Mosquito Occurrence Prediction

Don't get annoyed by mosquitos anymore!

Christine Cai, Kyoko Kurihara, Robert Bonglamphone

Summary of questions and results

1. **How has the habitat of mosquitoes (specifically *Aedes aegypti)* changed in the US over time?**
   In the time period of 1904 - 2023, Aedes aegypti first occurred and developed in the south of the US and then migrated to the southwest of the US.

2. **Which species of mosquitoes are likely to occur in the US in which months?**
   In the time period of 2002-2022, *Culex* tarsalis had a higher occurrence than other species. However, all 3 species share a pattern of being likely to occur starting from April, peaking in July and sticking around until October.

3. **How would environmental changes such as global warming and human activities affect mosquito occurrence in one state of the US in the future (California)?**
   An increase in the human population can lead to an increase in the population and density of mosquitoes, but their habitats usually do not change significantly.

Motivation

More than 200 types of mosquitoes live in the US, but about 12 types spread serious diseases to people, such as Zika virus, dengue, and more[1]. Some of those mosquito-borne diseases' reservoirs are wild animals, which makes it impossible to eradicate the pathogens from the world even with effective vaccination. Therefore, it is important to prevent these diseases by avoiding high-risk areas and protecting yourself from mosquito bites. However, global warming is a contributing factor in changing the environment and areas where mosquitoes can populate. Being able to predict mosquito occurrence based on climatic factors

[1] Mosquitoes in the United States

can teach us about how and where we should manage mosquito populations to reduce the risk of disease.

In our project, we choose **three common species** of mosquitoes that can spread germs (table 1) and analyze their occurrence across the US over time. Our goal is to reveal where and when mosquitoes are likely to occur for each species. In addition, we want to predict mosquito occurrence in the future, considering environmental changes and human activities.

Table 1. Three common species of mosquitoes in the US that can spread germs.

Species	Mosquito-borne disease	Habitat
Aedes aegypti	Dengue virus Yellow fever virus Chikungunya virus Zika virus[2]	Tropical, subtropical, and in some temperate climates. Live near and prefer to feed on people, they are more likely to spread these viruses than other types of mosquitoes.[3] Along Southwest US to Southeast US
Culex tarsalis	St. Louis Encephalitis (SLE) Western Equine Encephalitis (WEE)	Almost every environment in the U.S.[4] Commonly seen in California
Anopheles quadrimaculatus	Malaria	Near aquatic habitats.[5] Primarily in the eastern part of the country, from the East Coast to the Texas Panhandle The highest densities of A. quadrimaculatus are found in the southeastern United States.

Dataset

The mosquito occurrence datasets are from GBIF. GBIF stands for Global Biodiversity Information Facility, an international network and data infrastructure funded by the world's governments.[6] This dataset holds the occurrence record for hundreds of millions of species of

[2] Aedes aegypti - Factsheet for experts.

[3] Potential Range of Aedes aegypti and Aedes albopictus in the United States, 2017 | Mosquitoes | CDC

[4] ADW: Culex tarsalis: INFORMATION.

[5] ADW: Anopheles quadrimaculatus: INFORMATION.

[6] What is GBIF?

life on Earth. Occurrences are collected from various sources including museum specimens from the 18th and 19th century to DNA barcodes and smartphone photos.

In addition, we plan to use supplemental datasets such as temperature, rainfall, and population change over time.

Table 2. The summary of the dataset we will work on.

Dataset	**URL**
Aedes aegypti occurrence	https://doi.org/10.15468/dl.uk36y3 Zip file: zip file
Culex tarsalis occurrence	https://doi.org/10.15468/dl.da59sj Zip file: zip file
Anopheles quadrimaculatus occurrence	https://doi.org/10.15468/dl.y9xydv Zip file: zip file
Temperature	City time series | NOAA Eueka: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00024213/tavg/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 Fresno: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USH00043257/tavg/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 Los Angeles: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00023174/tavg/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 Sacramento: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00023232/tavg/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 San Diego: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00023188/tavg/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 San Fransisco: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00023234/tavg/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020

Rainfall	City time series | NOAA Eureka: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00024213/pcp/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 Fresno: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USH00043257/pcp/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 Los Angeles: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00023174/pcp/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 Sacramento: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00023232/pcp/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 San Diego: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00023188/pcp/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020 San Fransisco: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00023234/pcp/all/1/1895-2023.csv?base_prd=true&begbaseyear=1991&endbaseyear=2020
Population	https://dof.ca.gov/forecasting/demographics/estimates/
CA county boundary	https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.2018.html#list-tab-CUBOSPWGA8KDWFHBTN

Method

1. How has the habitat of mosquitoes changed in the US over time?
    a. Read in the dataset occurrence.csv using pandas. Load the map of the US using geopandas.
    b. Filter dataframe where “countryCode” column is “US”. Split the data by four groups, where each group represents the occurrences of *Aedes aegypti* (Yellow fever mosquito) in 30 years.

c. Based on the map of the US, plot dots to represent the occurrences of *Aedes aegypti* to show the change of habitat over time. For every 30 years starting in 1904:
    i. Using zip and Point type from shapely.geometry, create points according to latitude and longitude columns in the dataset.
    ii. Store the points we created in a new column of the original dataset.

d. Using subplots, create a 2x2 plot. Plot the US map on each of the grid and plot points that represent occurrences of *Aedes aegypti* accordingly*.* Set the background color of the US map to gray and add titles for each graph.

