Animal Packers

Avery Faller, Fanny Heneine, Crystal Lim
averyfaller@g.harvard.edu, fannyheneine@g.harvard.edu, crystallim@g.harvard.edu
AM207 – Spring 2016 – Professor: Verena Kaynig – TA: Rafael Galarza

Introduction

Animal populations in the United States are closely monitored and controlled, if necessary.

An extensive database of animal observations is available GBIF.org. They have millions of records of animal observations in the wild, spanning the globe from dozens of data sets.

Given this dataset, we set out to see whether we could use observations of a particular species to infer information about that species range, population total and other co-occuring species.

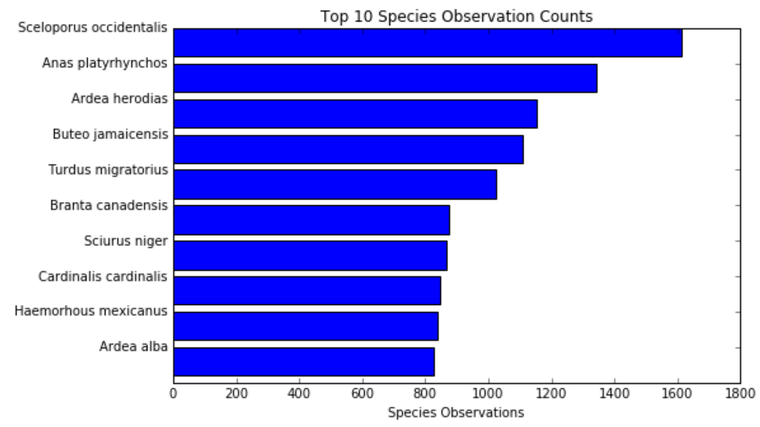
We used observations of deer in the United States, in particular, to test the dataset due to their well-documented population numbers and ranges.

Data

We examined the current range of deer in the US overall, and on a per-species level. Some questions we set out to answer include: What other animals appear frequently with the deer? How does the deer population change from year to year? How do people's observations of deer reflect their true numbers and distribution.

In the US, there were observations for 8 separate species of deer. We downloaded the observations as a CSV and imported the data into an iPython Notebook as a Pandas Dataframe.

There are large skews in the data for different species and for different years as scientists interests' and individual studies dictate which species and which locations get studied each year, and hence the observations in GBIF.org are heavily skewed to animals of interest to scientists. We kept this in mind as we conducted our experiments.



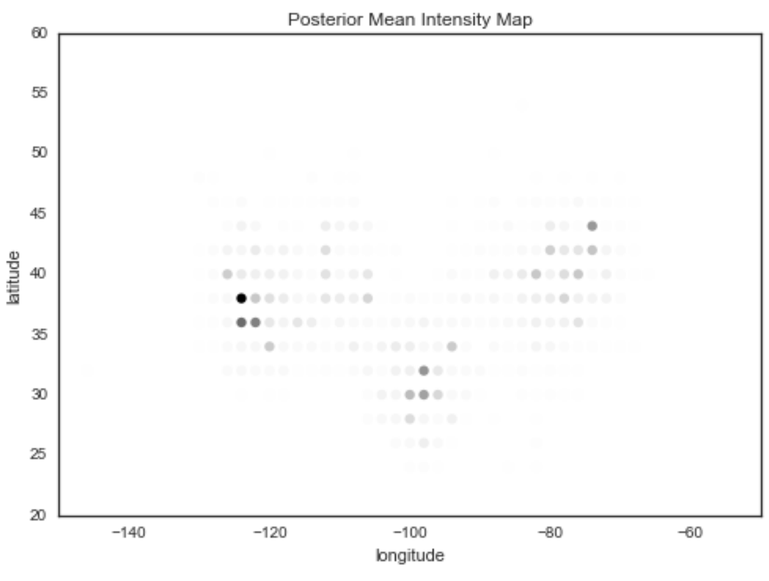
Top 10 US Species Observations, 2015

Above is a bar chart of the top 10 species in the US, ordered by number of observations. The most observed species in the database was the Western Fence Lizard, followed by Mallards,

Results

Intensity Map

Our analysis was broken into three sections. First we examined the lat & long of observations of deer within the US and tried to generate an intensity map of deer using Elliptical Slice Sampling. The result of this sampling can be seen below. The trick here was finding a way to spread the influence of a single observation over a large territory.

