PREDICTING ENDANGERED SPECIES HABITAT WITH CLUSTERING AND REGRESSION

Samuel Veliveli
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22903
rrs4bw@virginia.edu

Nick Garonne
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22903
nkj9hk@virginia.edu

December 7, 2024

ABSTRACT

We aim to identify areas where members of endangered species–specifically, the Shenandoah Salamander– frequent. We first use K-means clustering to build intuition about the problem and identify features that correlate to each other. Then, we perform regression analysis to predict the occurrence of the Salamander based on environmental features. Then, we apply this to a novel dataset in order to identify future areas for survey or preservation.

Code: https://colab.research.google.com/drive/1RbcapSwJ5UTKCBTZnGmwruB4l4lzlJsF?usp=sharing

1 Introduction

1.1 Motivation

The state of Virginia is home to many endangered species, including the Shenandoah Salamander. However, conservation resources are limited, and efforts should be put towards areas with the most potential benefit. Our proposal is to use clustering and regression analysis to better predict where the Shenendoah Salamanders can occur and thrive. By building a model that predicts occurrence based on physical features such as altitude and temperature, we hope to find areas where conservation efforts should be concentrated. Additionally, we plan to apply this model to areas where Shenendoah Salamanders are not known to occur in order to find candidates for species introduction.

1.2 Method

We first used k-means clustering to build up intuition about where salamanders tend to be found based on the physical features of where they have been found before. We clustered on slope, aspect, and elevation of past surveyed salamanders. Then, we built a regression model to consider a wider range of variables– slope, aspect, precipitation, average temperature, and elevation– to predict whether salamanders can be found in that location given those parameters. Finally, we ran the model on novel locations in order to predict the presence of previously unknown salamanders and to identify areas that would be hospitable for species introduction.

This data contains geographical information regarding observed distribution for the Shenandoah mountain salamander. It is accessible here: https://www.sciencebase.gov/catalog/item/6639093cd34edc29f40aed15[3]. It was collected by the USGS. There are two data files of relevance to us: the first lists species occurrences by the slope, aspect, and elevation of the occurrence. The second contains survey areas with their corresponding temperatures, moss coverage, precipitation, and cover, as well as the number of salamanders seen during the survey period.

1.3 Experiments

We used both a regression and clustering methodology for this project to start. For clustering, we wanted to analyze the factors in the salamanders' environment. The salamanders live on north- and northwest-facing talus slopes and in hardwood forests above 2,870 feet. They prefer the cool, moist conditions found at higher elevations, where the mountaintops are often covered in fog.

We found a dataset that had information regarding the elevation in meters, aspect in degrees, and slope in degrees, in areas where the salamander was found. We decided to use k-means clustering on these 3 values, to see what combination of these factors were related to each other with regards to the salamanders' presence. We used the elbow plot method as well as the silhouette score to determine the optimal number of clusters. Based on the elbow plot and the silhouette score (Figure 1), that number turned out to be 5. The specific silhouette score was 0.34.

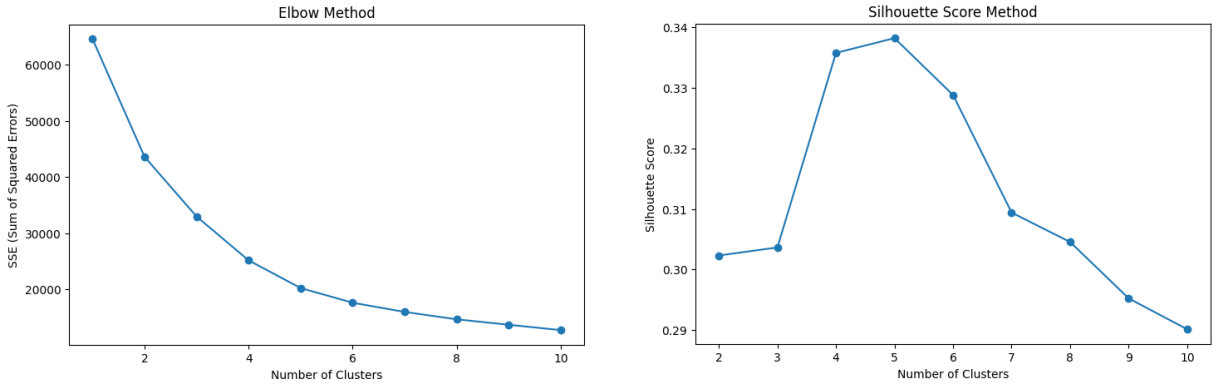

Figure 1: Clustering evaluation methods

Once the optimal cluster number was found, we ran a preliminary clustering with 4 clusters. We generated a 3d visualization (Figure 2) of our clustering along the elevation, aspect, and slope dimensions.

