IOSR Journal of Environmental Science, Toxicology and Food Technology (IOSR-JESTFT)
e-ISSN: 2319-2402,p- ISSN: 2319-2399.Volume 14, Issue 12 Ser. II (December 2020), PP 24-30
www.iosrjournals.org

Land Cover Feature Analysis for Mosquito Habitats

David W. Lin
AP Research, Seven Lakes High School, Katy, Texas, USA

Abstract:
Background*: Machine learning models were used to analyze the land cover features extracted from the Mosquito Habitat Mapper, to explore possible relationship between those features and the population distribution of mosquitos. We aim to find out what terrestrial and anthropogenic features on the land cover enhance the growth of mosquitos, while verifying some hypothesis implied by daily life experience.*
Materials and Methods*: Terrestrial and anthropogenic features were extracted from the Mosquito Habitat Mapper JSON file, which contains the data collected during the 2020 NASA SEES summer project. Python packages of text corpus processing and document-term conversion were used to process the data for machine learning. Two machine learning algorithms, Naïve Bayes and Ordinary Least Squares, were used to analyze the data.*
Results*: The Naïve Bayes model showed an accuracy that indicates the feasibility of using land cover features to predict the mosquito distribution, while the Ordinary Least Squares model identified land covers that have prominent impact to the mosquito larvae population.*
Conclusion: *Restricted by the data availability, the models' performance was limited, and this gives indicator about the research direction in the future to enhance data collection and organization.*

Key Word*: Mosquito Habitat, Land Cover Features, Machine Learning, Naïve Bayes, Ordinary Least Squares.*

Date of Submission: 14-12-2020 Date of Acceptance: 28-12-2020

I. Introduction

Scientific studies have revealed the importance of environmental modeling using land cover data. Extent work has been done in land cover map construction[1]. David Saah, et. al. presented an online tool for systemic data collection for land cover and use applications[2]. A tool for mapping land conversion was developed to use global land cover datasets to address the needs in identifying anthropogenic land conversion across a local region[3]. On the other hand, machine learning methods have been widely used in analyzing land cover data for various applications. Pardhasaradhi Teluguntla, et. al. used Random Forest method to analyze land cover image data to assess cropland products in China and Australia[4], and similar work was done in Southeast and Northeast Asia[5].

Mosquito habitat has been studied to address disease propagation issues[6]. It is important to find ecological routes in mosquito growth on anthropogenic land cover. NASA houses studies in using satellite data to track mosquito evolution[7]. In 2020 NASA SEES summer program, mosquito habitat data were collected to facilitate analytical studies. In this research, we use the GLOBE Mosquito Habitat Mapper data uploaded to the NASA SEES website. The dataset contains information about the terrestrial and anthropogenic features of mosquito habitats and observations of mosquito growth. The data were collected from various locations around the globe. Although there are studies about the ecology of mosquito habitats, it is informative to verify hypothesis about what terrestrial and anthropogenic features encourage the growth of mosquito population in different locational and geological settings. In particular, information about the influence of anthropogenic land cover features on the growth of mosquitos will be useful in policy making, administrative planning, and social and behavioral studies. The assessment of the machine learning models will also provide information about the quality of the data, and appropriateness of the research methods, and the directions for further improvement.

II. Material And Methods

The understanding of the growth of mosquito population is vital to human health, especially in disease propagation and control. Land cover features are important factors that define the environment in which the mosquitos breed and grow. These features are largely determined by natural climate changes and human activities. The study of the relationship between land cover features and mosquito growth will give insight to how human activities affect mosquito growth and what measures can be taken to assist disease control or environmental improvement.

