Keystone Species of the United States

Introduction

In practical terms, an ecosystem can be thought of as a network. Most elements of an ecosystem, whether they are flora or fauna, are directly or indirectly dependent on one another for sustenance. With the ever-looming threat of climate change and biodiversity loss that the planet is facing in modern times, it is now more important than ever to understand the influence that species have on their respective ecosystem. The ability to identify keystone species in a community could be critical to preserve biodiversity.

Protecting ecosystems is something that is important to me, especially those that exist in my home country and state. For this reason, I decided to take a data science approach to solve this issue by creating a BigCLAM community affiliation model from an undirected graph species network that identifies keystone species. This type of modeling can help various industries understand the full implications of harming a species, and hopefully act as a measure of persuasion against such action. While the models in this project specifically pertain to species and ecosystems in the United States, the methods of this project can be used to identify keystone species on a global scale, if data is accessible.

Data

For this project, I used metadata of North American wildlife camera trap images obtained from a data repository called the Labeled Information Library of Alexandria. This data has 3,382,215 objects. The data has 14 features: seq_no, id, filename, study, location, width, height, category_id, name, genus, family, order, class, common_name. Only the id and common_name features were used in the analysis.

Preprocessing/Exploratory Analysis

To preprocess the data, the data was loaded into a pandas data frame and any unnecessary rows were removed. First, all rows that had empty camera trap images were eliminated. After that, any rows where the value held in the common_name column equaled vehicle were removed. Finally, any rows where the value held in the common_name column of the data corresponded to domestic animals, specifically values of domestic dog, domestic cow, horse, and donkey were removed.

The camera trap name was extracted from every row of the data frame. This was done by taking the value held in the id column and parsing out the jpeg tag. For camera traps located in Colorado, the camera trap name was obtained by getting the substring that occurred before the last underscore of the string. For camera traps located in California and Florida, the camera trap name was obtained by getting the substring that occurred before the first underscore of the string. I stored these camera trap names sequentially in a list and then made a new column in my data frame called camera_traps to store the values.

After preprocessing, 610 unique camera traps and 44 unique animals were available for analysis.