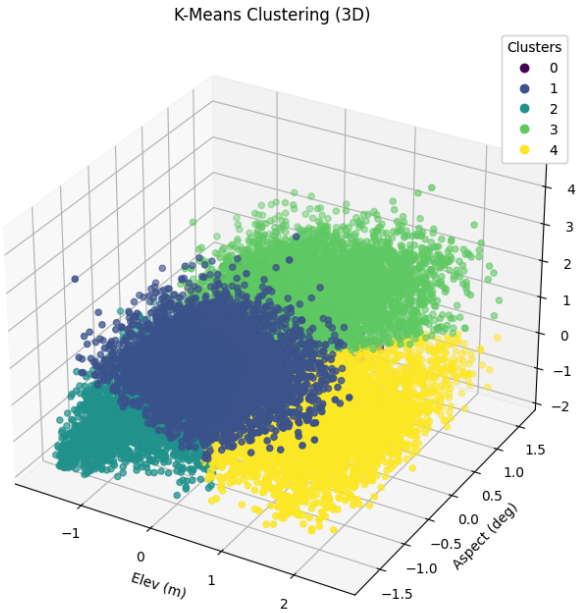

Figure 2: Preliminary clustering

Looking at this generated figure, we have some intuition for the way that the physical features of elevation, aspect, and slope correlate together. It seems like generally the distribution is pretty uniform. This suggests that features such as temperature and precipitation will be more important in the regression than these simple features.

For regression, we found a dataset that had information regarding factors like a region's precipitation, temperature, and the amount of cover it had, as well as the number of occurrences of the salamander. We aimed to predict the number of occurrences of the salamander in the test area based on these features. First, we removed non-relevant features such as date and other metadata. Then, we used both a random forest and decision tree regression, and did a grid search on both of these models to find the optimal hyperparameters. We also did a 5-fold CV mean squared error to evaluate both of these models. The best model turned out to be a random forest regressor with a test MSE of 3.15. We also did initial testing of a deep neural network with a 32 node hidden layer and a 16 node hidden layer which we trained for 50 epochs. However, the results were disappointing, with a test MSE of 6.4.

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.85	0.92	0.88	61
1	0.90	0.81	0.85	54
accuracy			0.87	115
macro avg	0.87	0.87	0.87	115
weighted avg	0.87	0.87	0.87	115

Figure 3: Evaluation of our chosen model

Once we had the binary classifier, we obtained data to process new locations. We obtained this data from two sources: the elevation, slope, and aspect data from a 15m resolution Digital Elevation Model (DEM) [2], and the climate data from the PRISM [3] dataset from the University of Oregon. In order to extract data from these, we used QGIS to create a grid of 10000 sample points over an area covering the US Mid-Atlantic and midwest. Then, we exported as a CSV and passed the data to our model.

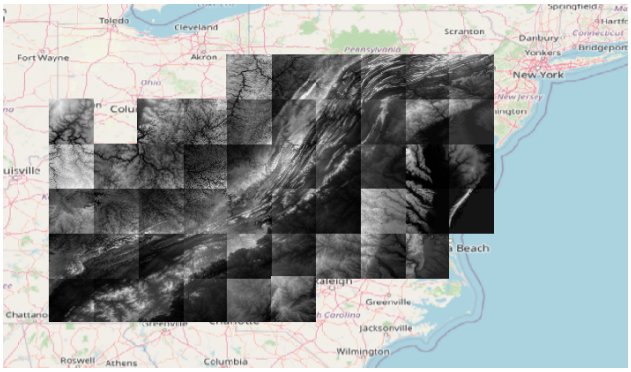

Figure 4: Survey area in grey

1.4 Results

As we mentioned before, we switched out topic to from clustering to predicting areas of future potential for the species. We used a random forest model and trained on features such as temperature, elevation, and precipitation. The model predicted locations in the Virginia, West Virginia, and North Carolina Appalachian Mountains, which makes sense given the salamanders tendency to live in elevated areas. These locations are potential starting points for habitat expansion or additional surveying. The map is shown in Figure 7.