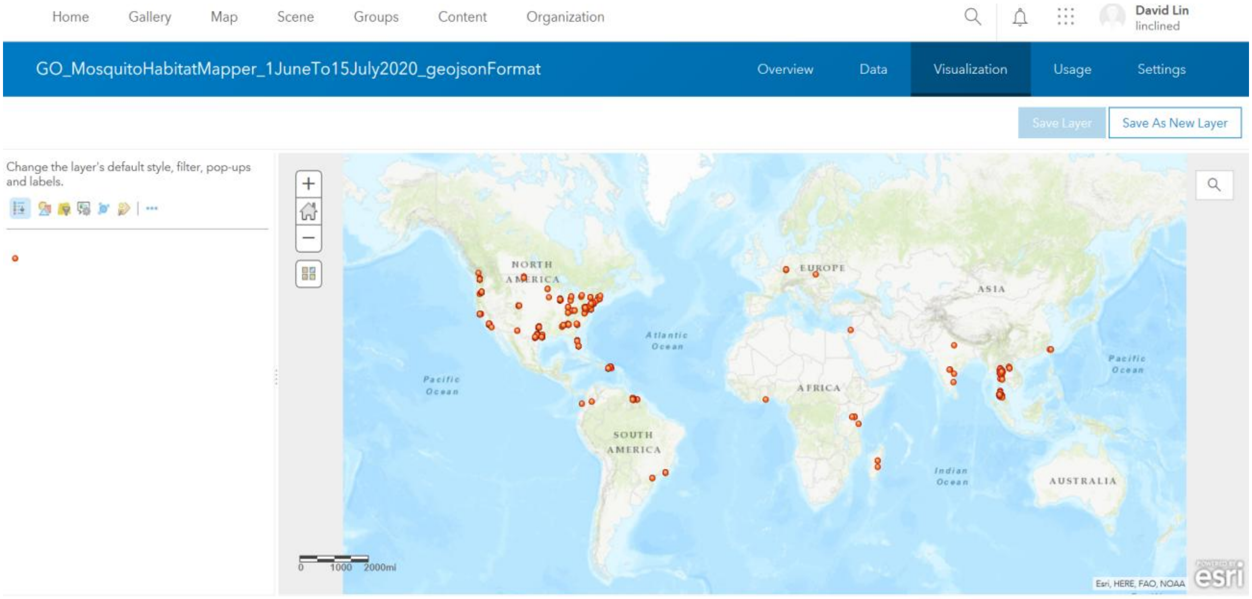
Research Questions: This research aims to find answers to the following questions:

1. Are there land cover features that affect the growth of mosquitos? If so, what are they, and how they differ from other land cover features?
2. With the given data support, is it feasible to build machine learning models to predict mosquito distribution? What data processing tools are needed in the process?

To answer the questions, proper data analysis shall be applied to verify the hypothesis. Data must support modeling. Given the uncertainty in the nature of the topic, the models may not lead to a firm conclusion to the research question, and therefore subsequent model evaluation is needed to validate the model and collect information about future improvement.

Procedure methodology

The Mosquito Habitat Mapper data used in this study are saved in file "GO_MosquitoHabitatMapper_1JuneTo15July2020_geojsonFormat" in JSON format. It collected data from sites in North and South America, Europe, Africa, Middle East, South and Southeast Asia. Figure 1 shows the sites of data collection.



Change the layer's default style, filter, pop-ups and labels.