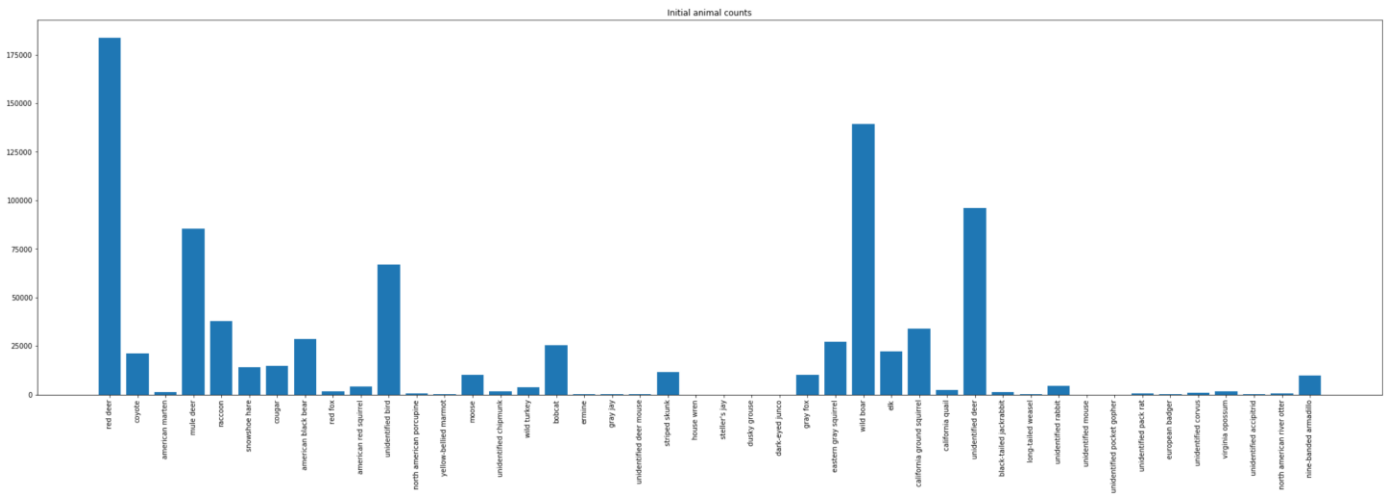

Figure 0: initial animal counts

Background research

After preprocessing, keystone species were predetermined by doing manual research. Out of the 44 unique animals, 7 species were classified as keystone. The keystone species that were found within my data were American black bear (ECOLOGICAL ROLE OF BEARS), cougar (Survey Mountain Lion Populations – 2019), snowshoe hare (NETN Species Spotlight), North American river otter (North American River Otter), bobcat (Bobcat.), coyote (Ortega), and American red squirrel (Posthumus). These 7 species were used as a baseline set to test how effective the BigCLAM model was at identifying keystone species.

Methods

The Apriori algorithm was used to find frequent pairs of animals in the data as well as support and confidence measures of the keystone species. All unique camera traps were used as baskets for the algorithm and all unique animals for a given camera trap were the content of each basket. Using a support threshold of 18.3 (3% of the total baskets), the counts of the unique animals in the dataset were determined to find which ones were frequent. After this, a lookup dictionary was created so the index of animals in the frequent animals list could be found. Then a list of triples was created, which contained a pair combination of every frequent animal as well as a count that was initialized as 0. All the baskets were looped through, and all frequent pairs of items were determined. The count of each frequent pair was found by finding it's index in the full upper-triangular array. The count of the triple at that index was then incremented by 1.

$$k = i \cdot \left(n - \frac{i+1}{2}\right) + j - i - 1$$

Figure 1: index of full upper-triangular matrix

Then, every pair of frequent animals that did not meet the support threshold were filtered out to obtain only frequent pairs of frequent animals. The support association rule of Apriori was used

to find out which of the frequent predetermined keystone species was most prevalent in the dataset.

$$support(X) = \frac{Number\ of\ baskets\ containing\ X}{Total\ number\ of\ baskets}$$

Figure 2: support association rule

To create an undirected graph species network, an empty affiliation matrix that used the number of frequent animals as the row and column dimensions was created. The list of frequent pair triples obtained from the Apriori step was used to fill the affiliation matrix.

$$mat[trip_0, trip_1], mat[trip_1, trip_0] = triple_2$$

Figure 3: affiliation matrix

To increase the level of sparsity in the matrix, every *[i,j]* and *[j,i]* elements within the matrix with a count less than 22 were set to 0 so the corresponding frequent pairs would not share an edge in the undirected graph species network.

The BigCLAM algorithm was used on the undirected graph species network to model community affiliations of animals. The frequent predetermined keystones species obtained from the Apriori step were ranked by their mean confidence scores and the top 3 ranked frequent keystones species from this measure were used as the centers of the BigCLAM algorithm.

$$confidence(X \rightarrow Y) = \frac{Number\ of\ baskets\ containing\ X\ and\ Y}{Number\ of\ baskets\ containing\ X}$$

Figure 4: confidence association rule

$$mean\ confidence(X) = \frac{\sum_i^{len(X_{afiliations})} confidence(X \rightarrow Y_i)}{len(X_{afiliations})}$$

Figure 5: mean confidence

A matrix *F* was created with 4 as the column dimension and the number of frequent animals as the row dimension. Matrix *F* was initialized by setting the values in the first three columns in each row to 1 and the fourth column with a background probability.

$$p = \sqrt{log\ 0.99 * -1}$$

Figure 6: background probability

Before running the algorithm, the step size was initialized to 10^{-5}. The number of iterations for the BigCLAM algorithm was set at 50,000. In each iteration, all nodes in the network were looped through. For each node, the gradient was initialized as an array of 4 zeros. The neighbors of each node were looped through, and the gradient was calculated.

$$\nabla l(F_u) = \sum_{v \in N(u)} F_v \left(\frac{1}{1 - \exp(-F_u \cdot F_v)} \right) + F_u - \sum_v F_v$$

Figure 7: gradient calculation

The row of matrix *F* corresponding to the node was then updated by adding the gradient. Any negative values in the row corresponding to the node were set to 0.

