**From Apocalypse to Action: Enhancing the "California Insect Barcoding Initiative" Strategy through Predictive Analysis of Mantis Distribution**

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

## Problem Definition

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.[3] Insects are fundamental to the health and functioning of ecosystems, playing a myriad of 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 also especially well 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.[5]

In an attempt to understand and mitigate these concerns, various organizations have initiated efforts to collect and analyze data regarding insect species across the globe. The primary challenge in doing so is informing these group efforts, such as the California Insect Barcoding Initiative (CIBI), on where to collect data in order to efficiently utilize their limited resources.

## Background

The Natural History Museum (NHM) of Los Angeles County has taken a proactive approach to address the insect decline by leading the ambitious California Insect Barcoding Initiative (CIBI), aiming to barcode every insect species in California. DNA barcoding uses primers to identify and categorize organisms based on a specific, conserved region of their DNA, and acts as a unique identifier for each species, enabling researchers to differentiate between even closely related species that might appear nearly morphologically identical.[4] 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 research seeks to complement NHM's efforts using data from the extensive Global Biodiversity Information Facility (GBIF) database. We mapped the distribution of mantises (superfamily: Mantodea) records across California, and identified gaps and clusters in sampling. Associations between clusters with other biologically significant predictors including ecoregion,

human population density, and climatological data were quantified. This serves to predict the regions that mantises are likely to be found which provides direction to CIBI for prioritizing sampling along with interesting insights into the biological needs and ecological role of mantises in California. This occurrence prediction map can be overlaid by CIBI with their 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 datasets**

To effectively inform researchers on where to focus their sampling efforts in California, we incorporated a variety of data from various sources. For the majority of data analysis, we keyed relevant data with county data, obtained from the Database of Global Administrative Areas, containing highly accurate and high resolution administrative boundary geographical data.[1] We used counties as a reference point because of their relatively uniform and computationally manageable size across California. Ecoregion data was gathered from the US Environmental Protection Agency (EPA) and filtered to include less specific ecoregions (level 3) rather than the very granular (level 4) ecoregion definitions, providing a more generalized classification of ecoregions within California.[2] Each county was then mapped to a set of ecoregions contained within its borders. Human population density data comes from the US Census Bureau 2020 American Community Survey, which we narrowed down to total population by county.[8] Biodiversity data 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.[6] After extensive filtering, this data provided us access to the coordinates of each mantis observation (4,556 records) in California which we then mapped to specific counties. Climatic data comes from Daymet, a data product hosted by Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) and supported by NASA and the U.S. DOE. Through their Thematic Real-time Environmental Distributed Data Services (THREDDS) Data Server, we obtained 129GB of accurate, high-resolution (1km x 1km) daily weather parameters across North America which we clipped to California counties.[7] For an overview of the data used in our analysis, see the Entity Relationship Diagram in Figure 1 below.