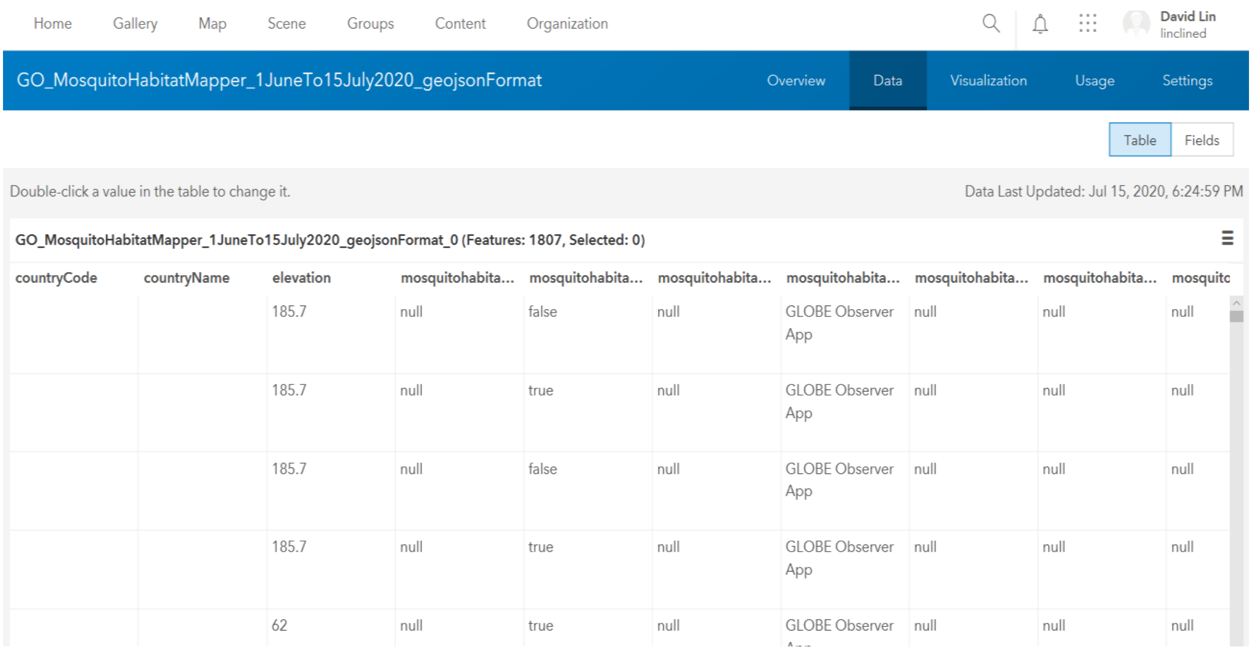
Figure 1. Mosquito Habitat Mapper data collection sites

The JSON file was downloaded and converted to a CSV file. The dataset contains 1807 rows, which reflects the number of data collection sites. It has data such as latitude, longitude, elevation, country code and name, mapper comments, larvae count, water source and type, and 3 logical variables indicating whether mosquito adult, pupae, and egg were observed. Figure 2 is a snapshot of data entries in the web page of GLOBE.

Terrestrial and anthropogenic features were extracted from the mosquito habitat mapper comments, water sources and types. Data in those 3 columns are text descriptions. Python packages "nltk" and "sklearn.feature_extraction.text" were used to tokenize strings, extract alphabetical words, convert words to lower case, removing stop words, and convert words to stem forms.

The existence and growth of mosquitos are reflected in 4 variables: the larvae count, and 3 logical variables indicating whether mosquito adult, pupae, and egg were observed. These 4 variables were combined into a single logical variable indicating whether there were mosquitos at the corresponding site. The criterium for the existence of mosquitos was either egg, larvae, pupae, or adult was observed.

The extracted words reflected terrestrial features such as "lake", "pond", "trough", and "flow"; and anthropogenic features such as "contain" (stem of "container"), "well", "cistern", "bottle", "pot", "artificial", and "ovitrap". Figure 3 shows the word clouds for the sites where mosquitos were observed (left), or not observed (right). The size of a word in word cloud reflects the frequency of its occurrence in the dataset. The bigger is the word size, the higher is the frequency of its occurrence.



GO_MosquitoHabitatMapper_1JuneTo15July2020_geojsonFormat

Double-click a value in the table to change it.

