Predicting *Culex* Mosquito Habitat and Breeding Patterns in Washington D.C. Using Machine Learning Models

Iona Xia, Neha Singirikonda, Landon Hellman, Jasmine Watson, Marvel Hanna

NASA STEM Enhancement in the Earth Sciences 2022

Abstract

Culex mosquitoes pose a large threat to humans and other species due to their ability to carry deadly viruses such as the West Nile and Zika Viruses. Washington D.C. in particular has a humid subtropical climate that is ideal as habitats for mosquito breeding. Thus, tracking the habitats and breeding patterns of mosquitos in Washington D.C. is crucial towards addressing local public health concerns. Although fieldwork techniques have improved over the years, tracking and analyzing mosquitos is difficult, dangerous, and time-consuming. In this work, we propose a solution to this issue by creating a *Culex* mosquito abundance predictor using machine learning techniques to determine under which conditions *Culex* mosquitoes thrive and reproduce. We used four environmental variables to conduct this experiment: precipitation, specific humidity, enhanced vegetation index (EVI), and surface skin temperature. We obtained sample data of these variables in the Washington D.C. areas from the NASA Giovanni Earth Science Data system, as well as mosquito abundance data collected by the D.C. government. Using these data, we created and compared four different machine learning regression models: Random Forest, Decision Tree, Support Vector Machine, and Multi-Layer Perceptron. For each model, we searched for the optimal configurations to get the best fitting possible. It was discovered that the Random Forest Regressor produced the most accurate prediction of mosquito abundance in an area with the four environment variables, with a mean average error of 3.3. It was also found that EVI was the most significant factor in determining the mosquito abundance. Models and findings from this research are going to be utilized by public health programs for mosquito related disease observations and predictions.

Keywords: mosquito breeding patterns, machine learning techniques, *Culex* mosquitoes, ecological variables

Research Question

How can we predict *Culex* mosquito breeding patterns in Washington D.C. with GLOBE and open-sourced data utilizing machine learning techniques?

Introduction and Review of Literature

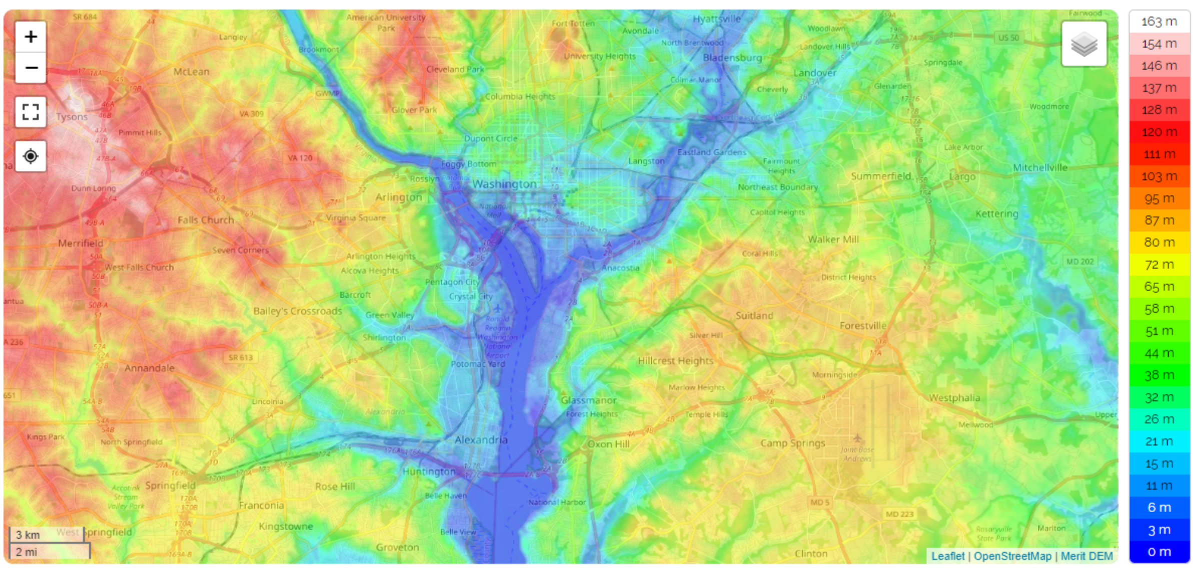
Culex mosquitoes are some of the most common species of mosquitoes in the world and can carry many diseases including the West Nile and Zika Viruses (Omodior et al. 2018). Currently, there are no licensed vaccines or medicines for such diseases. *Culex* larvae are native to Africa, Europe, and Asia, however they are currently populated virtually everywhere around the globe (Soh and Aik 2021). Thus, this makes several societies prone to these deadly diseases. Studies and projects have been done to limit their noticeable effects on human populations; however, very little has actually been able to be done due to the lack of understanding on how these viruses affect the body (Center of Disease Control 2020). Furthermore, there are currently no established systems that predict mosquito-borne disease outbreaks.