The Jaccard similarity between the top 7 most affiliated species in each community and the 7 predetermined keystone species was computed to measure how effective each community in the BigCLAM model was at identifying keystone species.

$$sim(A, B) = \frac{A \cap B}{A \cup B}$$

Figure 8: Jaccard Similarity

Results

The Apriori algorithm found 31 frequent animals and 210 frequent animal pairs. All 7 of the predetermined keystone species were frequent. The top 3 frequent keystone species ranked by mean confidence were North American river otter, American red squirrel, and cougar. The top 3 frequent keystone species ranked by support were coyote, snowshoe hare, and American black bear. The resulting rankings were displayed using the Matplotlib python library.

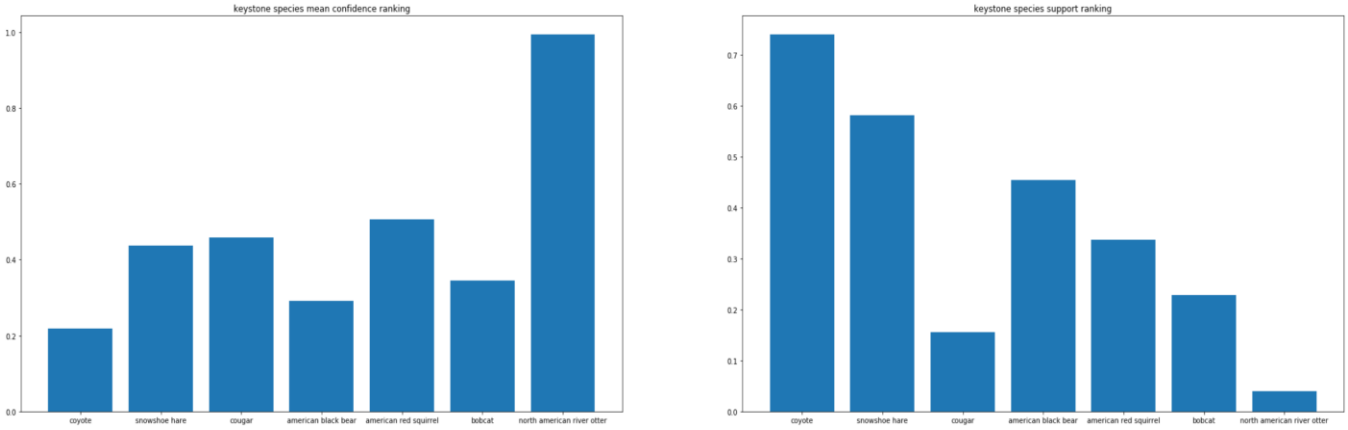

Figure 9: ranking of frequent keystone species by mean confidence and support

The final matrix that was used to create the undirected graph species network was 58% sparse.

The top 3 species affiliated with community 1 (initialized with North American river otter) were racoon, wild boar, and unidentified deer. The top 3 species affiliated with community 2 (initialized with American red squirrel) were snowshoe hare, red deer, and mule deer. The top 3 species affiliated with community 3 (initialized with cougar) were cougar, elk, and California ground squirrel. The resulting rankings were displayed using the NetworkX python library.

