Biodiversity in National Parks in the United States of America

Exploratory Data Analysis by Sangeetha Ramanuj and Shea O'Day

Questions

"Biodiversity" is the foundation of life on Earth. Though the estimated 8.7 million species on Earth may seem to dismiss the importance of one species' existence, each species is an integrated and significant part of their ecosystems. The National Geographic Society [2] defines conservation as "the care and protection of these 'natural' resources so that they can persist for future generations," including "maintaining diversity of species, genes, and ecosystems, as well as functions of the environment." As stated in the above definition, a major goal of conservation is to allow resources to persist for the future. The key is maintaining sustainable habitats for the preservation of these species.

In this paper, we explore the records of species throughout US National Parks to illustrate the biodiversity and ecosystem threats of national parks in the United States. We ask:

1. Does location and park size impact the biodiversity and number of species threatened in national parks?
2. How do different types of species make up the biodiversity of the national parks? Which national parks have the largest numbers of species with a conservation status?

To illustrate this, we look at recorded species counts in 56 of the United States national parks, and analyze the types of species that make up the parks.

Methods

We first identify the unit of analysis of this project, **(national parks)**, which is an independent variable. The metric of biodiversity for each national park is the **total species count** in it. Other dependent variables are park location (latitude, longitude) and type of species (birds, mammals, etc.). The other metric of interest is the **count of species with a conservation status**, i.e. the nature of threat to that species. The analytic data set contains 56 unique observations each representing a national park and 33 variables. Variables document its location, size, count of total species and specific species (mammals, amphibians, etc.), count of species with a conservation status and its specifics (count endangered species, in recovery, threatened, etc.). This was obtained from two independent data sets on parks `parks.csv` and species present in each park `species.csv` provided by the National Parks Service [1]. The species data set (containing 119,248 observations) was aggregated by park and category of species, and joined with the parks data set (containing geo-spatial attributes of 56 parks in USA) to obtain all relevant variables for the analysis. We cleansed data to eliminate observations with N/A values, and conducted full joins to ensure that no parks were dropped from either of the data sets.

Findings

Criteria	mean	median	standard_dev	iqr	maximum	minimum
Total Species	2129.429	1815.500	1202.955	985.500	6623.000	848.000
Mammal	69.05357	67.00000	32.24194	28.00000	212.00000	6.00000
Bird	260.73214	245.50000	89.66006	68.25000	531.00000	44.00000
Reptile	23.98214	16.50000	22.72963	33.50000	90.00000	0.00000
Amphibian	13.26786	8.50000	14.06542	10.25000	71.00000	0.00000
Fish	70.64286	23.50000	135.44787	48.50000	818.00000	0.00000
Vascular Plant	1164.6607	1091.5000	486.8466	581.0000	2761.0000	231.0000
Fungi	110.7679	2.0000	240.7847	75.7500	1363.0000	0.0000
Insect	256.2321	41.0000	473.1087	282.0000	2414.0000	0.0000
Latitude	41.23393	38.51000	10.90883	11.35250	67.78000	19.38000
Longitude	-113.23482	-110.98500	22.44029	18.17000	-68.21000	-159.28000
Acres	927928.1	238764.5	1709238.3	748349.8	8323148.0	5550.0
Total Conservation Status	84.250	80.500	36.826	31.000	247.000	21.000
Endangered	6.678571	4.000000	8.831981	4.250000	44.000000	0.000000
In Recovery	1.375000	1.000000	1.229375	1.000000	7.000000	0.000000
Species of Concern	68.62500	68.00000	27.52425	29.00000	177.00000	21.00000
Threatened	3.285714	2.000000	2.927700	4.000000	16.000000	0.000000

Table 1

In **Table 1**, the first row shows statistics across categories of species, followed by statistics by category, location (longitude, latitude), park area (acres) and conservation status (endangered, in-recovery, etc.). The mean total species count per park is around 2130, while the range is 5775, with a high standard deviation of 1203, indicating a vast biodiversity spread across parks. The IQR for vascular plants, insects, and fungi are also high, indicating that the central portion of the recorded data is spread out. The average number of species with a conservation status is relatively low (about 84) compared to the number of species, but goes up to 247 (maximum) in a park. The average number of endangered (<7) and threatened (<4) species are relatively low compared to the number of species per park, but go up to a high of 44 and 16, respectively.