Washington D.C. is known to have a humid subtropical climate which tends to be ideal for mosquito breeding habitats. Its annual rainfall average is 42 inches per year, which is 12 inches higher than the annual average precipitation nationwide which is 30.2 inches, thus making Washington D.C. highly prone to an abundance of mosquitoes because they thrive in relatively wet places (National Centers for Environmental Information 2022). Washington D.C.'s has an unique environment, with the Potomac River, a freshwater stream running through Washington D.C., serving this area and the areas around it as habitable mosquito microenvironments. Washington D.C's natural topography varies, as the area to the left of the Potomac River ranges from 100-140m above sea level while the area to the right of the Potomac River ranges from

40-80m (Topographic-Maps 2022, Figure 1). The first human case of West Nile Virus in the United States was reported in the District of Columbia (DC Health 2022). With 13 reported human cases in 2018 and 11 reported human cases in 2019 in a mere 68.34 mi^2 (Center of Disease Control 2020), it is evident that not only does Washington D.C. have a major public health mosquito threat, but that the West Nile Virus, and other mosquito-transmitted diseases, can become a serious problem in such a small area.

Figure 1: A representation of Washington D.C.'s natural topographyThe geographic center of Washington is near the intersection of 4th and L Streets NW. The highest natural elevation in the District is 409 feet (125 m) above sea level at Fort Reno Park in upper northwest Washington. The lowest point is sea level at the Potomac River.

Scale Unknown, (CC-BY-SA 3.0), Washington D.C., 2022.



Historically, machine learning models have proven to be useful tools in predicting trends and operational patterns. Machine learning is significantly useful in predicting breeding patterns of various animals, including mosquitoes. We chose four primary machine learning models: Random Forest, Decision Tree, Support Vector, and Multilayer Perceptron. Given its several capabilities, random forest regression models were recently used in predicting West Nile Virus positivity rates and abundance in the city of Chicago (Schneider et al. 2021). Random Forest algorithms can construct several decision trees to perform classification assignments that allow users to make highly accurate discrete predictions and solutions for their data sets. Moreover, the decision trees constructed by the Random Forest models can be used to create regression tasks that further help users predict continuous outputs for nonlinear inputs (Schonlau and Zou 2020). Because random forest is highly used specifically in geology and earth science, we predict that it would also do the best in our tests. Decision Tree regression, on the other hand, features only one decision tree analysis feature by partitioning data and fitting a simple model for each partition, and exists as a simpler method of Random Forest (Lou 2011). The Support Vector regression model is based on a linear regression that fits its data by a hyperplane in a higher dimension, allowing for it to recognize subtle patterns (Basak and Pal 2007). In addition, Multilayer Perceptron regression is a neural network algorithm that consists of various nodes, and has the capability to learn non-linear models, which could prove to be helpful (Murtagh 1991).

A previous study by Drakou et al. (2020) displayed that Cx. pipiens, Ae. detritus and Ae. Caspius mosquitoes increased as a result of precipitation. Francisco et al. (2021) found connections between the Dengue disease, another disease transmitted by mosquitoes, and environmental and landscape factors, including precipitation, land surface temperature, normalized difference of vegetation index, and road network. Other studies such as Madewell et al. (2019) have also found connections with urbanization. Thus, we found it useful to study various environmental factors with NASA Giovanni data and landcover with GLOBE.

Research Methods

Data Collection

We chose Washington D.C. as our primary area of interest (AOI) due to its high accessibility of open-sourced mosquito and environmental data, as well as its historical abundance of mosquitoes. We obtained data from four main sources: Giovanni, GLOBE, and Washington D.C. government data. The data we collected spans from April of 2016 to October of 2018. We began by looking into the NASA Giovanni Earth data collection website. This was where we gathered the daily data regarding our environmental variables: Average Surface Skin Temperature, Specific Humidity, Precipitation and EVI (Giovanni 2022). The Washington D.C. government provided open-sourced quantitative mosquito data in our respective AOI (Open Data DC 2021). From there, we found correlations between that data and the daily Giovanni variable data, and were able to create charts that compared the number of mosquitoes in Washington D.C. to our environmental factors.

