From Apocalypse to Action: A Data-Driven Strategy to Targeted Sampling for the “California Insect Barcoding Initiative”

Team 15: Andrews, Dylan; Chen, Arielle; Lehnen, Charles; Tinsley, Brian

Proposed project -

The rapid decline in insect populations, often referred to as the "Insect Apocalypse," is a pressing global concern. Recent studies suggest that we may have lost up to 75% of insect biomass since the 1980s, primarily due to human development, pesticides, and climate change (*Hallman et al. 2017*). Insects are fundamental to the health and functioning of ecosystems. They play myriad roles, from pollinators ensuring plant reproduction to decomposers breaking down organic matter. Their significance extends beyond just ecological services; they are deeply intertwined with numerous trophic levels, influencing the abundance and behavior of other species. Insects are especially suited for environmental impact assessment due to their high species diversity, omnipresence, and pivotal role in natural ecosystems. Their decline not only disrupts these systems but also serves as an indicator of broader environmental changes (*Rosenberg et al. 1986*).

The Natural History Museum (NHM) of Los Angeles County has taken a proactive approach to address this issue by leading the ambitious California Insect Barcoding Initiative (CIBI), aiming to barcode every insect species in California. DNA barcoding is a sophisticated technique that uses primers to identify and categorize organisms based on a specific, conserved region of their DNA. The resulting "barcode" sequence acts as a unique identifier for each species, enabling researchers to differentiate between even closely related species that might appear morphologically identical (*Hebert et al. 2003*). Through CIBI, the NHM is creating a comprehensive database that can serve as a baseline in monitoring California insect populations in the face of decline.

Our proposed project seeks to complement NHM's efforts using data from the extensive Global Biodiversity Information Facility (GBIF) database. We will map the distribution of mantises (superfamily: Mantodea) records across California. Gaps and clusters in distribution will be identified. Associations between clusters with other biologically significant predictors including ecoregion, human population density, and climatological data will be quantified. This will serve to predict the regions that mantises are likely to be found which will provide interesting insights into the biological needs and ecological role of mantises in California. We will overlay this occurrence prediction map with a map of CIBI sampling sites, resulting in a targeted recommendation on where NHM would most effectively focus their future mantis barcoding sampling resources. Altogether, this will serve as a proof of concept workflow that could be used by NHM for all other insect taxonomic groups to ensure that their insect barcode sampling resources are utilized efficiently and effectively, in doing so contributing to the monitoring and prevention of rapid population decline of insects.

Description of dataset -

Our data can be split into four different tables: biodiversity, human population density, climatological data, and ecoregion data. Initially we selected 7 insect groups (Lepidopterans, Sphingids, Odonates, Orthopterans, Blattodeans, Mantises, and Carabids) that include species simple enough for amateur biologists to identify so that data is accurate. By comparing these groups through preliminary analyses, we selected mantises (Mantodea) for this project because records are relatively uniformly distributed geographically (*Figure 1*) and between species (*Figure 2*). Our biodiversity data contains the observations of mantises by observers and has 4,556 records that include helpful columns such as species name, latitude, longitude, date, and elevation of the place and time of the observation. This dataset was sourced by the Global Biodiversity Information Facility (GBIF), an international data hub funded by governments worldwide to provide access to all types of data about life on Earth (*Telenius, Anders 2011*). The dataset is appropriate because it provides the mantis population data we need to assess how different factors affect their population distribution. Our human population density data comes from the United States Census, and provides us the boundaries of different neighborhoods as well as their populations in California, and will give us a good sense of the human population density in areas where we can see how it is correlated with the population density/ biodiversity of mantises in the area (*United States Census Bureau 2021*). The ecoregion data comes from the Environmental Protection Agency (*Griffith et al. 2016*), and contains information on the different ecoregions in California, which will be important to factor in when seeing how that impacts the population/diversity of mantises in certain regions. Finally, our climatological data is sourced from NOAA, and contains information such as daily temperature, precipitation, and weather types, which are all important factors to consider since some environments are more suitable for mantises than others.

