Exploration of species-location relationships

This model explores sub-populations of target species chosen earlier in Florida. Identifying regional populations of species allows to identify which groups might have been impacted by natural hazards such as extreme weather, within certain geographic areas. This model compares three machine learning algorithms: k-means, k-medoids and hierarchal clustering using silhouette score to determine he most suitable approach for the given species data.

K-means clustering works by randomly assigning each data point to a cluster and then computing the centroids. This is followed by the measurement of the distance between each data point and centroid and then reassignment to the closest centroid. Centroids are recalculated after each iteration. This is repeated until all observations are assigned to the closet centroid. [1] K-means algorithm is highly efficient which makes it suitable for great data sets, however, its main disadvantages include a tendency to favour spherical clusters which may not be appropriate for all sorts of data and sensitivity for outliers. K-Medoids algorithm works similarly to K-Means, but it chooses one of the data points as a medoid instead of finding centroids which makes it less sensitive to outliers – extreme values can’t affect the mean needed to find a centroid. However, this approach makes it also slower compared to K-means. [2]

Agglomerative clustering is a type of hierarchal closeting used by this model. At first, it treats all data points as separate clusters, and then, merges clusters within proximity. This is continued until the target number of clusters is obtained or all observations are merged into one cluster. Unlike in k-means and k-medoids algorithms there is no need to set the number of clusters before and if the distance threshold is set instead – clusters will stop merging when that distance between clusters is reached. [3,4] Dendrogram might be used to visualise Euclidean distance between clusters and choose the optimal number. Hierarchical clustering is more suitable for smaller data sets as it is very computationally expensive.

Silhouette Score was used to evaluate the performance of algorithms described above on target species in Florida. Silhouette Score is calculated by obtaining the mean distance from each data point to other data points in the same cluster and the mean distance between each data point and data points in the nearest other cluster. The overall silhouette score is the mean of silhouette scores from all points in the cluster, and it ranges from -1 to 1, 1 indicating well-defined clusters.

A graph with red lines and white text

Description automatically generated

A number with black text

Description automatically generated with medium confidence

(I will make this table nicer later)

Alternative ways of cluster evaluation include an Elbow Plot for k-means clustering and a dendrogram for hierarchical clustering. Literature reveals a range of concerns regarding Elbow plot [5,6,7] such as a decrease in the sum of squared errors (SSE) as the number of clusters (k) increases which often results in model failure on Real-World data sets. Therefore, this model uses the Silhouette Score as a more reliable alternative.  Dendrograms are a great way of visualising distances between clusters and visually choosing appropriate distance threshold between clusters. This model uses Silhouette Score to evaluate different numbers of clusters in hierarchal clustering, dendrogram might be added to expand this project in the future, but it was not added due to space constraints.

Following the Silhouette Score analysis, this model chooses the clustering algorithm that yielded the most well-defined clusters. Then the clusters obtained with the chosen algorithm are visualised to illustrate subpopulations of a given species in Florida.

A diagram of a number of dots

Description automatically generated

1. <https://www.datacamp.com/blog/machine-learning-models-explained>
2. <https://www.geeksforgeeks.org/k-means-vs-k-medoids-clustering/>
3. <https://www.geeksforgeeks.org/hierarchical-clustering/>
4. <https://scikit-learn.org/dev/modules/generated/sklearn.cluster.AgglomerativeClustering.html>
5. <https://doi.org/10.1145/3606274.3606278>
6. Milligan, G. W., and Cooper, M. C. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 2 (June 1985), 159–179.
7. Ketchen, D. J., and Shook, C. L. The application of cluster analysis in strategic management research: An analysis and critique. Strategic Management Journal 17, 6 (1996), 441–458.