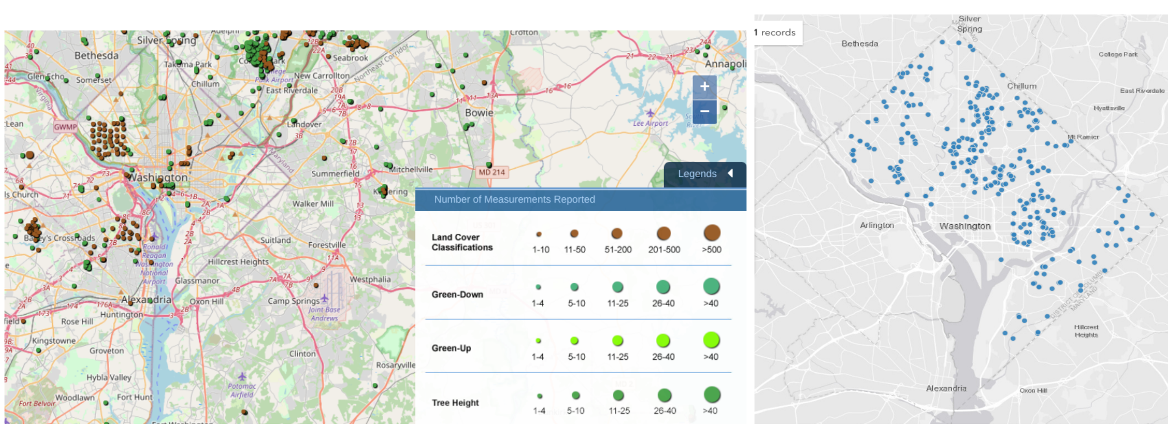
Initially, when trying to extract the data from GLOBE, we ran into the issue of inconsistent data, specifically because of its opportunistic nature. Although we had a hard time finding several data points in one area at first, we eventually found a way around this problem by looking at the data from a different perspective. Instead of looking at the mosquito mappers data, we began to look at the land cover data. This helped us immensely in our quest to answer our initial research question by giving us perspective on how the land cover pictures around Washington D.C. correlate to the environmental factors we used in our experiment. The availability of GLOBE citizen science data in Washington D.C. indicates the location as an area of interest for health programs and government efforts. We used the GLOBE data to analyze certain land cover observations in Washington D.C., which allowed us to determine which habitats correlate with specific environmental factors and therefore, the mosquito breeding patterns. GLOBE data was collected using land cover measurements updated daily, including green-down, green-up, tree height, and land cover classifications. Using the qualitative data of GLOBE and open source data, it was determined that the land cover points were generally concentrated in ward 3 of Washington D.C. with greater GLOBE citizen science activity. Using GLOBE and Washington D.C. open data, ward 3 displayed relatively low container removals (Figure 2). Therefore, we speculate that ward 3 will result in relatively greater mosquito breeding as there are lower container removals which are an ideal breeding habitat for *Culex* mosquitoes.

Figure 2: A representation of our data collected from both the government and citizen science.

Left: Washington D.C. land cover data updated daily made with GLOBE Observer. Legend includes land cover classifications, green-down, green-up, and tree height.

Global Learning and Observations to Benefit the Environment (GLOBE) Program, July 21, globe.gov

Right: Washington D.C. open access data from arcGIS of container removal updated monthly. Scale Unknown, Open Data DC in the Office of the Chief Technology Officer, Washington D.C., 2022



Data Analysis

Data Features

We were able to calculate the relationship between mosquito breeding patterns and our environmental variables that have been shown to influence mosquito breeding patterns in the past (precipitation, EVI, specific humidity, surface skin temperature). We chose these four environmental factors because we hypothesized that they would have the biggest effect on mosquito breeding patterns. We chose to calculate this relationship because it allowed for our machine learning models to predict Culex mosquito breeding patterns given certain environmental conditions in our AOI location. The visual derivatives of the mosquito abundance in our AOI location showed high variability over our selected time-span, and were utilized in our models to further identify the specific relationships. The means and ranges of all of the data used is shown in Table 1.

