A BLI and GenBank Metadata Study Integrating Terrain Classification of LandSat8 Images and Machine Learning Modeling to Determine the Impact of Landscape Variables on Mosquito Abundance

Shayan Joarder
Plainedge High School

Problem

- Global temperatures have increased by 1.3°C while water temperatures in the Long Island region have increased by 0.03°C (Rice 2014).
- Mosquitoes in general have been found to thrive in warmer climates.
- *Aedes albopictus*, the fifth most prevalent mosquito it is known for contracting Zika Virus and other mosquito-borne viruses (Bajwa 2018).

Research Goals

- Employ machine learning (ML) to identify variables of prevalence associated with mosquito distribution and prevalence.
- Use Partial Dependence Plots (PDP) which show non-linear dependence between a certain input variable and the response within the dependent variable.

Research Basis for Proposed Methodology

- PDP based protocol used to identify human footprints effects on animal extinction (Ramirez-Delgado *et al*., 2022)
- Machine learning to identify variables for mosquito prevalence → Lee *et al.* (2022) landscape and meteorological data within Seoul; Chen et al. (2019) socioeconomic variables within Baltimore
- Young *et al.* (2022) →identified landscape heterogeneity as a main variable for increase in *Aedes aegypti* specifically
- Use of ArcGIS and geographical services for landscape heterogeneity → (Murwira & Skidmore 2005)

Methods (Summary)

Mosquito Samples Metadata Collection → LandSat 8 Satellite Data Collection → Band Processing → Terrain Classification → Landscape Heterogeneity Calculation → Machine Learning Analysis

Methods – Collection and Processing of Sample Metadata and Satellite Band Data

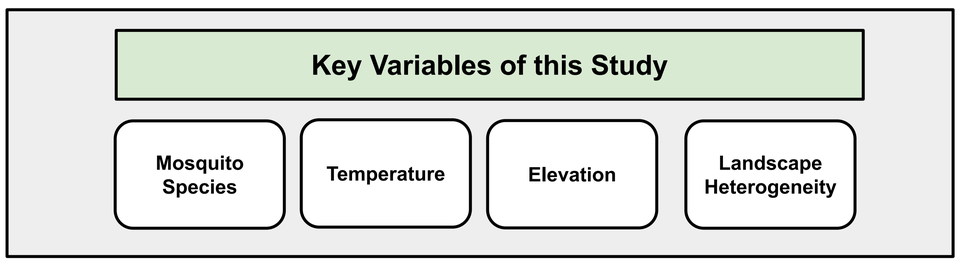


Figure 1. Key Variables Operationalized in the Study. The researcher decided to conduct a data science project that would apply sample data collected from previous Barcode Long Island projects along with data in GenBank. Mosquito species served as individual dummy variables (*Aedes albopictus, Aedes aegypti,* and *Culex pipiens)*. Temperature and elevation were interval data and derived using the coordinates provided for the sample by the BLI Sample Database and GenBank, Finally, landscape heterogeneity was calculated as a percentage through the use of terrain classification of LandSat 8 satellite data using the Semiautomatic Classification Plugin of QGIS. The final data analysis involved the creation of decision tree classification models and partial dependence plots with each individual mosquito dummy variable being the outcome variable and the remaining variables serving as the predictor variables.

Table 1

Mosquito Samples and Metadata Collected from Barcode LI Database and GenBank

Species	Coordinates (Latitude, Longitude)	Date Collected (year-month-day)	Temperature (Fahrenheit)	Elevation (Feet)	Landscape Heterogeneity
Aedes albopictus	30.630, -81.610	2021-09-29	75.08	28	Data Calculated Using QGIS
Aedes albopictus	40.869, -73.585	2022-10-05	56.77	217	
Aedes albopictus	1.401, 110.364	2019-09-27	84.00	68.90	
Aedes albopictus	15.086, -92.210	2020-09-12	77.00	4379.92	
Aedes albopictus	42.406, 18.641	2019-02-09	43.63	679.13	
Aedes aegypti	40.865, -73.598	2023-01-17	42.60	123	
Aedes aegypti	40.953, -72.904	2019-12-20	26.33	118	
Aedes aegypti	10.75, 78.69	2018-12-05	84.00	295.28	
Aedes aegypti	-3.934, 39.569	2016-12-05	82.00	482.28	
Aedes aegypti	33.03, 73.594	2016-10-24	82.00	908.79	
Culex pipiens	40.641, -73.968	2017-10-01	60.40	34	
Culex pipiens	40.864, -73.597	2023-01-04	53.79	131	
Culex pipiens	4.286, -2.639	2019-05-01	79.00	3733.60	
Culex pipiens	42.436, -2.597	2016-08-24	79.00	1525.59	
Culex pipiens	51.374, 0.791	2014-07-16	58.51	285.43	