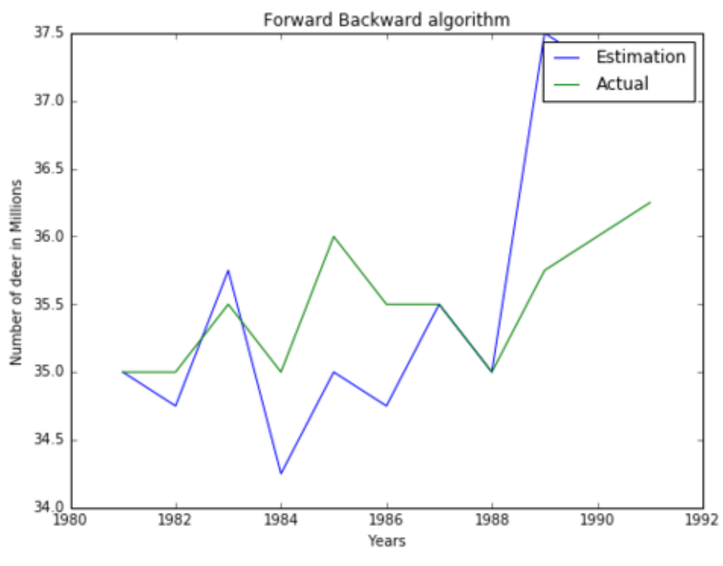


US Deer Intensity Map, 2015

Above is an intensity map for deer in the United States in the year 2015. The map was generated from 1280 observations.

Hidden Markov Model

Next, we attempted to predict the population of deer given the previous year's population using a Hidden Markov Model. We used several techniques (including Viterbi, Baum Welch and Forward-Backwards) in an attempt to model the data given the difficult nature of a continuous state space and very few actual data points to build up the transition and emission matrices.