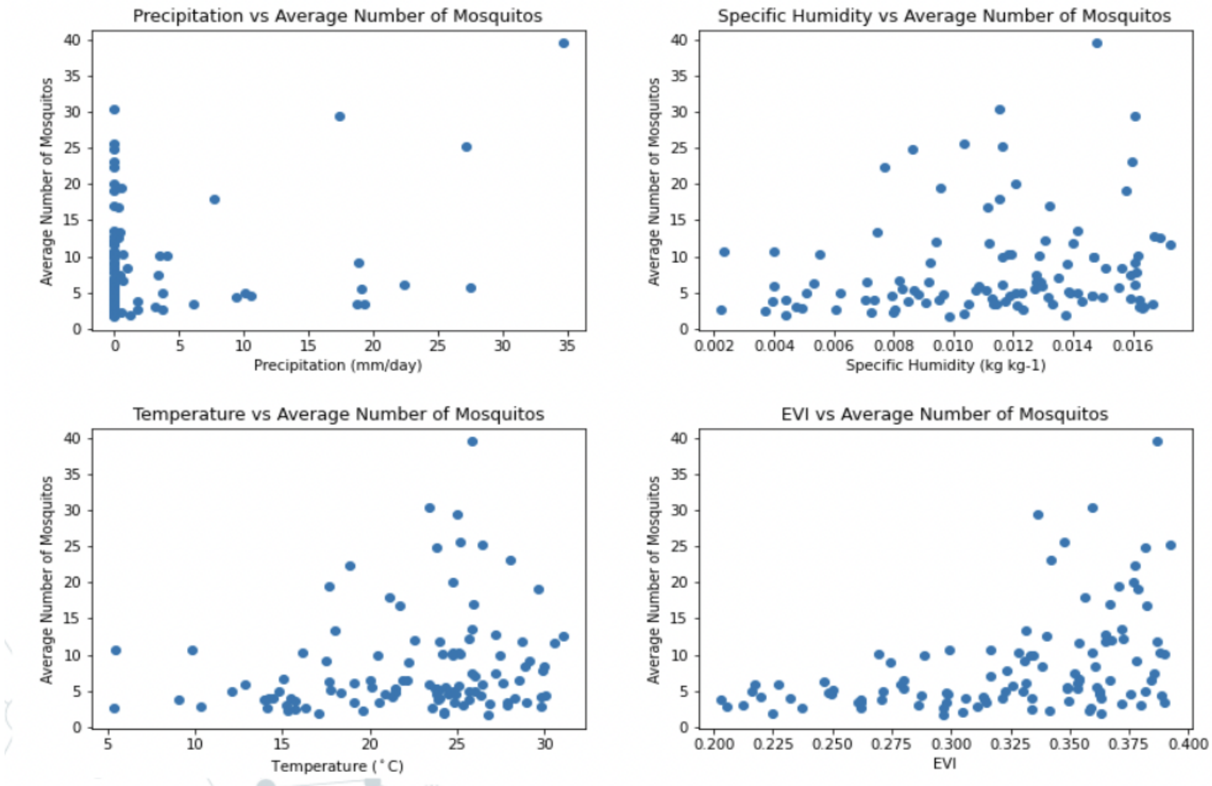
Table 1: The mean and ranges of inputs.

Factor	Mean	Range
Average Mosquitoes	8.3600	37.720
Precipitation (mm/day)	2.5939	34.623
Specific Humidity (kg/kg)	0.0113	0.0150
Average Skin Surface Temperature (℃)	22.352	25.639
EVI (Spectral Index)	0.3279	0.1882

Analyzing Trends and Lag

We compared each of our ecological variables with the daily mosquito data taken from the government of Washington D.C. and displayed the graphical relationships between each ecological variable and mosquito abundance in Figure 3. We found its p-value through the Ordinary Least Square Regression model. In order for us to consider a factor as being statistically significant, it needed to have a p-value < 0.05. We had previously hypothesized that there would be a lag in the data due to the lengthy incubation period of mosquito eggs, which would lead ecological variables to only show an effect in the population several days after (Environmental Protection Agency 2022).