2. Which species of mosquitoes are likely to occur in the US over time?

a. Read in dataset occurrence.csv for one mosquito species using pandas
b. Filter dataframe where “countryCode” column is “US”
c. Fillna values with 1 for “individualCount” column is NA
d. Sum the total of “individualCount” by month per year (Groupby “year”, “month”)
e. Filter “year” for 2002 to 2022
f. Repeat for steps **a-f** for each mosquito species dataset
g. Merge all mosquito species data into one DataFrame
h. Using plotly, plot a line graph from all mosquito species data, setting x-axis to “month”, y-axis to “individualCount” and color to “species”.
i. Add plotly’s animation frame bar to “year”, to see occurrence data for any year
j. Plot a line graph from *Aedes* species, setting x-axis to “month”, y-axis to “individualCount”, and color to “year”

3. How would environmental changes such as global warming and human activities affect mosquito occurrence in California in the future?

a. **Clean up mosquito occurrence datasets:** Read mosquito occurrence CSV files into DataFrames. Extract occurrences that are only in the US. Fill up NaN values in the “individualCount” column with 1. Drop any rows that contain Nan values.
Functions: `get_path`, `to_int`, `get_df_m`

b. **Generate a GeoDataFrame of mosquito occurrence in California:** Extract occurrences that are only in California using `filter_ca` function. Get the geometry with the longitude and latitude from the dataset. Return a GeoDataFrame of mosquito occurrence in California. If there is no occurrence in California, print “No occurrence in Califonia”.
Functions: `filter_ca`, `get_geometry`, `ca_geomosquito`

c. **Merge the mosquito GeoDataFrame and California county shape file:** Read the California county boundary shape file. Use the gdf.sjoin function to merge the shapefile with the mosquito occurrence GeoDataFrame. Remove unnecessary columns from the merged GeoDataFrame.

Functions: `get_map_ca`, `ca_occurrence`

d. **Clean up temperature and precipitation datasets:** Read temperature and precipitation CSV files for each city into separate DataFrames. Extract only the necessary columns from each DataFrame. Combine these DataFrames into one DataFrame with columns for each city's temperature and precipitation. Save the resulting DataFrame to the results folder. (Cities: Eureka, Fresno, Los Angeles, Sacramento, San Diego, San Francisco)
Functions: `get_path`, `get_df_t`, `get_df_p`, `generate_city_df`

e. **Clean up population datasets:** Read population Excel files into DataFrames. For the population data from 1947 to 1970, reassign column names and drop unnecessary columns. For the 1970-1979, 1980-1989, 1990-1999, 2000-2010, 2010-2019, 2020-2022 Excel files, change the format of the DataFrame so that the first column has the county and the second column has the population. Fill up the following Nan with each county name by looking at the location of Nan for the 1st column. If the county name is wrong, fix it. Drop any rows with Nan value. To make the formats of 1947-1970 and 1970-2022 the same, change the 1970-2022 DataFrames formats by writing new DataFrames. Merge all files into one population DataFrame over the years by county. Save the resulting DataFrame to the results folder.
Functions: `get_path`, `get_df_pop`, `pop50`, `column_name`, `clean_up_pop_df`, `change_style`, `combine_pop_df`

f. **Merge all datasets:** Remove the mosquito occurrence data before 1947 and after 2022. Create new columns, "population", "temperature", and "precipitation", in the mosquito occurrence GeoDataFrame. Fill up the population column with the value from the population DataFrame corresponding to the county and year of the mosquito GeoDataFrame. Fill up the temperature and precipitation columns with the values from the city DataFrame corresponding to the year, month, and capital of the area that the county belongs to (Table 3). Drop unnecessary columns. Drop NaN values. Return one GeoDataFrame with all necessary data.
Functions: `get_capital`, `add_0`, `merge_all_data`

Table 3. Groups of counties (Capital: **Bold**)

Group 1: Counties near Eureka (Northern California)	**Humboldt**, Del Norte, Lake, Mendocino, Modoc, Shasta, Siskiyou, Tehama, Trinity
Group 2: Counties near Fresno (Central California)	**Fresno**, Inyo, Kern, Kings, Madera, Mariposa, Merced, Mono, Tulare, Tuolumne
Group 3: Counties near Sacramento (Northern California)	**Sacramento**, Alpine, Amador, Butte, Calaveras, Colusa, El Dorado, Glenn, Nevada, Placer, Plumas, Sierra, Solano, Stanislaus, Sutter, Yolo, Yuba

Group 4: Counties near San Diego (Southern California)	**San Diego**, Imperial
Group 5: Counties near San Francisco (Bay Area)	**San Francisco**, Alameda, Contra Costa, Marin, Napa, San Mateo, Santa Clara, Santa Cruz, Sonoma
Group 6: Counties near Los Angeles (Southern California)	**Los Angeles**, Monterey, Orange, Riverside, San Benito, San Bernardino, San Joaquin, San Luis Obispo, Santa Barbara, Ventura

g. **Train a machine learning model to predict mosquito occurrence:** Trains a Random Forest regressor model with features of “LSAD”, “ALAND”, “AWATER”, “month”, “year”, “population, “temperature”, and “precipitation” to predict labels of “individualCount”, “decimalLatitude”, and “decimalLongitude” with a test size of 0.2. After training a model, compute and print a mean squared error. Pass a number from 1 to 30 to the tree’s depth hyperparameter “n_estimators” and returns a number that gives the minimum mean squared error. Plot the n_estimators vs. the mean squared error. Use the optimized depth after this. Functions: `prediction`, `decide_depth`

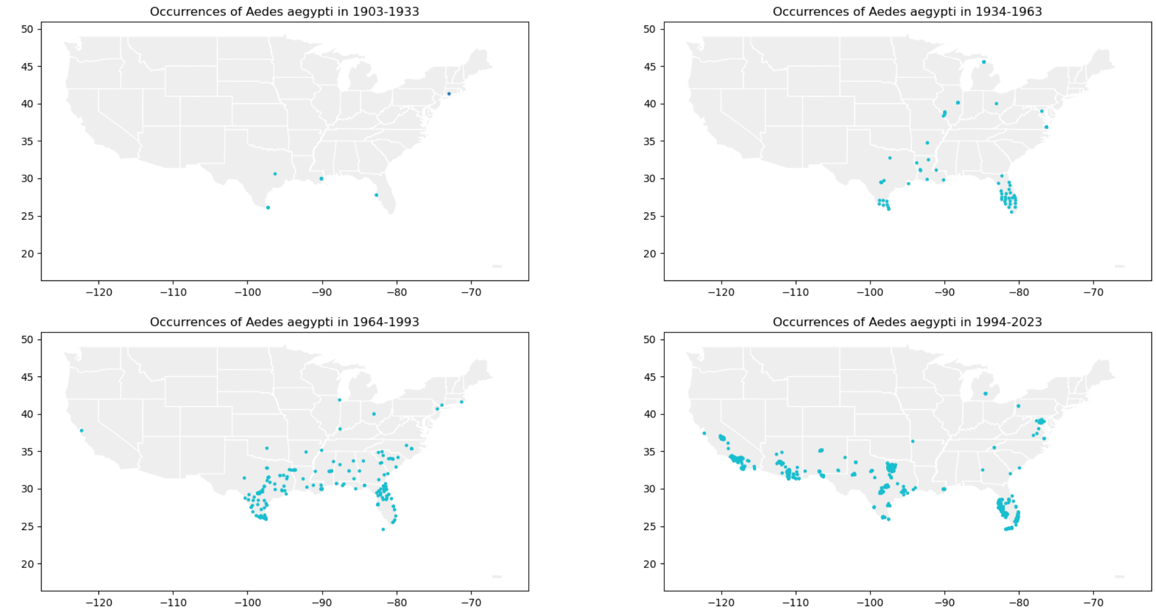
h. **Plot mosquito occurrence:** Plot the total past mosquito occurrence based on the longitude and latitude on a California map with a marker size of “individual count”. Plot the predicted mosquito occurrence on the test set in the same way. Finally, by returning features only when the return_features parameter is True in the prediction function, get features and change its “population”, “year”, “temperature”, and “precipitation” values if you want. Plot the predicted mosquito occurrence by the Random Forest regressor model on the new features. Funcitons: `prediction`, `plot_prediciton`

