ranking of rarer bird species, which shows as a potential proof for the significance of the impact of observation bias in the original eBird data.

Methodology

Density-based Spatial Clustering of Applications with Noise (DBSCAN) is a statistical clustering method which forms clustering by grouping together points that are within a certain Euclidean distance from one another. This distance parameter is referred to as eps. The noise aspect of DBSCAN with its associated minimum number of points parameter is not used for this project.

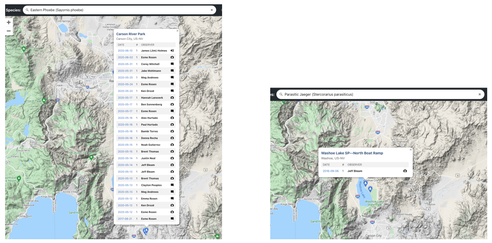


(a) 7 clusters found (b) 3 clusters found

Figure: Example of DBSCAN clustering on the same set of data points with different eps parameters.

eBird

eBird is a citizen-science project which gathers data on observations of birds across the globe though checklists submitted by birdwathcers. This data – arguably the largest avian data set in the world – forms the basis of widely used educational resources [1], and is also is used by scientists worldwide to study various aspects of avian biology (e.g., in 2020 over 90 peer reviewed publications used eBird data) [2]. Because the data is self-reported, certain species which attract the attention of bird watchers appear much more often in eBird checklists. Below are examples of one species which appear in many checklists vs one that appears in few.



(a) Eastern Phoebe (b) Parasitic Jaeger

eBird includes a rarity ranking on their website, with a species's rarity rank determined by the percentage of checklists that include such species.

Figure: eBird's own ranking system.

However, because of the over-reporting bias mentioned above, the number of checklist which contains a bird does not necessarily reflect a bird's rarity, but rather has a lot to do with a bird's popularity amongst birdwatchers. This project aims to use DBSCAN clustering to reduce the impact of over-reporting and thus generate rankings for birds more similar to the opinion of birding experts.

New Ranks

Table 1 shows the unclustered eBird ranking, DBSCAN clustered ranking using eps=10 miles and days/eps=7 days, and the ranking created by Dr. Hurtado for six species of birds.

Species	Unclustered	Clustered	Hurtado
Lesser Black-backed Gull	1	2	1
Gray Catbird	2	1	2
Brown Thrasher	3	5	5
Sabines Gull	4	4	4
Mew Gull	5	3	3
Pomarine Jaeger	6	6	6

Table: Bigger number means rarer.

Comparative Rarity Graphs

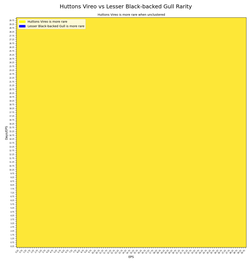


Figure: An example of two species whose comparative rarities are unaffected by clustering.

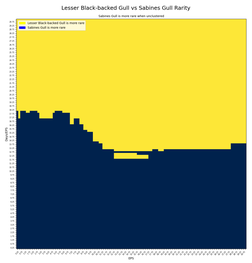


Figure: An example of two species whose comparative rarities are determined based solely on temporal clustering.

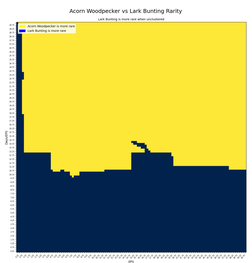


Figure: An example of two species whose comparative rarities are determined by both spatial and temporal clustering.

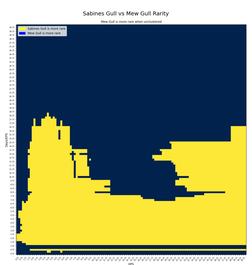


Figure: Rare examples of two species whose comparative rarities switch around in a chaotic pattern.

etc.

References