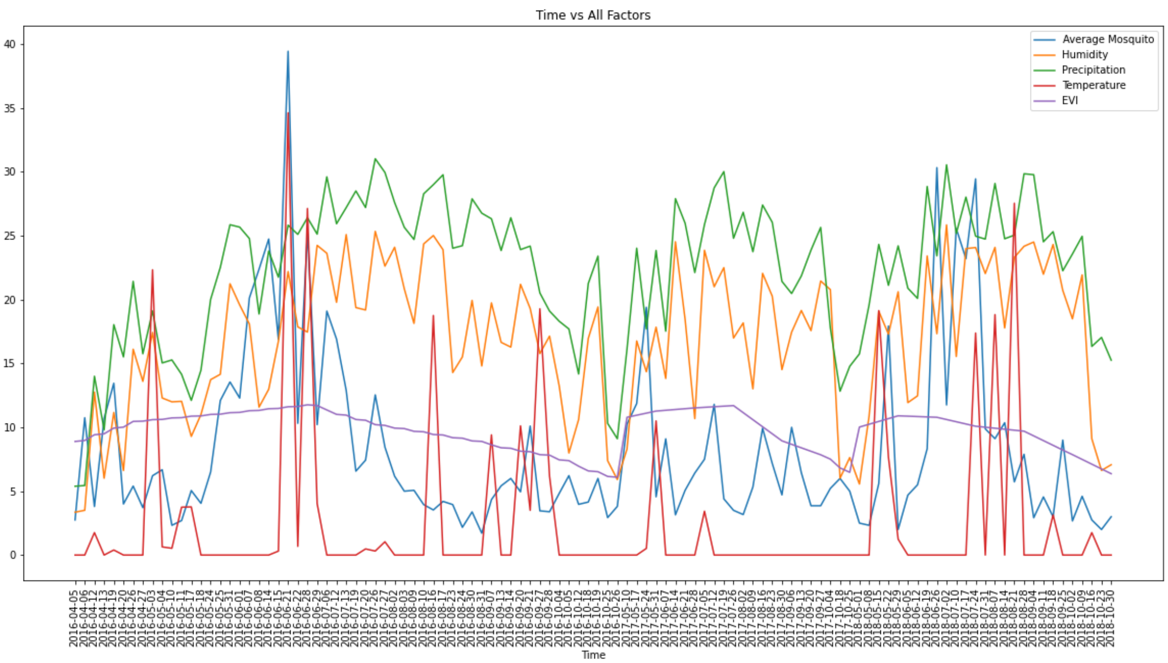
Figure 3: *Top Left:* The scatter plot shows the relationship between the precipitation and average mosquito populations. Precipitation has a p-value of 0.0023. *Top Right:* The scatter plot shows the relationship between the percentage of specific humidity and average mosquito populations. Specific humidity has a *p*-value of 0.0487. *Bottom Left:* The scatter plot shows the relationship between average surface skin temperature measured in ℃ and average mosquito populations. Average Surface Skin Temperature has a p-value of 0.0393. *Bottom Right:*The scatter plot shows the relationship between enhanced vegetation index (EVI) measured in Spectral Index (Band Ratio) and average mosquito populations. EVI has a p-value of 5e-6.



However, when we compared the difference between a one-week lag to no lag, we found that most of the data was more statistically significant (had smaller p-values) with no lag. Temperature was the only one that was more statistically significant with a one-week lag. Peaks in the data also matched better with no lag (Figure 4). In fact, certain factors such as humidity and precipitation would not have been statistically significant with the one-week lag. We compared the differences between the p-values of each ecological variable with and without the data lag in Table 2.

Figure 4: The line graphs show the relationship between time and precipitation, temperature, humidity, and enhanced vegetation index (EVI) with no lag time. EVI values have been

multiplied by 30 and humidity values have been divided by 5 for scaling purposes.



Furthermore, we received a p-value < 0.05 for every ecological variable, proving that they all have significant effects on mosquito populations. This rejects the null hypothesis stating that none of the ecological variables have a statistically significant relationship with the mosquito populations.

Table 2: The differences between the *p*-values of precipitation, temperature, humidity, and enhanced vegetation index (EVI) with and without the data lag. A statistical significance threshold of 0.05.

Variables	1 Week Lag	No Lag
Precipitation	0.349566	0.002316
Temperature	0.021255	0.039277
Humidity	0.581798	0.048652
EVI	0.000020	0.000005

Data Preprocessing

Because all of the data came from various satellites or sources, it was necessary for there to be a lot of data cleaning. The Washington D.C. government-collected mosquito data was measured twice a week for the majority of 2016, and then once a week for the rest of the time. The data contained both females and males of various types of mosquitoes. Because we were focused on predicting mosquito growth, we only kept data for female mosquitos, as well as focusing on *Culex* mosquitoes. Due to the varying amounts of mosquito traps per day, we took the average number of mosquitoes per trap per day for the dates at which mosquitoes traps were set. In total, we had 108 days from the months of May to October of 2016-2018. For each of these days, we collected data for EVI, average surface skin temperature, specific humidity, and precipitation. We found that there were a couple of holes for some of the environmental factor data. For these, we used a SciPy's interpolation method interp1d to fill in the holes, which we found to be the most accurate.

Training the Model

We trained a variety of models in order to find the best possible model. All models were from the SciKit-Learn python package, and we tested their hyperparameters using its Grid Search Cross Validation tool. We tested four models in total: the Random Forest Regressor model, the Decision Tree model, the Multilayer Perceptron model, and the Support Vector Regression model. The hyperparameters tested and chosen for Random Forest Regressor and Decision Tree specifically are listed in Tables 3 and 4. We decided to use a 70/30 training-testing split, using SciKit-Learn's function "train_test_split" to split them up randomly.