i. **Compute the above for three mosquito species.** The same process is applied to all three mosquito datasets. If there have been no occurrences in California in the past, print out “No occurrence in Califonia.”

Results

1. How has the habitat of mosquitoes (specifically Aedes aegypti) changed in the US over time?

We generated 4 plots to show the change of habitat of Aedes aegypti in the US. Each plot represents the occurrences of Aedes aegypti in a time range of 30 years. We chose 30 years as the time range because mosquitos migrate slowly over the years. With multiple attempts with various time ranges, we found that a 30-year time range works


Occurrences of Aedes aegypti in 1903-1933

Occurrences of Aedes aegypti in 1934-1963

Occurrences of Aedes aegypti in 1964-1993

Occurrences of Aedes aegypti in 1994-2023

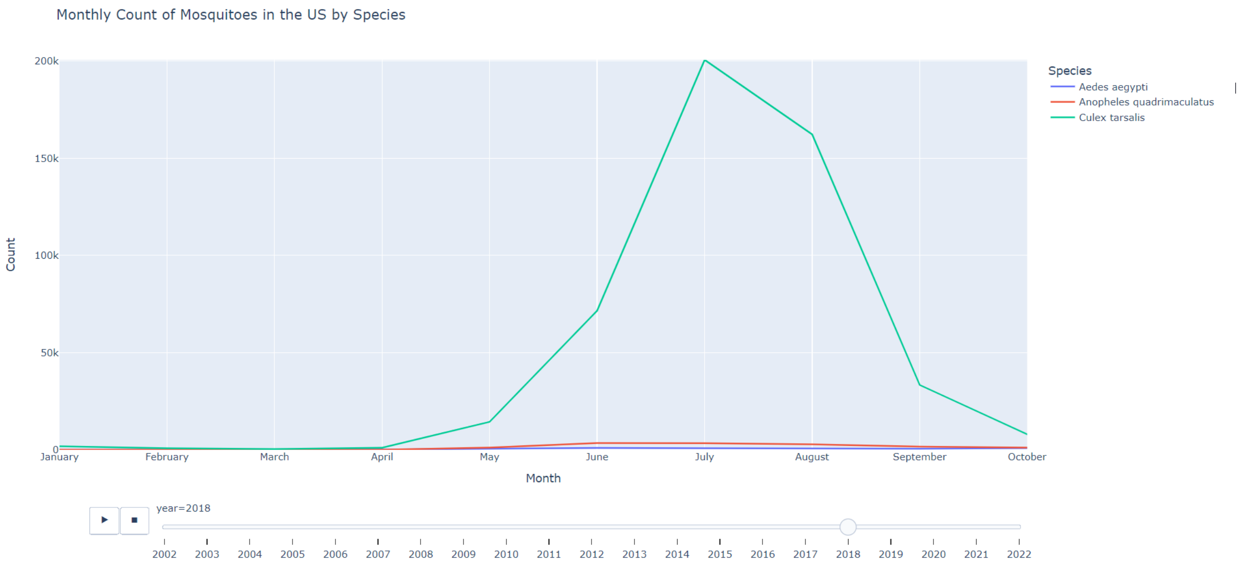
best to show the migration path of Aedes aegypti.

As shown in the plots, there were few occurrences of Aedes aegypti from 1903 to 1933, mostly located in the far south of the US, with one occurrence in the far east. During the next thirty years from 1934 to 1963, the occurrences of Aedes aegypti started increasing. Most of them occurred in Florida and spread throughout the northeast. For the next thirty years from 1964 to 1993, the occurrences of Aedes aegypti increased a lot in the southern states, including Florida, South Carolina, Georgia, Alabama, Mississippi, Louisiana, Arkansas, Oklahoma, and Texas. During the period between 1994 and 2023, the total occurrences of Aedes aegypti did not change much. But surprisingly, most Aedes aegypti moved from the southern states to the southwestern states, including Texas, New Mexico, Arizona, and California, while Florida was still a high-occurrence state of Aedes aegypti.

(Note: To make the graph more visible, we did not count the occurrences of Aedes aegypti in Hawaii and Alaska.)

2. Which species of mosquitoes are likely to occur in the US in which months?

Based on the graphs, *Culex tarsalis* has a higher occurrence than the 2 other species: *Aedes aegypti* and *Anopheles quadrimaculatus*. Therefore, we can infer they're more likely to increase in population. As seen with the chart below, the biggest occurrence out of this set of species was *Culex tarsalis* in July, 2018 with a count of about 200,000. This was about 5882% more than the next highest species count of the *Anopheles quadrimaculatus* with a count of about 3400. This datapoint is highly questionable and would need to be invalidated. In further investigation, the entries that were being collected that month were based out of Salt Lake, Utah. It's apparent that one researcher may count an occurrence differently from another. There's a bit of inconsistency in how count was being calculated by different researchers across the States and how frequently this data was being collected is unclear.