Data Last Updated: Jul 15, 2020, 6:24:59 PM

GO_MosquitoHabitatMapper_1JuneTo15July2020_geojsonFormat_0 (Features: 1807, Selected: 0)

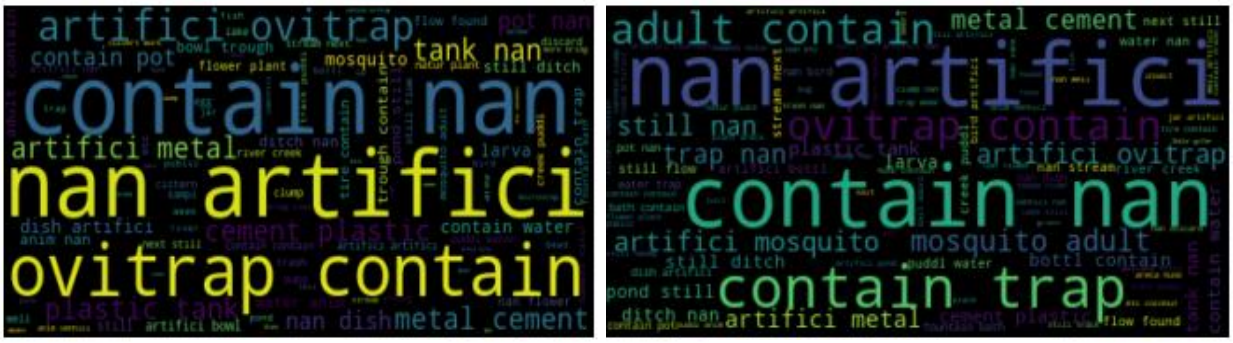
countryCode	countryName	elevation	mosquitohabita...	mosquitohabita...	mosquitohabita...	mosquitohabita...	mosquitohabita...	mosquitohabita...	mosquitc
		185.7	null	false	null	GLOBE Observer App	null	null	null
		185.7	null	true	null	GLOBE Observer App	null	null	null
		185.7	null	false	null	GLOBE Observer App	null	null	null
		185.7	null	true	null	GLOBE Observer App	null	null	null
		62	null	true	null	GLOBE Observer	null	null	null

Figure 2. Mosquito Habitat Mapper data entries

It can be seen that the differences in word frequencies in those two word clouds are not very prominent. This could imply difficulties in building a high quality model.

After feature extraction, the dataset was converted into a document-term matrix in which each entry is a 0-or-1 number indicating whether a feature (or term) appeared in a habitat site (document). After that, Naïve Bayes algorithm was used to build the classification model. I used the multinomial Naïve Bayes model in the “sklearn.naive_bayes” package. This experiment was to explore the feasibility of building a classifier that uses terrestrial and anthropogenic features to predict the existence of a mosquito habitat.

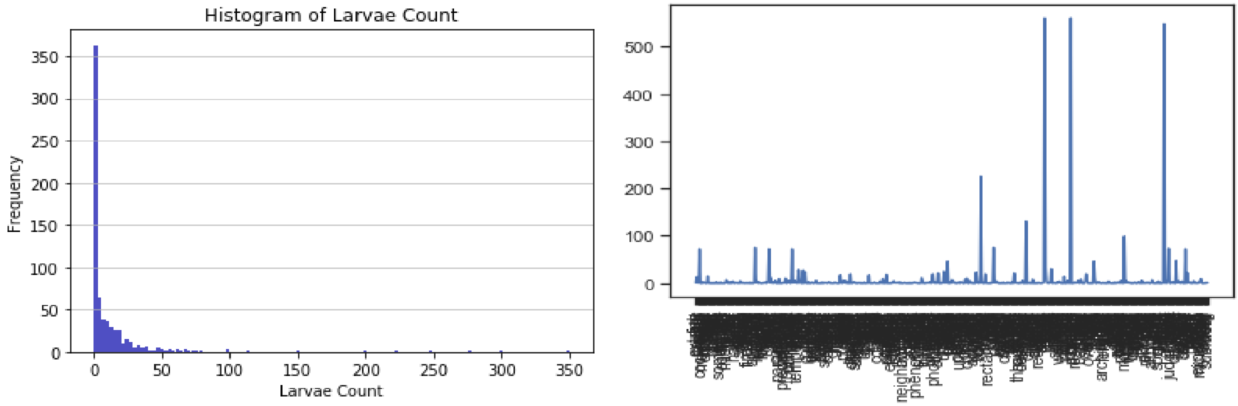
The second experiment was to identify the features that have high probably of encouraging the growth of mosquitos. To this end, the number of larvae was used as the dependent variable, and the latitude, longitude, elevation, and the binary indicators of feature existence in the document-term matrix were used as dependent variables. Since there are a lot of null entries in the number of larvae, habitats corresponding to a null value were removed from the dataset. In addition, features with less than 10 occurrences in the document-term matrix were removed. Figure 4 shows the distribution of the larvae count in the dataset (left) and the distribution of frequencies of feature occurrences (right).