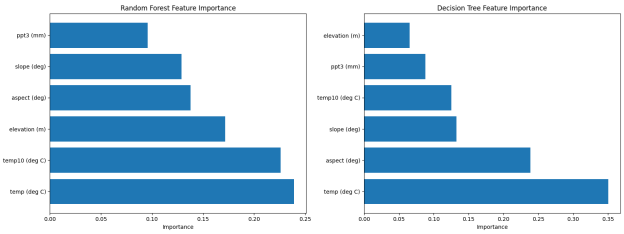

Figure 5: Feature importance for random forest model

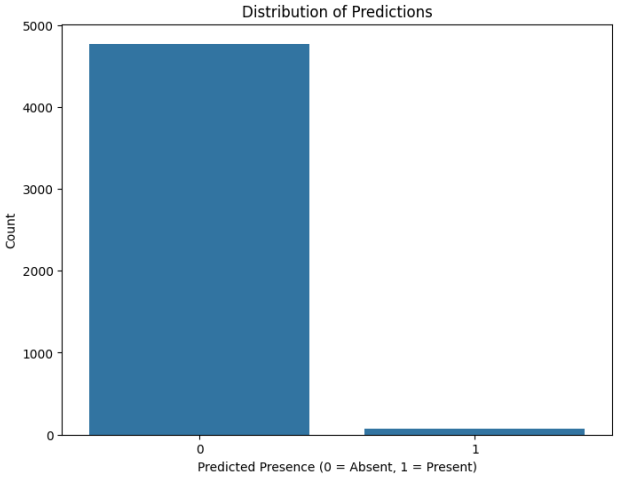

Figure 6: Distribution of Predictions

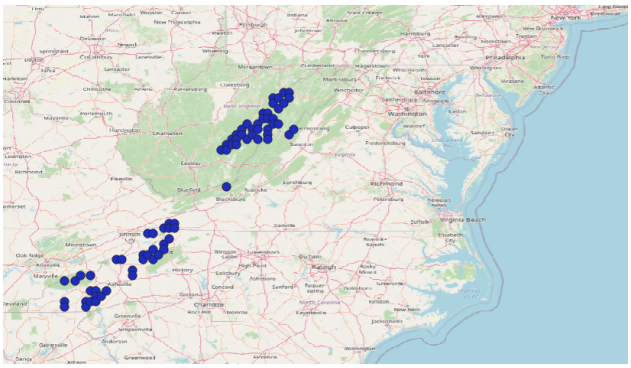

Figure 7: Locations predicted to have Shenedoah Salamanders

2 Conclusion

We were able to show an intersection with conservation ecology and machine learning. Often times, researchers in conservation biology need to prioritize funding, resources, and time, and our project showed a proof of concept of a methodology that could aid resource conservation in which locations should be prioritized. Our experiment was able to show potential areas in Virginia where the Salamander could potentially be located. Based on our results, it may be worth expending resources to survey the areas that we have found as hospitable to the Shenandoah Salamander. Even if no salamanders are found, if the opportunity arises to introduce the species to new habitats, the areas we identified would likely be hospitable to the species. We hope that our experiment can provide a framework for similar work to be done with other endangered species and that our work can be validated by field ecologists.

3 Contribution

Nick: Worked on tuning and feature engineering for models, research, and report

Samuel: Worked on predictive model, regression analysis, and clustering methodology. Helped edit report as well.

4 References

[1] Pang, Sean E. H., J. W. Ferry Slik, Damaris Zurell, and Edward L. Webb. 2023. "The Clustering of Spatially Associated Species Unravels Patterns in Tropical Tree Species Distributions." Ecosphere 14(6): e4589. https://doi.org/10.1002/ecs2.4589

[2] PRISM Climate Group. PRISM Climate Data, Monthly. Retrieved from https://prism.oregonstate.edu/6month/.

[3] U.S. Fish and Wildlife Service. (n.d.). Shenandoah Salamander (Plethodon shenandoah). Retrieved from https://www.fws.gov https://www.usgs.gov/data/data-support-updated-range-map-plethodon-shenandoah-and-evaluating-support-multiple-models

[4] U.S. Geological Survey. Digital Elevation Model 1/3 Arc-second. Retrieved from https://apps.nationalmap.gov/downloader/.