Since the data that was collected was inconsistent, we've chosen the time frame to analyze the mosquito population to be between 2002 and 2022. This 20-year time span had more data readily available to understand trends on which months mosquitoes are occurring more frequently.

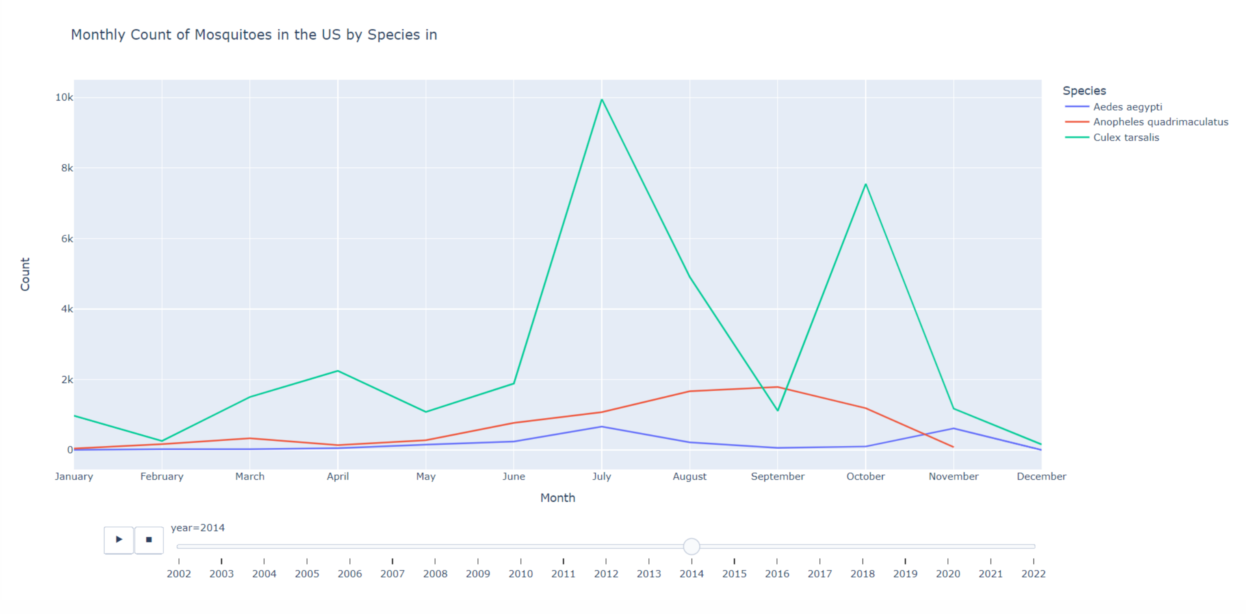
The plot below shows monthly counts for the year 2014 for all 3 species and the data could be interpreted more easily without *Culex* tarsalis skewing the y-axis on counts. Although this plot shows the monthly counts for all species in one specific year, there's a noticeable pattern among the species throughout the 2002-2022 timespan for when they start to increase in number and then drop off.

Aedes aegypti start to increase in count in the month of April and then peak in July.
Anopheles quadrimaculatus, tend to steadily rise in count starting from April and peaking out in September.
As for *Culex tarsalis*, their species tend to follow a similar pattern to that of the *Aedes aegypti* where their population rises in springtime around May before peaking out in the month of July. Their count continues to stay somewhat high all the way into October.

Throughout this time period of 20 years, these trends seem to be consistent.

Some likely reasons as to why these months are the best for mosquito populations is because the climate is ideal for them to thrive in. Rising temperatures starting in spring and lasting until early fall, mosquitoes can create breeding grounds in those areas. This is possibly why a lot of data is sourced from areas near Los Angeles because of its urban environment and hot weather. *Aedes aegypti*, for instance, would have an increase in numbers because of the hurricane season that takes place in States such as Florida.

Some exceptions to this was when the count data was low for certain years such as 2002-2005 and 2020-2022. The count across all species was anywhere below 100 when it should be higher than 500 counts. One possible reason for this decrease in count data could have been related to the quarantine policy from the pandemic and a lack of active researchers in the early 2000's.



3. How would environmental changes such as global warming and human activities affect mosquito occurrence in California?

   The mosquito occurrence data for *Aedes aegypti* in California contains individual counts, decimal longitude, latitude, month, and year of the occurrence. The occurrence location is classified into one of the counties in California. The mosquito occurrence data is combined with environmental data (temperature and precipitation) and human population data corresponding to its county, month, and year. Table 4 shows the head of the GeoDataFrame with all collected data for *Aedes aegypti*.

Table 4. Example of features and labels for *Aedes aegypti*

NAME	LSAD	ALAND	AWATER	individualCount	decimalLatitude	decimalLongitude	month	year	population	temperature	precipitation
Kern	06	21062456129	78764641	1.0	35.37315	-119.018713	10.0	2020	905241.0	71.6	0.0
Ventura	06	4771968316	947365005	1.0	34.277651	-118.743227	9.0	2021	838630.0	68.4	0.0
Alameda	06	1909598013	216923745	1.0	37.776419	-122.276015	10.0	1983	1160500.0	64.3	0.1
Alameda	06	1909598013	216923745	1.0	37.77769	-122.277342	10.0	1983	1160500.0	64.3	0.1

After removing the "NAME" column from the table, the columns "LSAD", "ALAND", "AWATER", "month", "year", "population, "temperature", and "precipitation" are used as features to train a Random Forest regressor model. Random Forest is well-suited for classification tasks with a large number of features. The model is selected with the purpose of including all features to make predictions since the impact of each feature is unclear. The model predicts labels of "individualCount", "decimalLatitude", and "decimalLongitude".