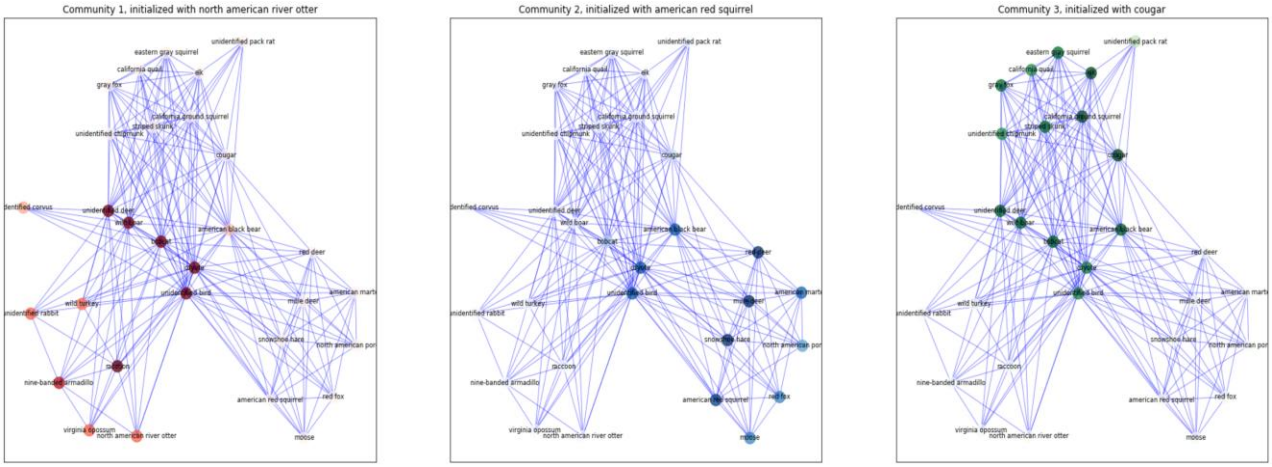

Figure 10: resulting community affiliation networks after BigCLAM

The top 7 most affiliated members of communities 1, 2, and 3 had Jaccard similarity scores of 0.17, 0.40, and 0.27 respectively with the 7 predetermined keystone species. Reported values were rounded to the second decimal place.

Conclusions

The final BigCLAM model has its limitations. For one, only a fraction of the biodiversity in the United States is represented in the data as only 44 useable species and 3 locations were present. Furthermore, many of the useable species were classified as 'unidentified' due to what I assume to be limitations of the convolutional neural network that was used to determine the identity of the animal in each image. This could distort the model because some of the unidentified species could be classified by biologists as a keystone specie themselves. The accuracy of this type of modeling is partially dependent on the functionality of the methods used to identify the species in each image.

Despite the vague identification of some of the species and the overall small number of species in the data, I would say that my BigCLAM model did an overall good job of identifying keystone species in an ecosystem, based on my measure of Jaccard similarity score between its top 7 members and the 7 predetermined keystone species. The seemingly low Jaccard similarity score values are expected, as the number of keystone species that exist within a given ecosystem may be variable, especially for the limited number of keystone species available in this data. I would say that a Jaccard similarity score of 0.2 or above indicates successful community modeling as that would imply that the model could identify 2 or more keystone species in that community. A rather informal measure of identifying successful community modeling would be checking if any of the predetermined keystone species exist in a more exclusionary ranking number, such as the top 3 most affiliated members. I would say that communities 2 and 3 successfully identified keystone species, as both had a Jaccard similarity score above 0.2 and 1 predetermined keystone specie in their top 3 most affiliated members.

The Apriori step of analysis acted as an effective preprocessing step for BigCLAM. The array of triples created in this step effectively acted as a list of edges for the undirected graph

species network. Filtering out non-frequent items in the dataset contributed to the sparsity of the affiliation matrix. Although the original paper for BigCLAM suggests initializing centers using locally minimum neighborhoods (Yang, Leskovec 6), using the Apriori confidence association rule to find mean confidence scores to rank the predetermined species seemed to work well for initialization of centers.

While this project was focused on identifying keystone species, the resulting model has potential to predict actual ecosystems. A potential way to test the BigCLAM model for ecosystem prediction would be to obtain sets of all unique animals in each state, partition the graph network, and run a Jaccard similarity measure on the members of each subgraph and the location-based sets. If the current graph were to be partitioned, more location data would be necessary as there are currently points of overlap present in all 3 of the modeled communities that would prevent partitioning.

Works Cited

"Bobcat." *Bobcat | Office for Environmental Programs Outreach Services*, University of Kentucky.

"ECOLOGICAL ROLE OF BEARS." *Project Coyote*, https://projectcoyote.org/carnivores/bear/.

"NETN Species Spotlight." *National Parks Service*, U.S. Department of the Interior.

"North American River Otter." *Smithsonian's National Zoo*, 11 July 2018, https://nationalzoo.si.edu/animals/north-american-river-otter.

Ortega, Mallory. "Human Wildlife Interactions." *Coyotes*, Utah State University, https://extension.usu.edu/wildlife-interactions/featured-animals/coyotes.

Posthumus, Erin. "Can Red Squirrel Middens Influence Species Diversity?" *Conservation.arizona.edu*, University of Arizona, 11 May 2020.

"Survey Mountain Lion Populations – 2019." *Yosemite Conservancy*, 15 Jan. 2020, https://yosemite.org/projects/survey-mountain-lion-populations-2019/. – cougar

Yang, Jaewon, and Jure Leskovec. *Overlapping Community Detection at Scale: A Nonnegative Matrix ...* https://cs.stanford.edu/people/jure/pubs/bigclam-wsdm13.pdf.