Table 3: The hyperparameters for the Random Forest regressor including the values tested and chosen by GridSearchCV.

Hyperparameter	Values Tested	Chosen Value
‘bootstrap’	True, False	True
‘max_depth’	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None	60
‘max_features’	'auto', 'sqrt'	‘sqrt’
‘n_estimators’	100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, 1900	300
‘min_samples_leaf’	1, 2, 4	1
‘min_samples_split’	2, 5, 10	2

Table 4: The hyperparameters for the Decision Tree regressor including the values tested and chosen by GridSearchCV.

Hyperparameters	Tested Values	Chosen Value
‘splitter’	‘best’, ‘random’	‘random’
‘max_depth’	1, 3, 5, 7, 9	9
‘min_samples_leaf’	1, 2, 3, 4, 5, 6, 7	4
‘min_weight_fraction_leaf’	0.1, 0.2, 0.3, 0.4, 0.5	0.1
‘max_features’	‘auto’, ‘log2’, ‘sqrt’, None	‘auto’
‘max_leaf_nodes’	10, 20, 30, 40, 50, 60, None	60

Results

When testing our four models, we decided to measure them in two different metrics: the mean absolute error (MAE) and the root mean square error (RMSE). These two measurements are both important in different ways. MAE measures the average magnitude of the difference of the error, without caring about the direction. RMSE, on the other hand, measures the square root of the average of the squared differences. While MAE is more often used, especially when used for comparing model statistics, it has been found that RMSE is better used to represent the model performance when the error is in Gaussian distribution. This is why we found it necessary to measure both in order to understand which model performance performed the best (Chai and Draxler 2014). When comparing all four, we found that Random Forest performed better in both MAE and RMSE. Support Vector Machine performed the worst in RMSE, and Multi-Layer Perceptron performed the worst in MAE. Table 5 details the scores for each of the models in MAE and RMSE. However, all models did perform similarly.

Table 5: Regression Data: Shows random forest, decision tree, support vector, and Multi-Layer Perceptron Regressor's data.

Model	Mean Absolute Error	Root Mean Square Error
Random Forest Regressor	3.27046	5.14630
Decision Tree Regressor	3.40613	5.30988
Support Vector Regressor	3.51837	6.08155
Multi-Layer Perceptron Regressor	3.92544	5.40554

Figure 5: Random Forest Regressor: The relationship between predicted number of mosquitoes by model and actual number of mosquitoes for the Random Forest Regressor. *Left:* A plot of prediction by the model vs actual number of mosquitoes. *Right:* Another representation with the prediction by model being plotted in blue and the actual value plotted in orange.

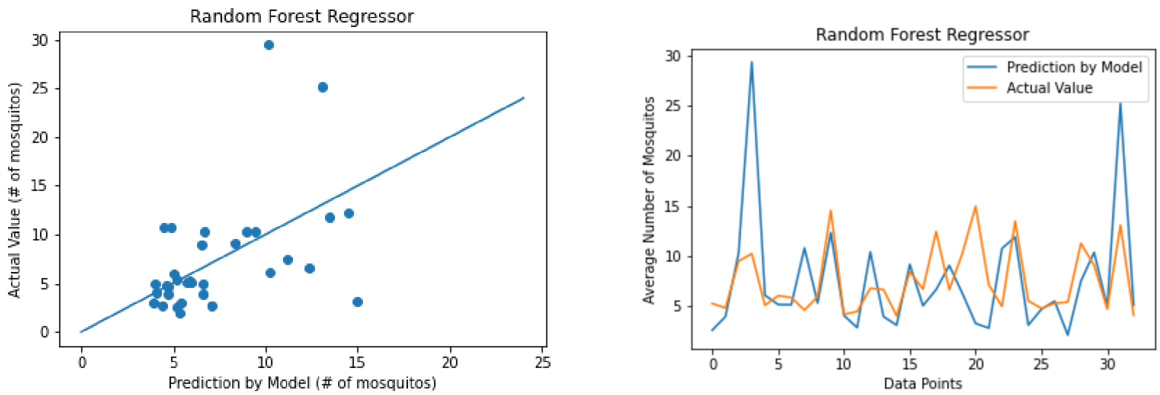


Figure 6: Decision Tree Regressor: The relationship between predicted number of mosquitoes by model and actual number of mosquitoes for the Decision Tree Regressor. *Left:* A plot of prediction by the model vs actual number of mosquitoes. *Right:* Another representation with the prediction by model being plotted in blue and the actual value plotted in orange.