The tree's depth hyperparameter "n_estimators" is optimized after the trials from 1 to 30 (Fig 3). The predicted "individualCount", "decimalLatitude", and "decimalLongitude" of mosquito occurrence are plotted on a Califonia map (Fig 4). The left figure shows the total recorded occurrence in Califonia. The size and color of each dot correspond to the value of "individualCount". The center figure visualizes the predicted mosquito occurrence on the test feature set. The mean squared error is 8.620183984509008, which is reasonable since we are dealing with large values for longitude and latitude. The right figure shows the prediction on the new features, which is defined by the following code:

—

```
new_features1['population'] = new_features1['population'] + 10000
new_features1['year'] = new_features1['year'].apply(int) + 50
new_features1['temperature'] = new_features1['temperature'] + 5
new_features1['precipitation'] = new_features1['precipitation'] + 1
```

—

The new features assumes that there is an increase in population, temperature, precipitation, and a shift towards 50 years. Compared to the other figures, the prediction on the new features shows a large number of individual counts, indicating an increase in the mosquito population and density. The habitat of *Aedes aegypti* remains similar. Mosquito-borne diseases are more likely to spread when mosquito density is high. The *Aedes aegypti* mosquito, which prefers to feed on people in tropical or subtropical climates[7], is known to carry diseases such as Dengue fever and Yellow fever. If the population and density of this mosquito increase in the city in the future, the risk of mosquito-borne diseases spreading will be high.

[7] Potential Range of Aedes aegypti and Aedes albopictus in the United States, 2017 | Mosquitoes | CDC

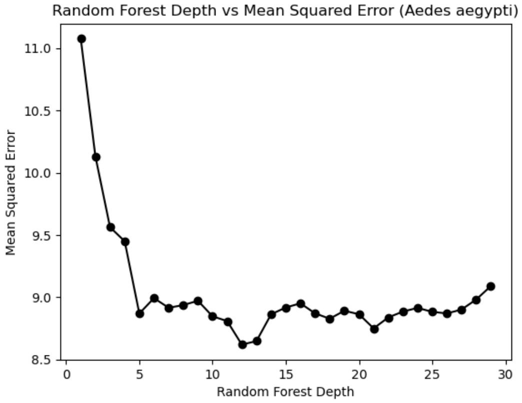


Fig 3. Optimization of the hyperparameter (*Aedes aegypti*)

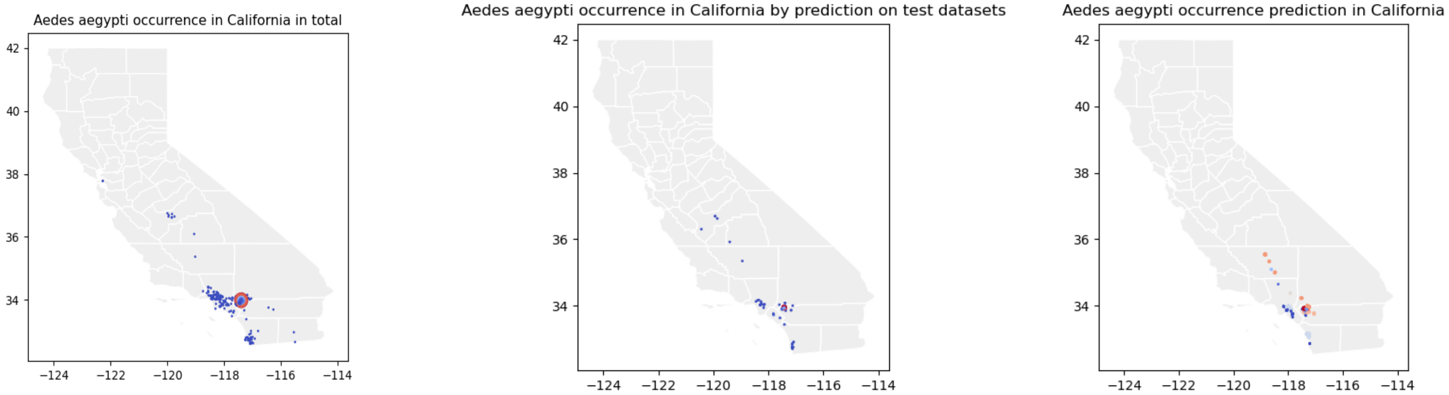

Aedes aegypti occurrence in California in total
Aedes aegypti occurrence in California by prediction on test datasets
Aedes aegypti occurrence prediction in California
42
40
38
36
34
−124
−122
−120
−118
−116
−114

Fig 4. Prediction plots for *Aedes aegypti*

The second mosquito, *Anopheles quadrimaculatus*, has not been recorded in California in the past.

For the third mosquito, *Culex tarsalis*, the same procedure is applied to merge datasets and train a Random Forest regressor model (Table 5, Fig 5, Fig 6). This results in a high mean squared error of 2236.1663013894304. The reason for the high error is that the model is missing occurrences of this mosquito in a high-latitude area near Eureka. In order to increase the accuracy of predictions across the entire state of California, a larger sample size is required. If we could ignore the error, the model predicts that the population of *Culex tarsalis* is concentrated in the Central and South areas of California, even when both population and environmental factors change. This suggests the government can implement long-term policies to protect against malaria, a mosquito-borne disease carried by *Culex tarsalis*, for people living in these areas.

Table 5. Example of features and labels for *Culex tarsalis*

NAME	LSAD	ALAND	AWATER	individualCount	decimalLatitude	decimalLongitude	month	year	population	temperature	precipitation
Napa	06	1938114186	104300794	1.0	38.602527	-122.435787	8.0	1974	88900.0	64.3	0.0
Napa	06	1938114186	104300794	1.0	38.602527	-122.435787	8.0	1974	88900.0	64.3	0.0
Napa	06	1938114186	104300794	1.0	38.602527	-122.435787	9.0	1974	88900.0	63.1	0.0
Napa	06	1938114186	104300794	1.0	38.602527	-122.435787	9.0	1974	88900.0	63.1	0.0

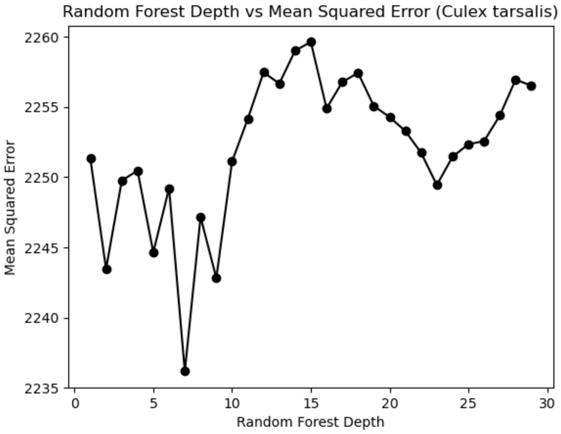


Fig 5. Optimization of the hyperparameter (*Culex tarsalis*)