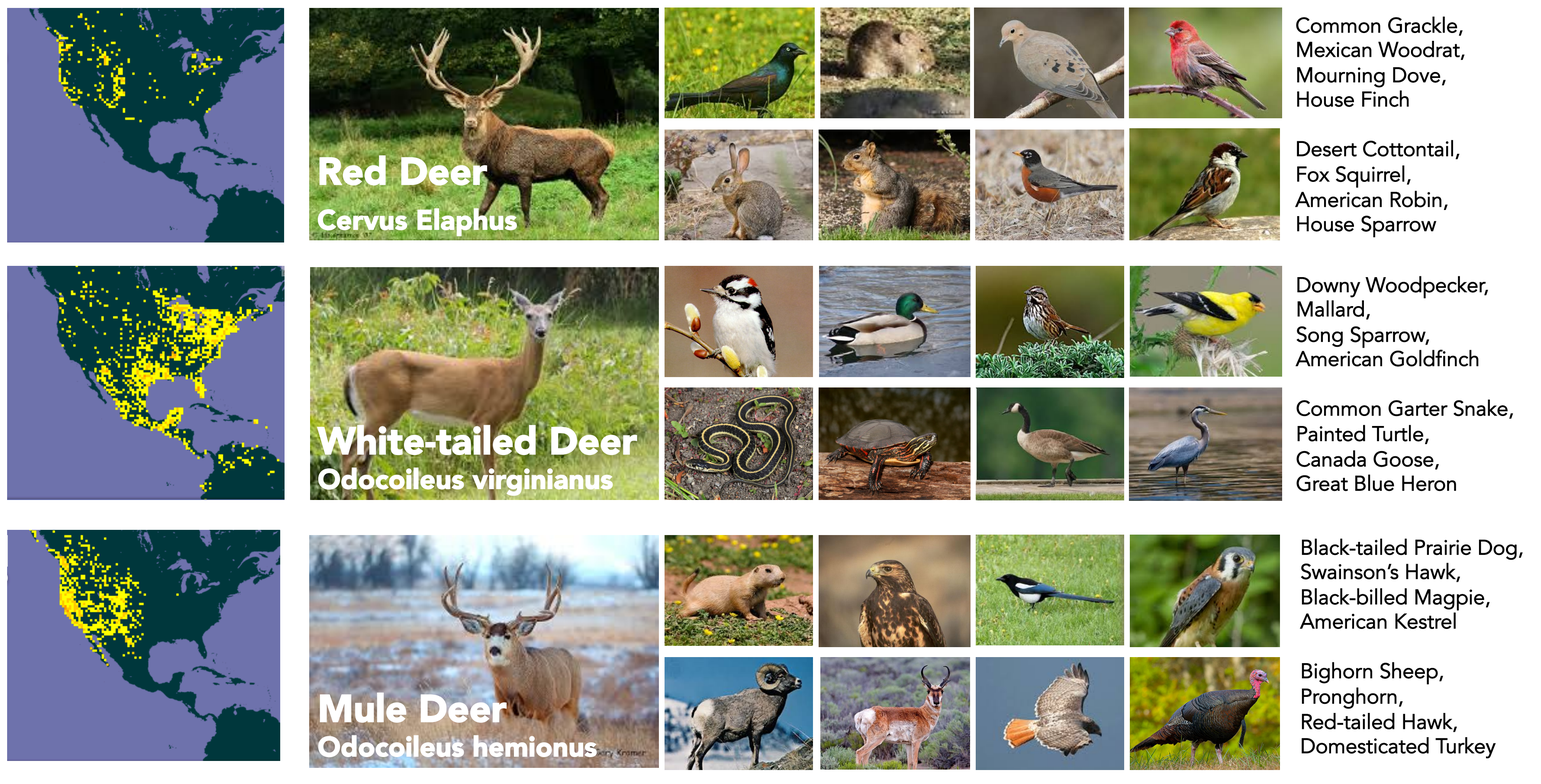
US Deer Total Population, 1980-1991

Above is the predictions of deer populations for our test set from 1980-1991.

Expectation Maximization

Next, we attempted to predict the population of deer given the previous year's population using a Hidden Markov Model. We used several techniques (including Viterbi, Kalman Filters and Forward-Backwards) in an attempt to model the data given the difficult nature of a continuous state space and very few actual data points to build up the transition and emission matrices.

Finally, we used Expectation-Maximization to cluster each deer species with other mammals, birds, reptiles and amphibians using the observed animals' latitude and longitude. We then selected the topic that each deer most closely associated with and found the top species in that topic. To the right we show three of the deer species and the other species that clustered with them. As can be seen from the maps, the deer all occupy fairly distinct geographical areas and thus have correspondingly varied associated animal species.



Conclusions

One aspect of this dataset that was made readily apparent by our analysis, was the bias presented in data collected by scientists studying specific species. This became most apparent in our HMM predictions. In order to get good results, it was necessary to normalize each year's data by the number of observed mammals that year over the total number of mammals present in all of the years. This helped account for the bias in any one particular year of far greater or fewer numbers of observations, and thus captured the real trends more accurately in the deer population. However, this brings up concerns about the validity of the data.

On the other hand, given a minimal number of observations, we were able to accurately predict co-occurring species using EM without performing any normalization.

Citations and Links

"The Decline of Deer Populations - Deer Friendly." The Decline of Deer Populations - Deer Friendly. Web. 03 May 2016. <http://www.deerfriendly.com/decline-of-deer-populations>.

Images from: picgood.xyz, wikipedia.org, ibc.lynxeds.com, pinterest.com, britannica.com, gbif.org