Table 1. Mosquito Samples and Metadata Collected from Barcode LI Database and GenBank. I first identified mosquito specimens published to the Barcode Long Island Database and GenBank. For each sample, metadata was collected in the form of the coordinates (latitude, longitude) and date of collection. Coordinates were inputted into Weather Underground (wunderground.com) to derive the historic temperature at the time of sampling. Coordinates were also inputted into the website CalcMaps (calcmaps.com) to derive the elevation of the sample location. The column labeled 'Landscape Heterogeneity' (highlighted in yellow required the processing of LandSat8 data from the Earth Explorer database through the QGIS software.

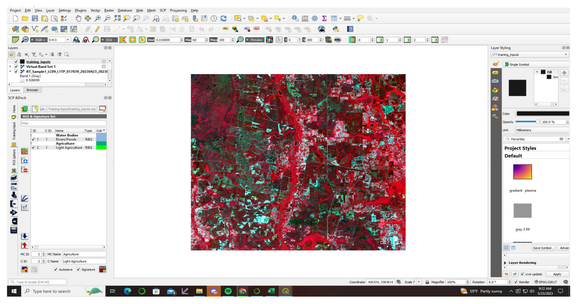
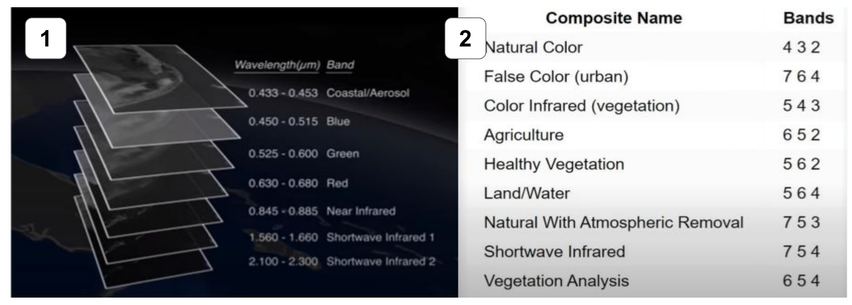


Figure 2. Satellite Data Preprocessing for Training Inputs. Following the protocol of GeoLabs and the SCP documentation, Landsat 8 data were imported into QGIS for preprocessing using the Semiautomatic Classification Plugin (SCP). Landsat data took the forms of different bands, each specific to a wavelength of light. Bands were clipped and modified for reflectance. Shown above is a false-color composite (RGB 5-4-3) image of Sample 1 (Florida). SCP enables unsupervised classification of regions of interest, which were manually selected by the user as macroclasses and classes.



Composite Name	Bands
Natural Color	4 3 2
False Color (urban)	7 6 4
Color Infrared (vegetation)	5 4 3
Agriculture	6 5 2
Healthy Vegetation	5 6 2
Land/Water	5 6 4
Natural With Atmospheric Removal	7 5 3
Shortwave Infrared	7 5 4
Vegetation Analysis	6 5 4

Figure 3. LandSat 8 Band Data and Composites. Following the protocol of GeoLabs and the SCP documentation, Landsat 8 data were imported into QGIS for preprocessing using the Semiautomatic Classification Plugin (SCP). Landsat data took the forms of different bands [1], each specific to a wavelength of light. Using SCP, bands were clipped and modified for reflectance. The appearance of the resulting composite layer could be changed by selecting which bands formed the composite [2]. The researcher explored different band compositions in order to identify the combination that best showed contrasts between different terrain features. Once the desired band composition was achieved, SCP enabled the classification of regions of interest, which were manually selected by the user as macroclasses and classes. Figure 1 and 2 are courtesy of GeoLabs on YouTube (https://www.youtube.com/watch?v=HKNS-wsc7Io).

