Nehemya McCarter-Ribakoff

Dr. He

MATH 448 T/Th 11:00 AM

23 February 2017

Project Proposal

Members: Nehemya McCarter-Ribakoff

About the Data

The data is provided courtesy of the National Park Service. The National Park Service Species List Database is managed and updated by staff at individual national parks..

species.csv contains information about species identified within United States National Parks. The set contains the following columns: Species ID, Park Name, Category (Mammal, Reptile, etc.), Order, Family, Scientific Name, Common Names, Record Status. There are 119,248 observations.

parks.csv outlines basic information about each of the included National Parks. This set is relevant for geographic information, as it contains the following data: Park Code, Park Name, State, Acres, Latitude, Longitude. There are 56 observations (56 parks).

Source: https://www.kaggle.com/nationalparkservice/park-biodiversity

Problem: Unsupervised learning, inference/clustering

I would like to study the National Parks from a couple difference angles. For inference, which predictors, if any, have a relationship with biodiversity (ie latitude, park acreage, state, etc.), and what model best captures the relationship between biodiversity and its predictors? I would also like to cluster the parks by similarity to see how different parks are similar (ie biodiversity, dominant taxa (which parks have similar species/families/etc.), etc.) and what predictors, if any, correlate to these similarities.

Comments/Concerns

NPS withholds information regarding some endangered/threatened species for their safety. NPS clarifies that the data are a work in progress and time spent collecting data varies from park-to-park.