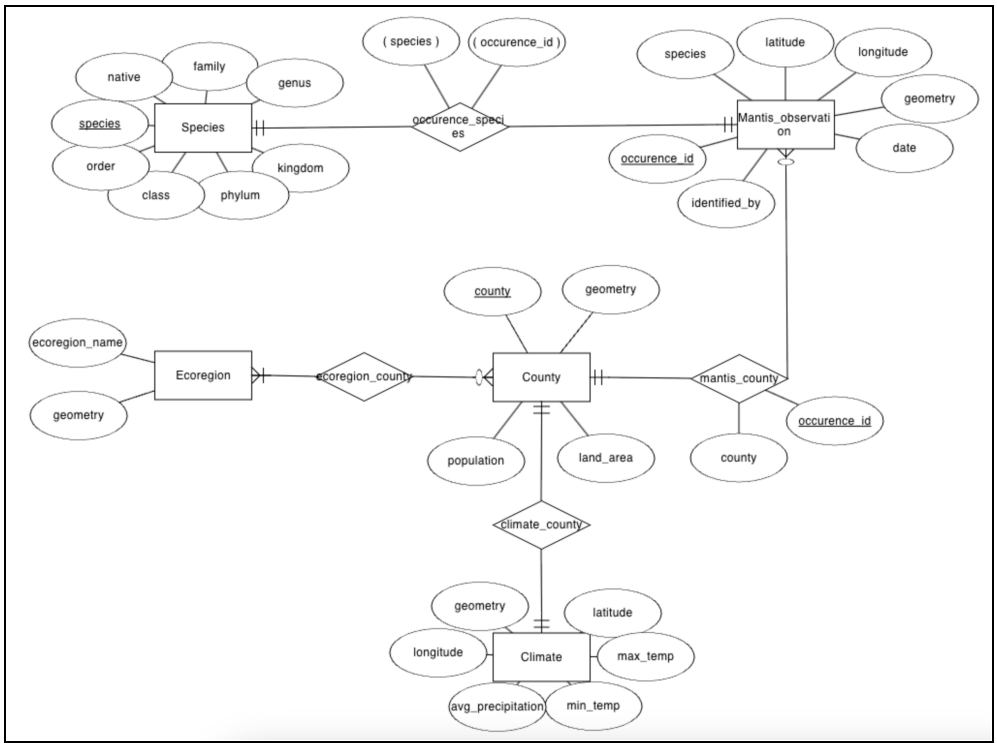



***Figure 1:*** *An Entity Relationship Diagram showing the relationships between biodiversity, climatic, human population, and ecoregion data after preprocessing data for relevant attributes*

## Methods and Results

After downloading, cleaning, and aggregating all relevant data with Python in Google Colab, we created a normalized relational database as outlined by the ER Diagram in Figure 1. To preview some of our data, we created heatmaps to show the distribution of human populations, mantis observations, ecoregions, and climatic data across California. To view these heatmaps, refer to our ‘Pre-proj-team15’ and ‘Prog-proj-team15’ documents.

As part of our initial data analysis, we performed four different linear regressions to gain a better understanding of the correlation between human and mantis populations across California counties. We first isolated the necessary data and exported it to Excel for ease of graphing plots and trend lines. The linear regressions compared human population to a) individual mantis count per county, b) mantis species count per county, c) percentage of introduced mantises per county, and d) percentage of native mantises per county.

The linear regression for human population vs. individual mantis observation count per county had a correlation coefficient of 0.944 and $R^2 = 0.891$ (Figure 2), indicating a strong positive correlation between higher mantis observation counts in counties and larger human populations.

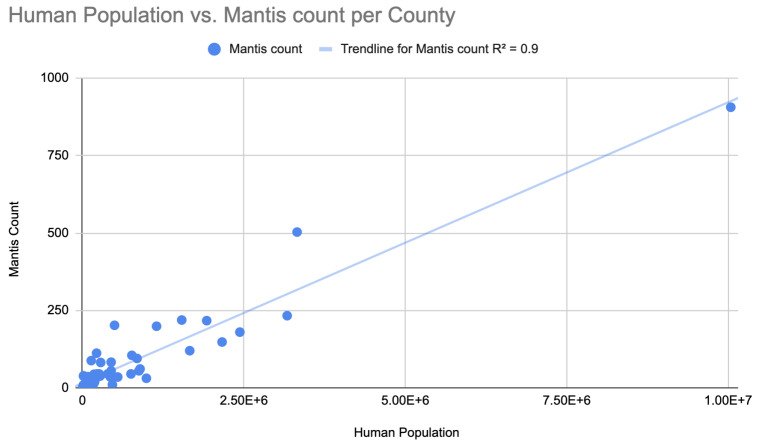



***Figure 2:*** *A linear regression of mantis counts vs. human population*

However, a county's higher human population did not suggest higher mantis biodiversity. The linear regression for human population vs. mantis species count per county had a correlation coefficient of 0.609 and $R^2 = 0.371$. When comparing human density to the percentage of introduced vs native species, the linear regressions suggested a weak correlation. Regression for introduced mantises per county resulted in a negative correlation coefficient of -0.326 and $R^2 = 0.106$. Human population vs. percentage of native mantises per county resulted in a correlation coefficient of 0.326 and $R^2 = 0.106$.

We chose to do decision tree analysis because it works well with a mixture of continuous/discrete and categorical parameters , and gives a simple, predictive, categorical result (predicted presence or absence). We also liked that it would weigh the importance of different parameters and give summary statistics on model quality. We constructed a primary table with county information including ecoregion, human population, climatic, and biodiversity data, dividing the dataset into training and test segments. After training the decision tree classifier using the *sklearn.tree* library we visualized and saved the trees, evaluating their performance through metrics like accuracy, precision, recall, and F1 scores for each species using cross-validation with folds to enhance robustness (Figure 3). This methodology was also applied to differentiate between introduced and native species. Our goal was to predict locations of more sampling for rare, native species.