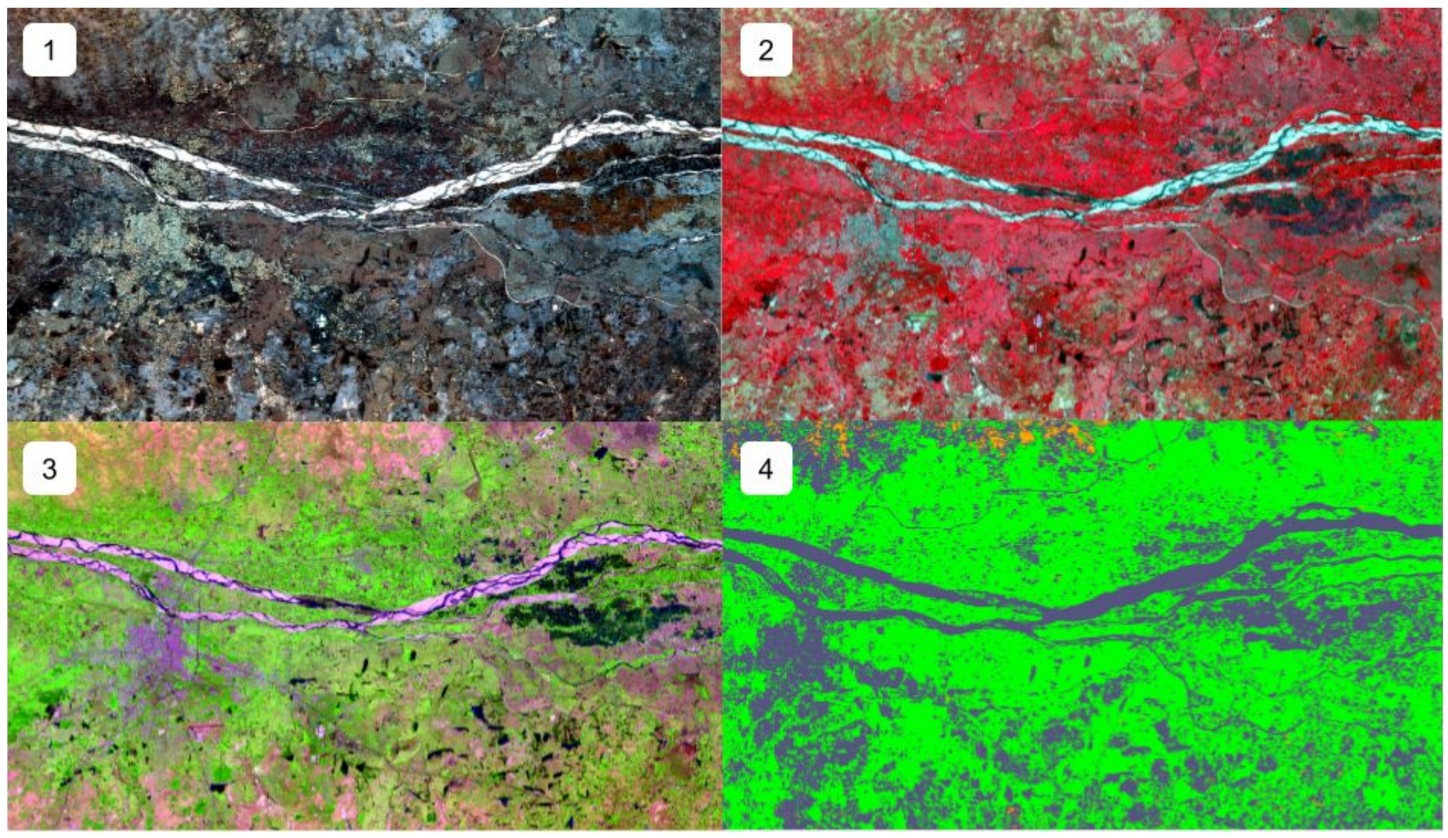
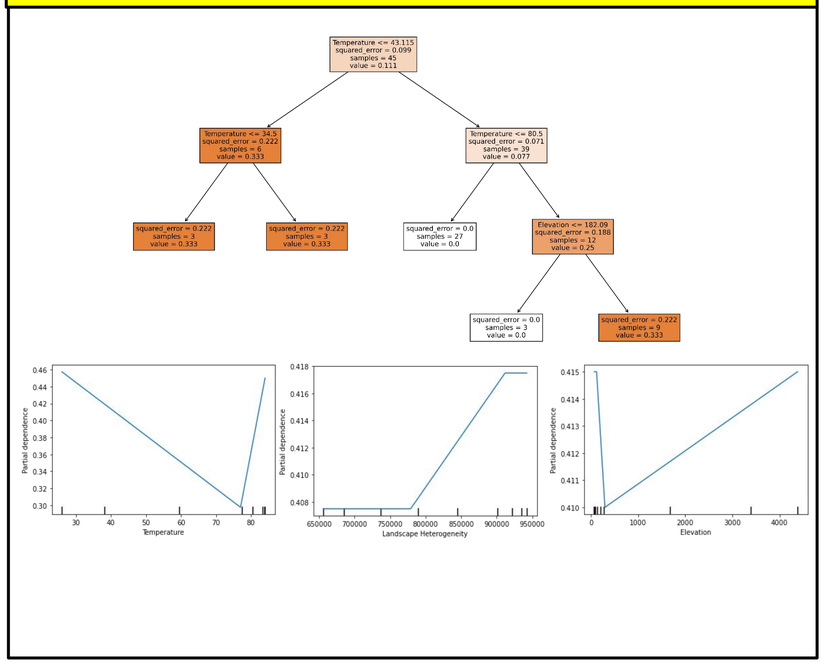