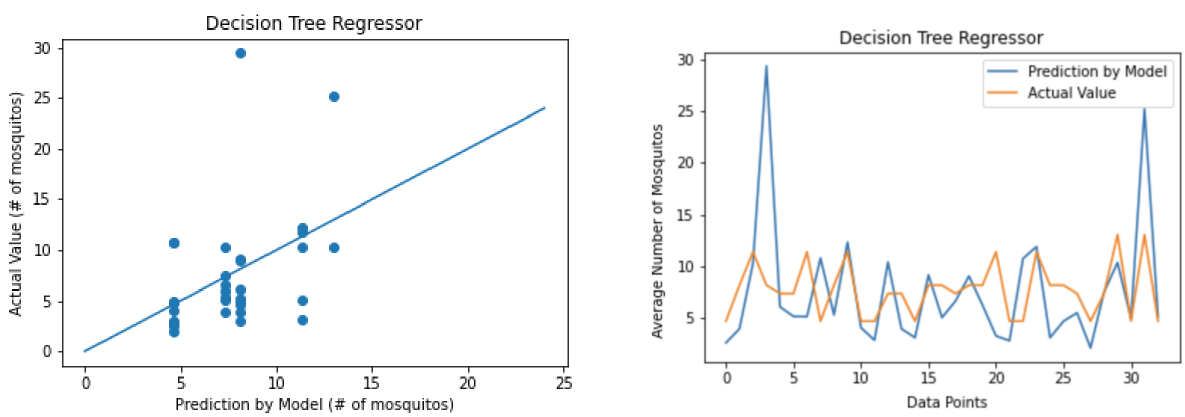


Figure 7: The relationship between predicted number of mosquitoes by model and actual number of mosquitoes for the Support Vector Regressor. *Left:* A plot of prediction by the model vs actual number of mosquitoes. *Right:* Another representation with the prediction by model being plotted in blue and the actual value plotted in orange.

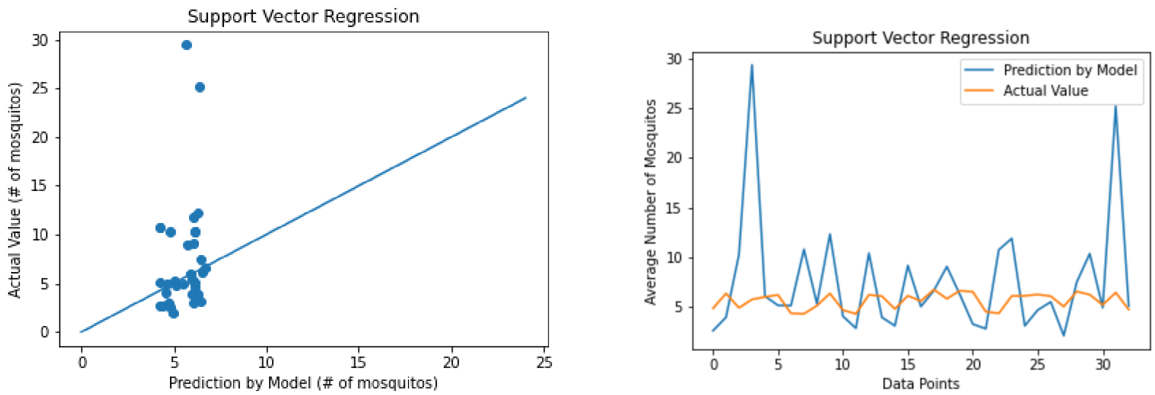
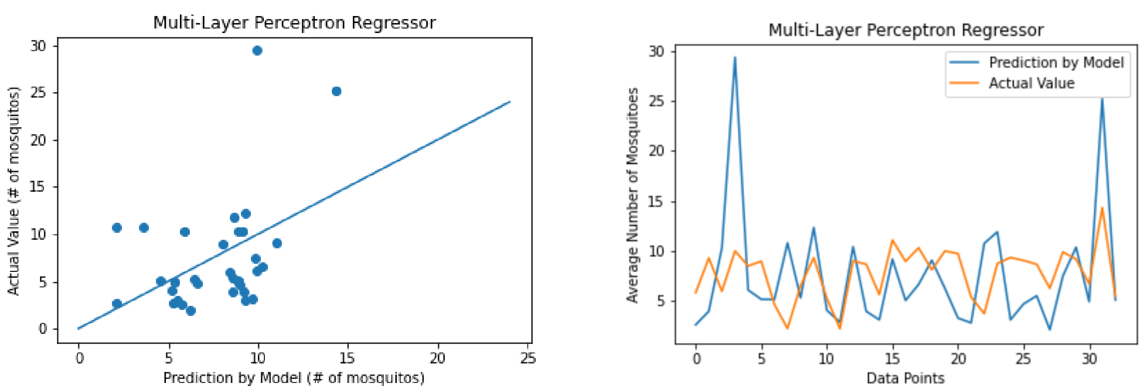