| Species | Accuracy | Precision | Recall | F1 Score | Status | Count |
|---|---|---|---|---|---|---|
| Thesprotia graminis | 1.000000 | 0.000000 | 0.000000 | 0.000000 | native | 1 |
| Litaneutria chaparrali | 0.916667 | 0.000000 | 0.000000 | 0.000000 | native | 2 |
| Yersiniops newboldi | 0.916667 | 0.000000 | 0.000000 | 0.000000 | native | 3 |
| Litaneutria skinneri | 0.833333 | 0.500000 | 0.500000 | 0.500000 | native | 4 |
| Litaneutria minor | 0.833333 | 0.000000 | 0.000000 | 0.000000 | native | 7 |
| Litaneutria ocularis | 0.750000 | 1.000000 | 0.400000 | 0.571429 | native | 54 |
| Litaneutria pacifica | 0.583333 | 0.666667 | 0.333333 | 0.444444 | native | 102 |
| Stagmomantis californica | 0.250000 | 0.142857 | 0.250000 | 0.181818 | native | 336 |
| Stagmomantis limbata | 0.500000 | 0.500000 | 0.666667 | 0.571429 | native | 1713 |
| Hierodula patellifera | 0.916667 | 0.000000 | 0.000000 | 0.000000 | introduced | 1 |
| Tenodera sinensis | 0.750000 | 0.500000 | 0.333333 | 0.400000 | introduced | 44 |
| Miomantis caffra | 1.000000 | 0.000000 | 0.000000 | 0.000000 | introduced | 65 |
| Iris oratoria | 0.666667 | 0.750000 | 0.500000 | 0.600000 | introduced | 495 |
| Mantis religiosa | 1.000000 | 1.000000 | 1.000000 | 1.000000 | introduced | 1613 |

***Figure 3:*** *Table of accuracy, precision, recall, and F1 score metrics for each species*

The model was most effective for three species of interest: *Litaneutria skinneri*, *Litaneutria pacifica*, and *Litaneutria ocularis*. The *Litaneutria pacifica* model has relatively low accuracy but has balance between precision and recall. *Litaneutria ocularis* and *Litaneutria skinneri* have high accuracy and balance between precision (low false positive rate) and recall (actual positives are correctly assigned), indicating an effective model. Further analysis was conducted to generate concrete suggestions for CIBI. Here is an example for *L. skinneri* (Figure 4 and 5).

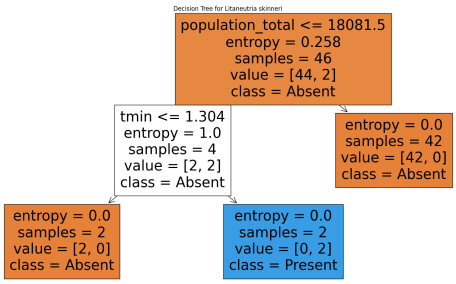



***Figure 4:*** *Decision tree for L. skinneri*

Feature Importances for Litaneutria skinneri:

| | Importance |
|---|---|
| population_total | 0.662984 |
| tmin | 0.337016 |
| tmax | 0.000000 |
| Eco_Northern_Basin_and_Range | 0.000000 |
| Eco_Coast_Range | 0.000000 |
| Eco_Central_California_Foothills_and_Coastal_Mo... | 0.000000 |
| Eco_Sonoran_Basin_and_Range | 0.000000 |
| Eco_Cascades | 0.000000 |
| Eco_Eastern_Cascades_Slopes_and_Foothills | 0.000000 |
| Eco_Southern_California_Mountains | 0.000000 |
| Eco_Southern_California/Northern_Baja_Coast | 0.000000 |
| Eco_Central_California_Valley | 0.000000 |
| Eco_Sierra_Nevada | 0.000000 |
| Eco_Mojave_Basin_and_Range | 0.000000 |
| Eco_Central_Basin_and_Range | 0.000000 |
| prcp | 0.000000 |
| Eco_Klamath_Mountains/California_High_North_Coa... | 0.000000 |

***Figure 5:*** *Feature Importances for L. skinneri*

Of these parameters, human population more strongly affects likelihood of presence of *L. skinneri*, which may have biological implications that humans have a larger impact on them than the natural effect of temperature (Figure 5).