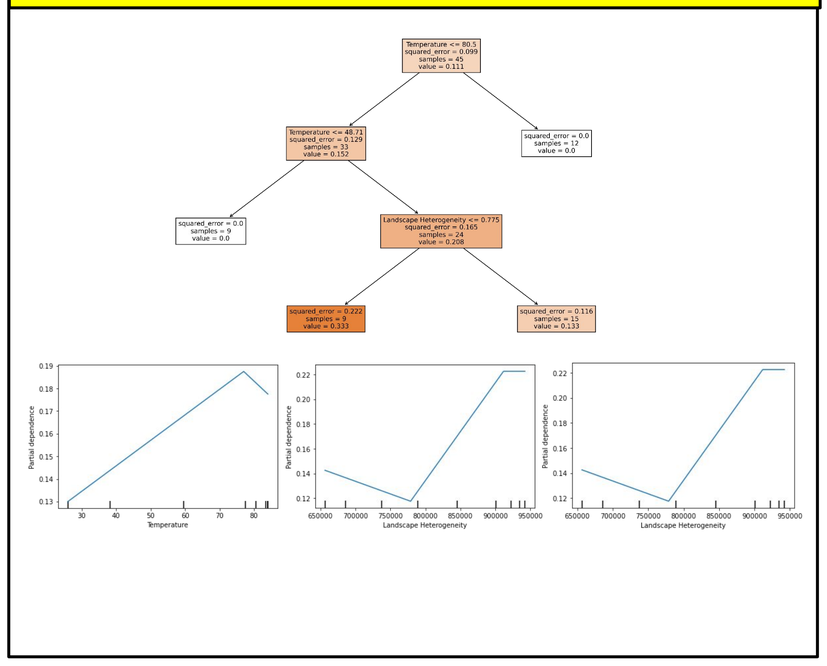
Figure 4. Composite Band Layers and the Progression to Landscape Classification. As described in Fig 2, different combinations of three satellite image bands yield different composite images of a location. Image 1 is a combination of the green, blue, and coastal wavelength bands, yielding a cold-intensity composite. Image 2 combines near-infrared, red, and green bands to create a color infrared composite that accentuates vegetation. Image 3 shows a composite of near-infrared, shortwave infrared, and green bands. This composite yielded the best contrast of urban structures, vegetation, and barren earth and was the composite selected for image classification. Using SCP, urban structures, vegetation, and barren earth were designated as the different classes and regions of interest comprising pixels of each class were selected. The resulting output is shown in Image 4, with urban structures appearing as gray, barren earth as orange, and vegetation as green. Following this terrain classification, a classification report yielded the respective percentages of total pixels for each class which populated the final column of the dataframe (Table 1).

Results – Decision Tree Classification and Partial Dependence Plots

Aedes aegypti



Culex pipiens



Aedes albopictus