Mosquitos were observed

Mosquitos were not observed

Figure 3. The word clouds for the mosquito habitat sites



Distribution of larvae count　Distribution of feature frequencies

Distributions of larvae count and feature frequencies

A pair grid was created to investigate the correlation between locational features and the larvae count. The locational features include latitude, longitude, and elevation. The pair grid is in Figure 5.

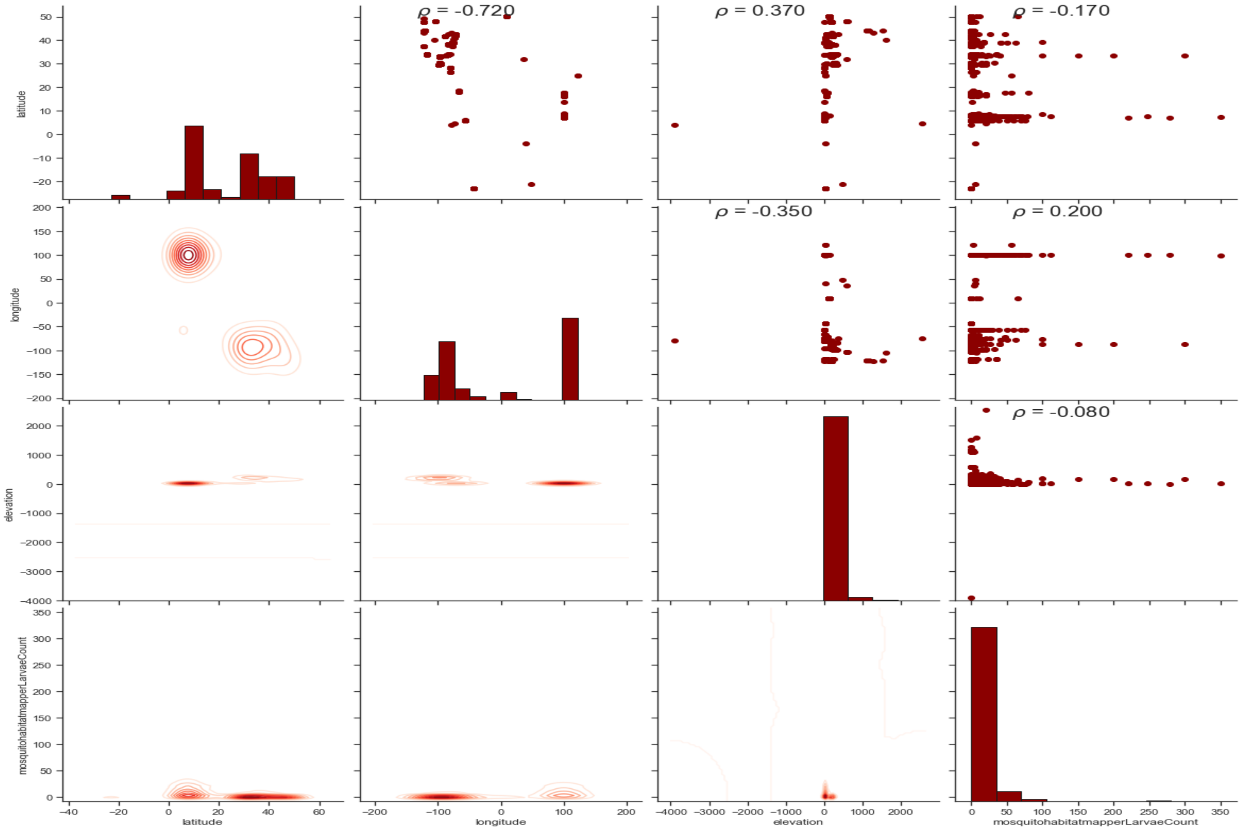
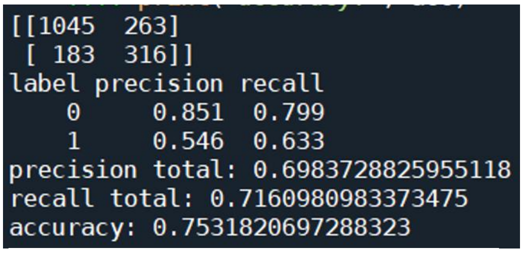