We then filtered for counties that match criteria from the decision tree (Counties with population_total < 18081.5 and tmin > 1.304) to identify counties to target for future sampling. For *L. skinneri,* the model predicted presence in 3 counties where the species is already found (Inyo, Mariposa, Sierra), and one county, Trinity, where it has not yet been found (Figure 6).

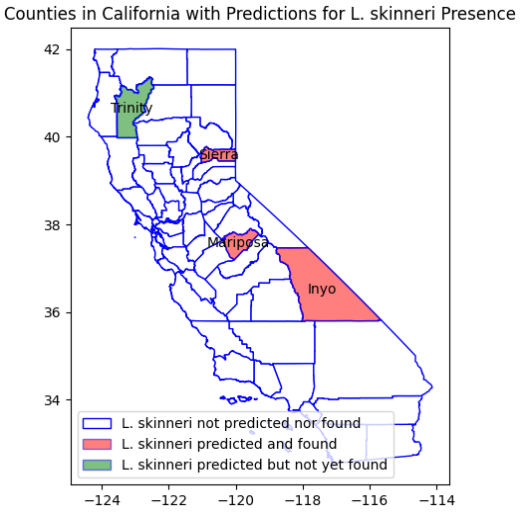
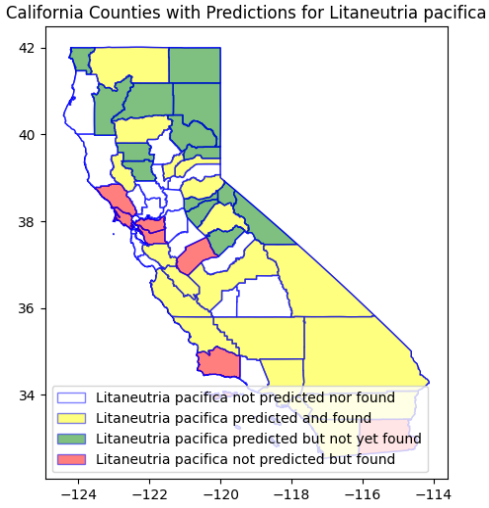



***Figure 6:*** *Map of decision tree's predictions for L. skinneri presence per county*

For species *L. pacifica* and *L. ocularis,* the model's predictions for presence in each county can be seen below (Figures 7 and 8).



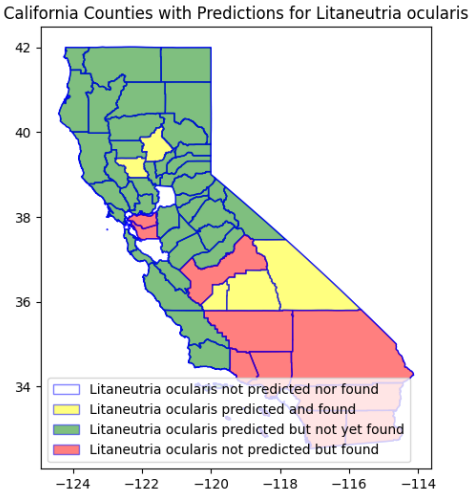




***Figure 7:*** *Predictions for L. skinneri presence per county* ***Figure 8:*** *Predictions for L. ocularis presence per county*

Introducing non-native mantis species to California in a short amount of time can harm an ecosystem and dramatically change the landscape for many native inhabitants of the land. To determine whether non-native species were taking over areas that were previously dominated by native species, we wanted to see if we could build a classifier to distinguish between native and non-native species by looking at their location (longitude, latitude), the climate around them (represented by the average maximum temperature, average minimum temperature, and

precipitation level of the county), as well as the human population density surrounding them. Our dataset had a fairly even split between native versus non-native species (2222 native records versus 2218 non-native records to be exact), so a null model would have a classification accuracy of around 50%.With two distinct classes, we decided that a Support Vector Machine would be an appropriate choice to build a classifier. After splitting up the data into train and test sets, we trained two different types of SVMs: one with an RBF kernel and one with a linear kernel. The idea behind this was that since a linear kernel simply tries to find a plane splitting the two classes, it would be easily interpretable and help simulate drawing an actual real boundary between native and non-native species, while the RBF kernel model, because of its non-linear decision boundary, would represent the optimal model we could create but might not translate to making much sense for real-world use. After using k-fold cross validation to assist with tuning the hyperparameters (parameters C and Gamma for the RBF kernel model, parameter C for the linear kernel model), we obtained the following confusion matrices representing the test results on both models (Figure 9 and 10):