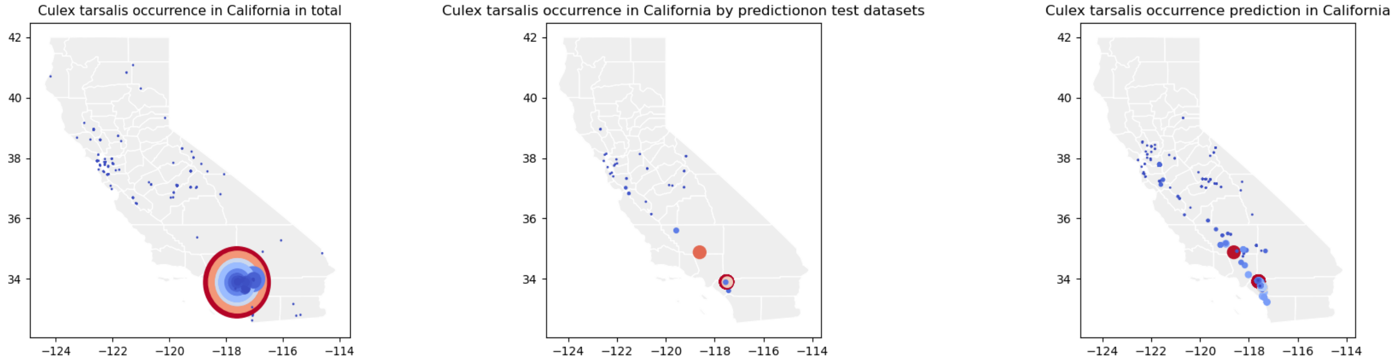

Culex tarsalis occurrence in California in total
Culex tarsalis occurrence in California by predictionon test datasets
Culex tarsalis occurrence prediction in California
42
40
38
36
34
−124
−122
−120
−118
−116
−114

Fig 6. Prediction plots for *Culex tarsalis*

Impact and Limitations

Positive Impacts

Our analysis helps visualize and build a narrative for the everyday person to be able to see how these mosquito populations are thriving in certain regions of the States especially in urban areas. Disease carried by mosquitoes do take a toll on public health and the numbers can help show people to stay vigilant and practice mosquito prevention. Ultimately, the most valuable takeaway from our results is not necessarily the increase of occurrences of mosquitoes over time, but the need for transparency for mosquito data. Our results show a huge bias for mosquitoes to occur in certain places of the states, but that's only because of the data that is readily available. This shows that if we had access to better data. Mosquito control boards do collect and hold occurrence data, but it's not usually open to the public. Building awareness of this problem would help strike a narrative and push for better occurrence and so scientists and researchers alike would be able to study the ecology of mosquitoes and better predict the spread of mosquitoes and their impact.

Negative Impacts

Someone might misconstrue and take our findings as face value to believe that mosquitoes will increase and migrate as shown in our prediction models. People could draw an immediate conclusion to say that the maps represent areas where mosquito-transmitted diseases are likely to occur when really the maps are only meant to show mosquito migration patterns.

Mosquito populations are one factor residents might consider when living in a certain area. If presented with data about mosquitoes thriving in their communities, then it could plummet a homeowner's property value and people would move to a place with less mosquitoes (Yong). This is actually one reason why mosquito control boards would be hesitant to make their occurrence data public because this type of analysis could trigger people to sue them for not sufficiently controlling the population.

Limitations

Our biggest limitation was the lack of variety in places where *Aedes aegypti* and *Culex tarsalis* species samples were collected. Both *Aedes aegypti* and *C*ulex tarsalis data was primarily collected in Riverside, CA which is within 60 miles from Los Angeles. Of course, our prediction model would be biased to see that area as a massive breeding ground for mosquitoes. There are a few outliers in northern parts of California such as Clear Lake, CA where occurrence data was found in those regions too. Since our data mostly consisted of occurrences near Los Angeles, it hugely underrepresents other populations where mosquitoes can be found.

In looking at the map of the United States as a whole, one could conclude that *Aedes aegypti* were first spotted in states such as Texas and Florida and have increased in numbers over the century while also spreading into California. Once again, we're only limited by whoever and where they're collecting samples from. There are only 2 counties in Florida where data for *Aedes aegypti* was being collected from: Manatee and Lee. It overlooks areas where mosquitoes are known to live like the Florida Keys. Although the maps signify that's a westward migration trend, people shouldn't believe that to be 100% true. One possible reason for this is that simply, nobody was collecting samples in California in the earlier time periods.

Finally, data accuracy also limited us to what the true number of occurrences were. Standards were placed to properly document and collect information about the location and time of the occurrence, but many of the researchers who sampled mosquito data counted the number of mosquitoes differently. In the line plot against all mosquito species, the largest occurences for *Culex tarsalis* took place in July, 2018, and most of these counts were in the thousands based out of Salt Lake, Utah. Within that same time period, places such as San Antonio, Texas had no entries because collecting the number of occurrences is not entirely enforced. If there’s a ridiculous spike in occurrences, then the likely scenario is that a researcher was using a different method to count their samples.

Challenge Goals

1. **Machine Learning**: We trained a Random Forest regressor model to predict the latitude, longitude, and individual counts of mosquito occurrence by optimizing the hyperparameter "n_estimators." To determine the optimal hyperparameter for the tree's depth, the function "decide_depth()" in the mosquito.py file passes a number from 1 to 30 to the hyperparameter and returns the number that results in the minimum error. Additionally, the results are visualized by plotting the predicted longitude and latitude on a map. The values of features can be edited to enable predictions in hypothetical scenarios. By using a model that was not taught in class, optimizing its performance, visualizing the results, and enabling manual feature editing, we have achieved this challenge goal.

2. **New library:** With seaborn and matplotlib, you can plot and save images, but we realize that they’re static and if you wanted to extract a value for a certain plot, you’d need to eyeball it. So, we wanted to find a way to make our plots more interactive and easier to see the data when hovering over certain data points. Our goal was to plot all occurrences of each species on a monthly basis on a single graph. In addition, we wanted to give the plot user input for the year, so that we could visualize how occurrences were changing over time. This led us to implement plotly as a new library. With the mouse cursor, you could hover whichever plot on the graph and it can provide more details and also toggle on and off different species to view. Also, we added a slider widget that changes to the selected year for the occurrence data. Overall, it was necessary to read through the documentation and look through their examples to understand how it is used and can be incorporated into our plots.