Figure 5. Pair grid

The pair grid displays the histogram of each variable on the diagonal, and the scatter plot and the Pearson correlation between each pair of variables on the upper triangle of the grid. An uneven distribution can be observed from the histogram of each variable, especially the elevation. Moreover, the correlations between each of the locational variables (viz., latitude, longitude, and elevation), and the larvae count are -0.17, 0.2, and -0.08, respectively, indicating that there is no strong correlation between each of the locational variables and the larvae count. These factors do not favor a high quality linear regression model based on those three locational variables, with longitude having the highest probability of enhancing the prediction power of the model.

III. Result

The confusion matrix that shows the classification result of the Multinomial Naïve Bayes model is in Figure 6. In Figure 6 are also the precision and recall for each class, where label 0 and 1 represent class "mosquito not observed" and "mosquito observed", respectively. The averaged precision, recall, and overall accuracy are shown thereafter. Each precision, recall, and accuracy is higher than the chance level (0.5). The performance of classifying class "mosquito not observed" (precision: 0.851, recall: 0.799) is higher than classifying class "mosquito observed" (precision: 0.546, recall: 0.633).



```
[[1045  263]
 [ 183  316]]
label precision recall
    0     0.851  0.799
    1     0.546  0.633
precision total: 0.6983728825955118
recall total: 0.7160980983373475
accuracy: 0.7531820697288323
```