Figure 1: Total Species vs Park Size

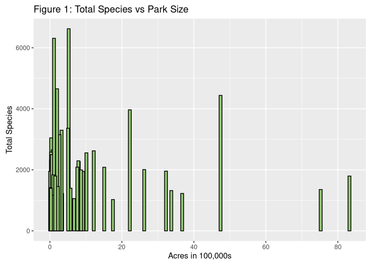
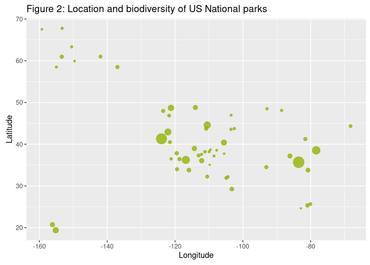
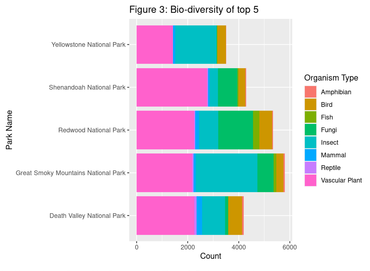


Figure 2: Location and biodiversity of US National parks



The biodiversity in relation to park size and location is described in **Figure 1** and **Figure 2**, respectively. From **Figure 1**, we see that the relatively smaller parks (< 500,000 acres) are the most bio-diverse with up to 6500 species, while there are very large parks (7.5-8.5 million acres) with low biodiversity (< 1900 species). The most bio-diverse parks are distributed in the middle-to-eastern part of USA (see **Figure 2**). It is clear from these, that park location does have an impact on bio-diversity, more than park size. Small sized parks located in places with favorable conditions tend to be more bio-diverse, while parks in places with extreme conditions (Alaska) tend to be less bio-diverse. One exception to this is Death Valley, which has extreme conditions, but yet one of the most bio-diverse.

Figure 3: Bio-diversity of top 5



Looking at the top 5 most bio-diverse parks in **Figure 3**, Redwood's species are more evenly distributed, although Great Smoky Mountains has the most species (mostly vascular plants and insects). Furthermore, in **Figure 4**, the parks with greater numbers of species with a conservation status tend to be in the Southwest region of USA. The Great Smoky Mountains – the most bio-diverse park – has around 100-150 species with a conservation status, whereas Death Valley – the 5th most bio-diverse park – has around 200-250. Death Valley has the most species with some degree of threat. Endangered species exist in each park, however, Death Valley and Redwood have the most. Each of these parks has a small number of species in-recovery, with Redwood having a noticeable count (see **Figure 5**). Moreover, in **Figure 6**, amongst the least bio-diverse parks, Petrified Forest and Dry Tortugas have the most number of endangered species, with very few species in recovery.

Figure 4: Location and Conservation Statuses of US National Parks

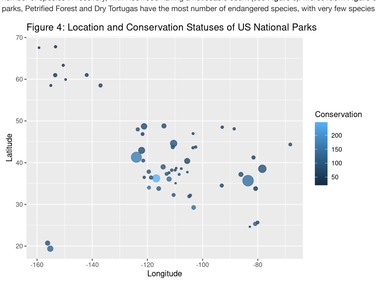


Figure 5: Conservation statuses of top 5

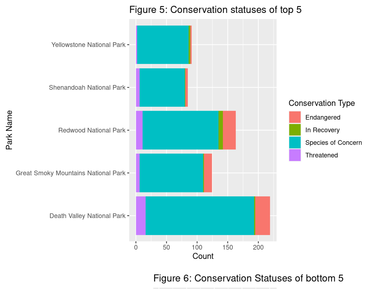
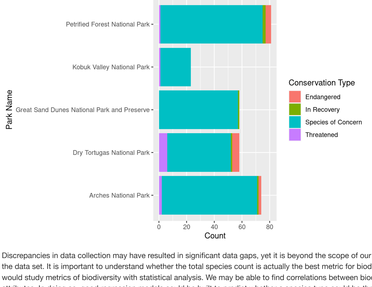


Figure 6: Conservation Statuses of bottom 5



Discrepancies in data collection may have resulted in significant data gaps, yet it is beyond the scope of our project to confirm or deny validity of the data set. It is important to understand whether the total species count is actually the best metric for biodiversity. To pursue this further, we would study metrics of biodiversity with statistical analysis. We may be able to find correlations between biodiversity or threat with other park attributes. In doing so, good regression models could be built to predict whether a species type could be threatened based on the park in which it resides and its various attributes.

References

1. National Park Service. (2017). Biodiversity in National Parks. Available from https://www.kaggle.com/datasets/nationalparkservice/park-biodiversity.
2. National Geographic Society.(2023) Biodiversity. Available from https://education.nationalgeographic.org/resource/biodiversity/.