3. **Multiple Datasets**: To predict mosquito occurrence and account for the impact of environmental changes and human activities, we combine several datasets using the merge_all_data function in mosquito.py. Specifically, we merge datasets on California county boundaries, mosquito occurrence, population by county over time, the average temperature in each area, and precipitation in each area, including DataFrames and GeoDataFrames. First, we combine the California county boundary dataset with the latitude and longitude of mosquito occurrence to classify occurrences by county. Next, we add population, temperature, and precipitation data to the DataFrame using county, years, and months as indexers. To obtain the necessary population, temperature, and precipitation data, we combine data files from different years into a single file before merging them with the occurrence data. Since we merged approximately 21 datasets for each species from different sources, we have achieved this challenge goal.

4. **Messy data**: The seven population datasets from different years cited from the Department of Finance in California state have different formats. To merge them into one DataFrame, the clean_up_pop_df function in mosquito.py changes their structure and removes unnecessary NaN values to handle each dataset's format. The column_name function helps to reassign column names. Since the population Excel file from 1947 to 1970 has a completely different style, the other six population datasets are reorganized by creating and editing a new DataFrame. After cleaning up the 1947-1970 population data using the pop50() function, we combine all population datasets in the combine_pop_df() function. Since this requires a lot of pre-processing, we have achieved this challenging goal by merging the population datasets.

Work Plan Evaluation

1. Data clean up and additional dataset finding
    a. Estimated: (3 hours) Read in the dataset successfully.
    b. Actual: (6 hours) Because we used multiple datasets and the seven population datasets focusing on California are messy. We spent a lot of time writing clean-up functions for our datasets. Also, because we split our tasks by group members, we accidentally did some overlapping work in reading the datasets, which led to extra time for this part.
2. Research questions that are listed in *Summary of Research Questions* (Estimated: 8 hours)
    a. Research question 1
        i. Estimated: (2 hours) We will extract data using pandas and create a plot to visualize the change over time.
        ii. Actual: (8 hours) Even though this is the easiest question, it still took more than the estimated time to complete because we were not proficient in plotting geographic plots. We spent some time reviewing material learned in class before we started exploring this question. Around 3 hours of the total time was spent on testing. Because we used Google Data Studio to test our results, which is a brand-new tool, it took more time than we expected to get familiar with its features.
    b. Research question 2
        i. Estimated: (2 hours) Since question 2 is simply extracting data using pandas, it will take the least time among all the questions.
        ii. Actual: (10 hours) It turns out that we decided to use a new library *plotly* to generate interactive plots, which made this task much more time-consuming than we expected. Also, the plots were giving unusual results where the *Culex tarsalis* species had a much higher occurrence than the other species. So, it took more time truly understanding the data and developing a rationale for why these outliers exist.
    c. Research question 3

        i. Estimated: (4 hours) Since we need to find an appropriate model for machine learning, we might spend most of our time exploring this question.
        ii. Actual: (15 hours) This is the hardest question to answer. Just as with the other two questions, we also spent more time exploring this question. The most unexpected challenge meet in this question is the messy datasets. We spent most of our time cleaning up datasets and combining them. Also, we implemented machine learning techniques while exploring this question. Since there were two challenges in this question, it took us the longest time to finish this question.

3. Research challenging questions which are listed in the *Challenge goals*.
    a. Estimated: (6 hours) Find a way to combine multiple datasets. Find a new library to plot the US map. Find an appropriate model for the prediction. Plot predicted data from the model onto the US map.
    b. Actual: (20 hours) Based on our mentor's advice, we made subtle changes to our challenge goals. We decided to use a machine learning model that was not taught in class (a Random Forest regressor model). We did not use a new library to plot the US map. Instead, we used a new library to show the monthly count of mosquitos in the US. Also, cleaning up messy datasets is an unexpected challenge that we met and conquered.

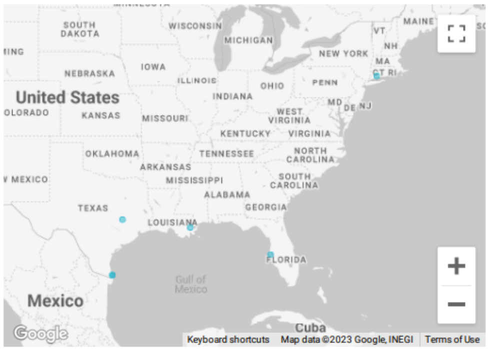
Testing

1. How has the habitat of mosquitoes (specifically Aedes aegypti) changed in the US over time?

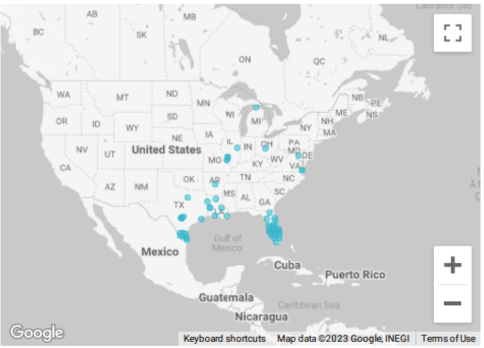
To test that the graphs that we plot for this research question are correct, we uploaded the original *Aedes aegypti occurrence* dataset to Google data studio. We used the built-in feature to plot bubble maps, which each represent the occurrences of Aedes aegypti in thirty years. For each of the bubble maps, we used the filter feature to filter down to the data we want. We first filter the location to "US", then we apply the time range of thirty years, and we also add a filter to exclude occurrences in Hawaii and Alaska. Here is the link to the bubble maps we created. Since the original dataset has two separate columns that each represent decimal longitude and decimal latitude, I added a new column using the built-in function that combines these two columns (Latitude, Longitude). We plot the bubble map with points by setting the new column as our location. I also set the column *eventDate* as the tooltip, so when you hover over the points, you will see the date of the occurrences of Aedes aegypti.