Figure 6. Naïve Bayes model performance

The following table shows the Ordinary Least Squares model. The R-squared value is 0.115, indicating low performance of the model. By examining the p values (in column "P>|t|"), we can identify variables that have high prediction power. A p value indicates the probability for the prediction made by the variable to be equal to the chance level. When the p value is close to 0, such probability is low. We can see that "longitude", "pot", "well", "bottle", "cistern", "ovitrap", and "dish" are among such variables. As noted, longitude has the highest correlation with the larvae count. The other high performing features are all anthropogenic land cover features, implying the influence of human activities in the growth of mosquito habitats.

```
                                OLS Regression Results
=============================================================================================
Dep. Variable:     mosquitohabitatmapperLarvaeCount   R-squared:                       0.115
Model:                                          OLS   Adj. R-squared:                  0.069
Method:                               Least Squares   F-statistic:                     2.505
Date:                              Fri, 24 Jul 2020   Prob (F-statistic):           7.64e-06
Time:                                      22:51:08   Log-Likelihood:                -3273.8
No. Observations:                               690   AIC:                             6618.
Df Residuals:                                   655   BIC:                             6776.
Df Model:                                        34
Covariance Type:                          nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.7604     14.345      0.541      0.589     -20.408      35.929
fish           2.6723     11.042      0.242      0.809     -19.010      24.354
water         -0.0879      6.607     -0.013      0.989     -13.062      12.886
trap         -11.8607     13.740     -0.863      0.388     -38.841      15.119
stream        -1.7525      6.122     -0.286      0.775     -13.774      10.269
bug           -2.7114      8.048     -0.337      0.736     -18.514      13.091
elevation      0.0006      0.005      0.119      0.905      -0.009       0.010
still         -1.7976     15.315     -0.117      0.907     -31.871      28.276
cement        -4.0081      9.800     -0.409      0.683     -23.252      15.236
ditch         10.0921     11.523      0.876      0.381     -12.534      32.718
tank          -4.0081      9.800     -0.409      0.683     -23.252      15.236
egg            0.7298      9.686      0.075      0.940     -18.291      19.750
natur         -3.9639     15.725     -0.252      0.801     -34.841      26.913
mosquito      -7.0848      6.705     -1.057      0.291     -20.251       6.081
grass          1.2880     11.081      0.116      0.908     -20.471      23.047
lake           1.3737     11.661      0.118      0.906     -21.524      24.271
pot           -5.9645      2.853     -2.090      0.037     -11.567      -0.362
etc           -3.9639     15.725     -0.252      0.801     -34.841      26.913
tire          -2.7275      8.531     -0.320      0.749     -19.480      14.025
flower        30.9182     30.540      1.012      0.312     -29.049      90.886
plant         -0.4195     29.319     -0.014      0.989     -57.989      57.150
metal         -4.0081      9.800     -0.409      0.683     -23.252      15.236
found          3.3100     14.003      0.236      0.813     -24.186      30.806
artifici      13.5219     18.360      0.736      0.462     -22.529      49.573
river         -1.7525      6.122     -0.286      0.775     -13.774      10.269
well          -8.2902      4.968     -1.669      0.096     -18.045       1.464
creek          5.7032     35.224      0.162      0.871     -63.463      74.869
latitude      -0.0435      0.106     -0.409      0.683      -0.252       0.165
bottl        -16.6685      8.923     -1.868      0.062     -34.189       0.852
```

```
contain          2.6746      30.640       0.087       0.930     -57.490      62.840
plastic         -3.8624      28.941      -0.133       0.894     -60.691      52.966
bait            -1.2934      10.238      -0.126       0.900     -21.396      18.809
longitude        0.0397       0.021       1.906       0.057      -0.001       0.081
bowl            -7.2934      21.212      -0.344       0.731     -48.945      34.358
trough          -7.2934      21.212      -0.344       0.731     -48.945      34.358
cistern         -8.2902       4.968      -1.669       0.096     -18.045       1.464
ovitrap         -9.2753       4.283      -2.166       0.031     -17.685      -0.866
adult            0.7932      14.251       0.056       0.956     -27.189      28.775
puddl           -1.9701      39.735      -0.050       0.960     -79.993      76.053
dish            -5.9645       2.853      -2.090       0.037     -11.567      -0.362
larva            2.3840       6.629       0.360       0.719     -10.632      15.400
next            -1.7525       6.122      -0.286       0.775     -13.774      10.269
pond             4.0635      10.646       0.382       0.703     -16.841      24.968
flow            -1.7525       6.122      -0.286       0.775     -13.774      10.269
anim             0.8627      41.985       0.021       0.984     -81.579      83.305
==============================================================================
Omnibus:                      899.657   Durbin-Watson:                   1.701
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           112942.701
Skew:                           6.776   Prob(JB):                         0.00
Kurtosis:                      64.195   Cond. No.                     1.15e+16
==============================================================================
```

IV. Discussion

The model performance is largely constrained by the data quality. In this study, terrestrial and anthropogenic features are extracted from texts in the mapper comments and the water sources and types. Given that data are collected from sites all over the sphere, consistency in writing those texts is not guaranteed. In addition, those texts may not give complete information, either. A large volume of null values in the dataset further degrades the completeness of the information. It is suggested that standard tools be used in data collection and a central platform be used for data reporting in the future. Moreover, the distribution of data collection sites is not even geographically. It is desirable to collect data in those regions with scarce density of data collection sites.

The fact that almost 2/3 of larvae counts are null largely chopped down the useful dataset and created mal-balanced dataset. The extremely clustered distribution of elevation data basically crippled the usefulness of the elevation feature. This results in that the conclusion made about elevation may not be valid.