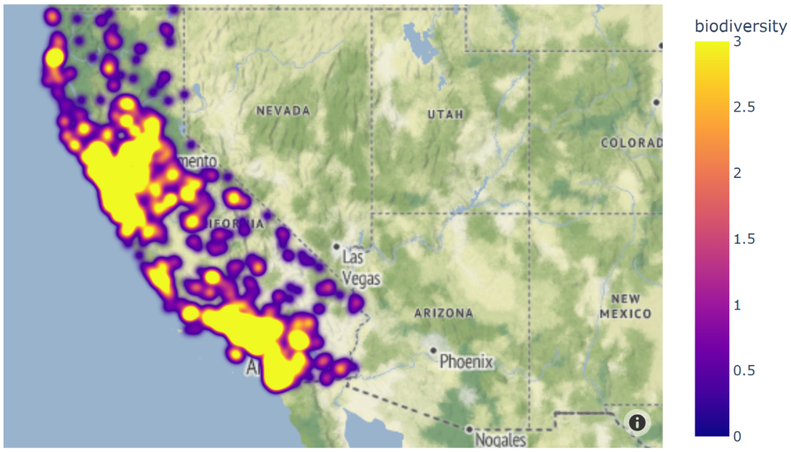

Figure 1: *Species biodiversity of Mantodea in California from GBIF records*

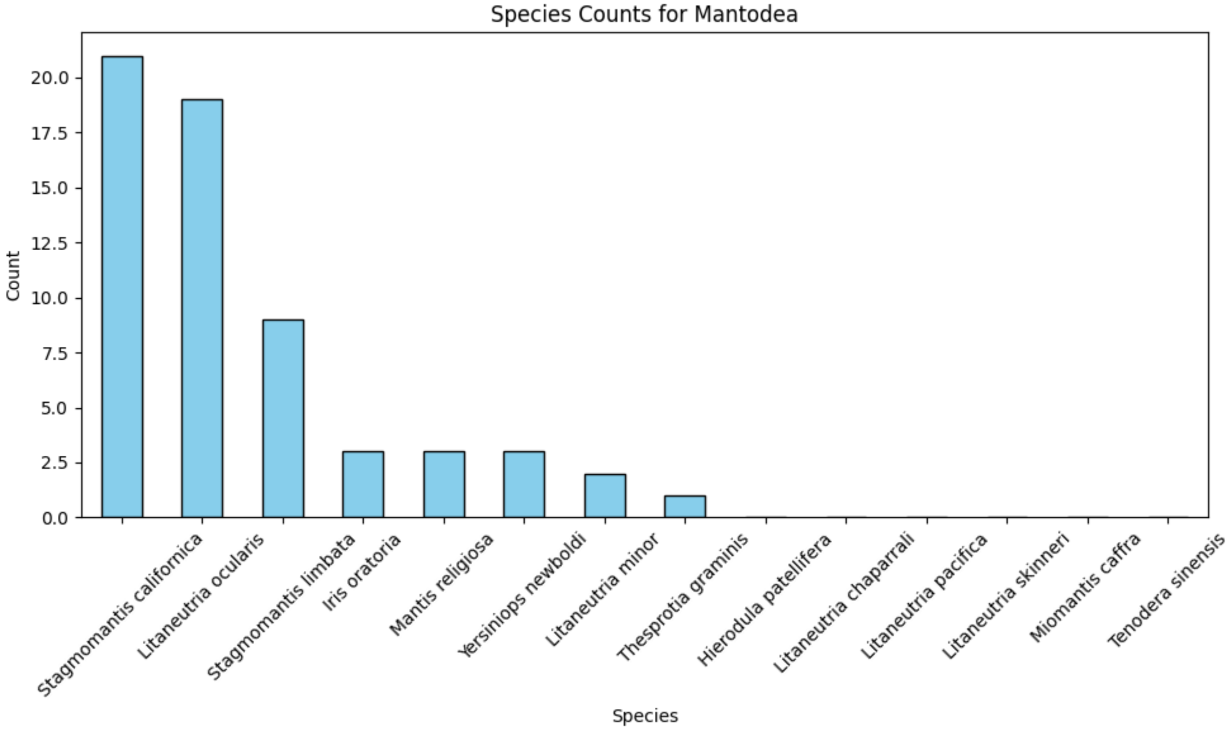

Figure 2: *Species counts for Mantodea in California from GBIF records*

Project plan -

Data Acquisition and Cleaning:

Our initial step involves acquiring the necessary datasets from the GBIF, the United States Census, the Environmental Protection Agency, and NOAA. Once acquired, our primary focus will be on the biodiversity data, specifically the mantis observations.

Given the vastness and diversity of the data, cleaning and refining it is necessary to ensure accuracy in our analyses. The following steps will be undertaken for data cleaning:

1. Identification Criteria: We will begin by filtering out records that are not identified to the species level. It's essential to have specific species data to ensure the precision of our analysis.

2. Geographical Data Criteria: Any samples missing latitude or longitude will be removed. Accurate geolocation data is crucial for our mapping and distribution analysis.

3. iNaturalist Data Quality Criteria: We will utilize the "Research Grade" classification from iNaturalist as a benchmark for data quality. Only samples that meet the "Research Grade" criteria will be retained. This ensures that the observations we include are verified and agreed upon by a majority of the community, ensuring their reliability. The criteria for "Research Grade" can be found at the provided link.

4. Expertise-Based Filtering: Recognizing that certain insect species might be challenging for amateurs to identify accurately, we will remove iNaturalist samples for those species. This step is taken to further enhance the reliability of our dataset.

Data Analysis and Mapping:

Post-cleaning, the refined data will be analyzed to map the distribution of mantises across California. Using advanced mapping tools and software, we will identify and highlight gaps and clusters in the distribution. The cleaned and processed data will then be cross-referenced with human population density, climatological data, and ecoregion data to draw correlations and insights. This comprehensive approach will provide a holistic view of the factors influencing mantis distribution in California to guide the NHM's future sampling strategies.

Division of Labor -

We plan to divide the labor evenly by 4 data types: biodiversity data, human population density data, climatological data, and ecoregion data. Gathering, cleaning, analyzing, and visualizing data will be done by one group member per data type. By splitting it up by the data types, we hope that each member will be an "expert" at their assigned data type so we can collaborate on how to most effectively combine and gain insights from the data. Domain expert Charles Lehnen will add domain specific knowledge to analysis.

References -

Griffith, Glenn E., et al. "Ecoregions of California." US Geological Survey Open-File Report 1021 (2016): 1-45.

Hallmann, Caspar A., et al. "More than 75 percent decline over 27 years in total flying insect biomass in protected areas." PloS one 12.10 (2017): e0185809.

Hebert, Paul DN, et al. "Biological identifications through DNA barcodes." Proceedings of the Royal Society of London. Series B: Biological Sciences 270.1512 (2003): 313-321.

Rosenberg, David M., H. V. Danks, and Dennis M. Lehmkuhl. "Importance of insects in environmental impact assessment." Environmental management 10 (1986): 773-783.

Telenius, Anders. "Biodiversity information goes public: GBIF at your service." Nordic Journal of Botany 29.3 (2011): 378-381.

United States Census Bureau. "TIGER/Line Shapefiles." U.S. Department of Commerce, Released August 6, 2021. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html#list-tab-790442341.