By comparing the bubble maps generated by Google data studio (first plot shown below) and the plots that we plot by Python (second plot shown below and shown in this section), we can see that the two plots are the same. Since the bubble maps generated by Google data studio are based on the unfiltered original dataset, we can conclude that the plot we create using Python is correct.

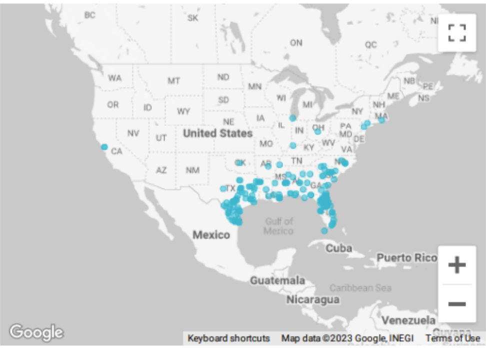
Occurrences of Aedes aegypti in 1903-1933



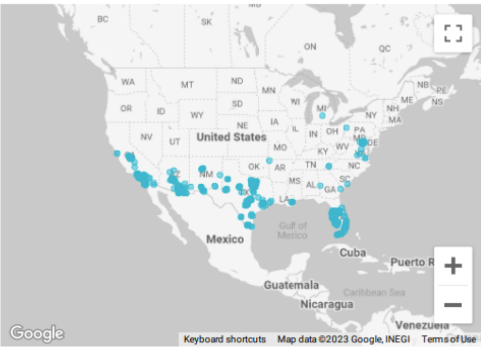
Occurrences of Aedes aegypti in 1934-1963



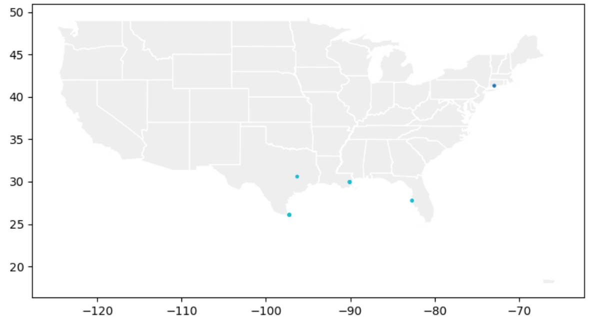
Occurrences of Aedes aegypti in 1964-1993



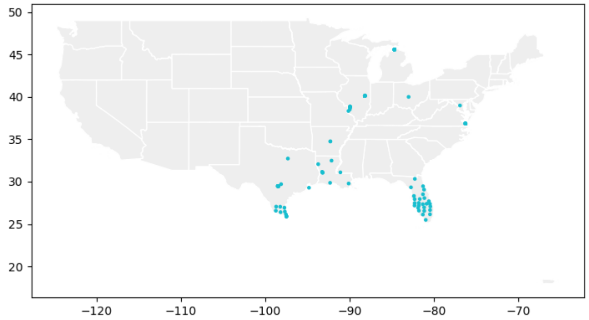
Occurrences of Aedes aegypti in 1994-2023



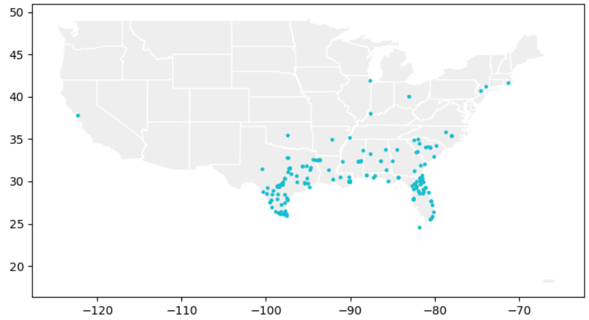
Occurrences of Aedes aegypti in 1903-1933



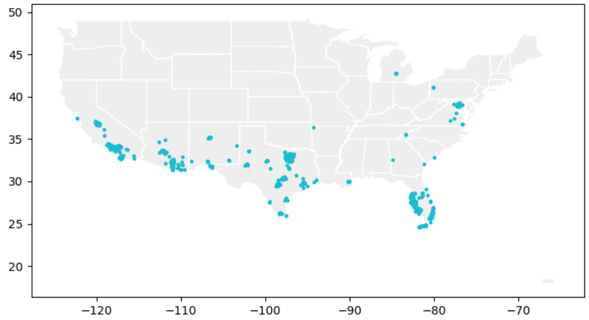
Occurrences of Aedes aegypti in 1934-1963



Occurrences of Aedes aegypti in 1964-1993



Occurrences of Aedes aegypti in 1994-2023



2. Which species of mosquitoes are likely to occur in the US over time?

   To test whether we were receiving the right dataframe where each species took all individual counts and added them all together per month for each year, we used an assert equals function. The value for the max monthly count was known and it was 200568. This happened in July 2018 for the *Culex tarsalis* species occurrence dataset. So, we used those filters on that data so that it can return the value of that month. It returned the dataframe and the value matched them exactly with 200568.

3. How would environmental changes such as global warming and human activities affect mosquito occurrence in California?

   Since we trained a machine learning model and visualized the predictions, it has been difficult to verify the generated plots' correctness. As a solution, we wrote test functions using assert statements in test.py to test whether the functions in mosquito.py return the expected results. We focused on testing the functions related to reading CSV files, merging DataFrames and GeoDataFrames, and machine learning.

   To test reading in CSV files to DataFrames, we checked the type of results and confirmed column names. The downloaded datasets for question 3 required a lot of pre-processing and careful data cleanups. We checked that the temperature and precipitation CSV files were cleaned up and merged properly. Combining seven population datasets will likely cause errors, so we also tested whether the merged data has the proper columns.

   For the prediction part, it is difficult to know the accuracy of the generated plots above the returned mean squared error. Therefore, we tested whether the prediction function returns expected values with different inputs. We conducted all the tests mentioned above and concluded that our code generated the expected results.

Collaboration

We worked on this project on our own. We referenced resources online to help us achieve our goals.

Work Cited

Yong, Ed. “The Case for Sharing All of America's Data on Mosquitoes.” *The Atlantic*, 24 August 2017, https://www.theatlantic.com/science/archive/2017/08/mosquito-data/537735/. Accessed 12 March 2023.

https://animaldiversity.org/accounts/Anopheles_quadrimaculatus/