Given these limitations, nevertheless, our study supports both hypothesis we intended to verify. We are able to build models to predict the existence of a mosquito habitat based on land cover features, and we are able to identify the features that have the highest influence to the population of mosquitos. Finer and more meaningful study can be done with support in the form of more complete data collection in the future.

V. Conclusion

Systematic mosquito habitat study has been an important research area, given its impact to healthcare and disease control. NASA's Mosquito Habitat Mapper program provides a platform to collect land cover data for mosquito habitat study. Thanks to research mentor Peter Nelson, a mosquito habitat mapper dataset was made available for this study. The dataset contains land cover data collected from mosquito habitats around the globe. However, due to the inconsistency in data collection protocol, the dataset suffers a large amount of missing information. Despite of this limitation, machine learning methods proved their usefulness in verifying some hypothesis. In my experiment, Naïve Bayes algorithm successfully built a model that shows positive capability in predicting the existence of mosquito habitat based on extracted terrestrial and anthropogenic features; and the Ordinary Least Squares regression algorithm identified features that have highest impact to the mosquito larvae population.

While this preliminary research produced promising results, its limitation due to the incompleteness of the dataset is also noticeable. Further studies are needed in more systematic data collection and finer data analysis using multiple models. One point of improvement is fine tuning the feature extraction method such that features are considered in environmental context. This will enable the model to reveal more useful information about how to manipulate anthropogenic land covers to influence the growth of mosquitos.

Acknowledgement

The NASA 2020 SEES Virtual High School Internship program, Mosquito Mappers, was sponsored by the NASA Earth Science Education Collaborative through an award to the Institute for Global Environmental Strategies, Arlington VA. by NASA under Award NNX6AE28A. I would like to thank the NASA STEM Enhancement in Earth Science Summer (SEES) Internship program coordinated by Margaret Baguio and Celena Miller, Texas Space Grant Consortium, as well as my project mentors, Peder Nelson, Dr. Russanne Low and

Cassie Soeffing. The author would also express gratitude to all the peer interns of the Mosquito Habitat Mapper program.

References

[1]. Saah D, et. al., Primitives as building blocks for constructing land cover maps, Int J Appl Earth Obs Geoinformation, Elsevier, 85 (2020), 101979.

[2]. Saah D, et. al., Collect Earth: An online tool for systematic reference data collection in land cover and use applications, Environmental Modelling & Software, Elsevier, 118 (2019), 166-171.

[3]. Jacobson A, et. al., A novel approach to mapping land conversion using Google Earth with an application to East Africa, Environmental Modelling & Software, Elsevier, 72 (2015), 1-9.

[4]. Teluguntla P, et. al., A 30-m landsat-derived cropland extent product of Australia and China using random forest machine learning algorithm on Google Earth Engine cloud computing platform, ISPRS Journal of Photogrammetry and Remote Sensing, Elsevier, 144 (2018), 325-340.

[5]. Oliphant AJ, et. al., Mapping cropland extent of Southeast and Northeast Asia using multi-year time-series Landsat 30-m data using a random forest classifier on the Google Earth Engine Cloud, Int J Appl Earth Obs Geoinformation, Elsevier, 81 (2019), 110-124.

[6]. Rejmánková E, et. al., Chapter 13. Ecology of Larval Habitats, in: Anopheles mosquitoes - New insights into malaria vectors, InTech, (2013), 397-446.

[7]. Patel K, Of Mosquitoes and Models: Tracking Disease by Satellite, NASA Earth Observatory, online story at: https://earthobservatory.nasa.gov/features/disease-vector?src=eoa-features, July 9, 2020.

David W. Lin. “Land Cover Feature Analysis for Mosquito Habitats.” *IOSR Journal of Environmental Science, Toxicology and Food Technology* (IOSR-JESTFT), 14(12), (2020): pp 24-30.