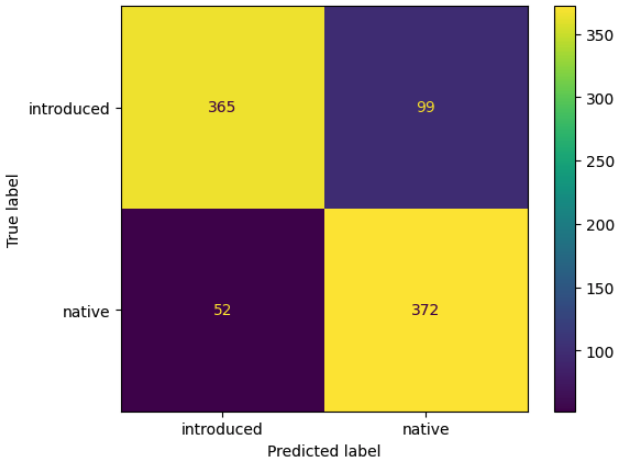



***Figure 9:*** *RBF kernel SVM*

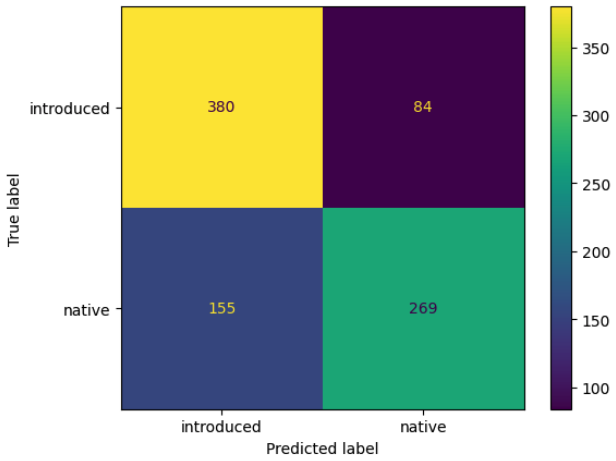



***Figure 10:*** *Linear kernel SVM*

As expected, the RBF kernel SVM outperformed the linear kernel SVM, with respective test accuracies of 0.83 and 0.73.

## Conclusion

From our decision tree results, we recommend CIBI to sample more in Trinity county for *L. skinneri*. If additional resources are available, sample in Sierra, Mariposa, and Inyo counties as well. For *L. pacifica,* focus sampling in Northern California and for *L. ocularis* in central and

Northern California. If resources are very limited, Trinity county would be a good place to look for all three. For future analysis, pre- and post-pruning should be tested to see if better fitting models can be developed for other native species to make suggestions to CIBI for where sampling should be conducted for the other 6 native species. Further analysis should be conducted on the biological reasons why parameters with high feature performance predict presence of these species which will give insights into their physiology and ecology.

From our SVM results, we can conclude while some distinctions can be made between native and non-native species using either kernel, there are enough samples that were misclassified to suggest that there are native and non-native species cohabiting the same environments, so it would be useful to gather more observations to confirm that if native and non-native species are cohabiting the same ecosystem, and what effect this might have on not only the mantises themselves through competition but also other animals in the ecosystem.

To fine tune future analyses, we can repeat all methods at the census block or tract level as the key for more evenness between entries. This will increase the size of data greatly and make analysis more computationally expensive, but would further improve our suggestions for CIBI.

**References**


1) *Database of Global Administrative Areas (2022). GADM database of Global Administrative Areas, version 4.1. [online] URL: https://gadm.org.*
2) *Griffith, Glenn E., et al. "Ecoregions of California." US Geological Survey Open-File Report 1021 (2016): 1-45.*
3) *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.*
4) *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.*
5) *Rosenberg, David M., H. V. Danks, and Dennis M. Lehmkuhl. "Importance of insects in environmental impact assessment." Environmental management 10 (1986): 773-783.*
6) *Telenius, Anders. "Biodiversity information goes public: GBIF at your service." Nordic Journal of Botany 29.3 (2011): 378-381.*
7) *Unidata, (2023): Integrated Data Viewer (IDV) version 6.2u1 [software]. Boulder, CO: UCAR/Unidata. http://doi.org/10.5065/D6RN35XM*
8) *U.S. Census Bureau (2020). Total Population, 2020 American Community Survey 5-Year Estimates Detailed Tables. Retrieved from [https://data.census.gov/](https://data.census.gov/)*