Figure 8: The relationship between predicted number of mosquitoes by model and actual number of mosquitoes for the Multi-Layer Perceptron Regressor. *Left:* A plot of prediction by the model vs actual number of mosquitoes. *Right:* Another representation with the prediction by model being plotted in blue and the actual value plotted in orange.



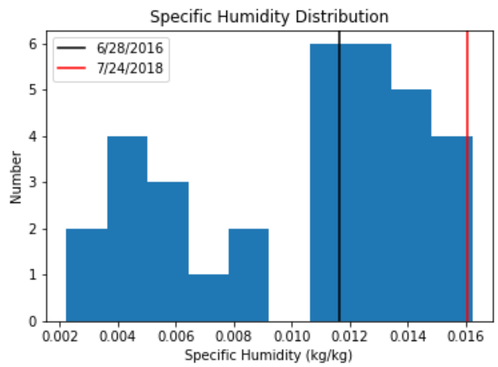
We also plotted a comparison of the predictions and the real mosquito values for all of the models in Figures 5, 6, 7, and 8. We found that all models struggled to correctly predict the values for the days July 24th, 2018, which had an average of 29.45 mosquitos, and for June 28th, 2016, which had an average of 25.27 mosquitos. Most models predicted that July 24th 2018 would have about an average of 10 mosquitos, and June 28th 2016 would have an average of 15 mosquitos. These two points had the highest value of all of them. In fact, they are about ten mosquitos more than the next largest. Furthermore, when these two points were removed, the Random Forest Regressor was able to significantly improve, shown in Table 6. The MAE was able to improve 0.65, while the RMSE was able to improve 1.71.

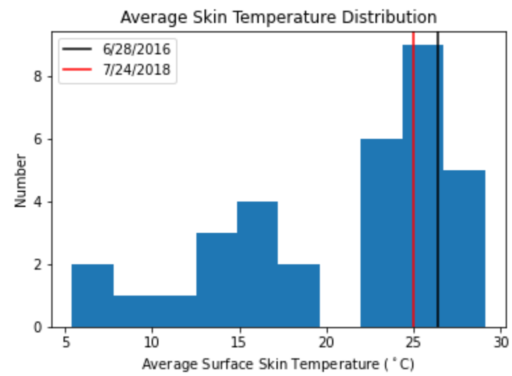
Table 6: The scores for Random Forest Regressor with and without the outliers.

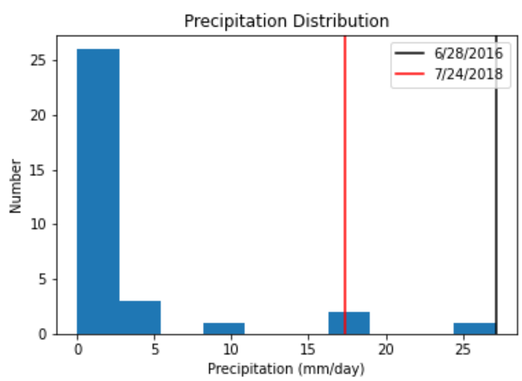
Condition	Mean Absolute Error	Root Mean Square Error
With Outliers	3.27046	5.14630
Without Outliers	2.62273	3.43861

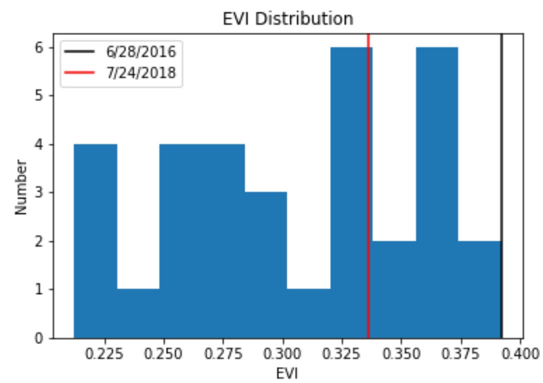
When we looked at the environmental factors for these, shown in Figure 9, we found that while these two data points did have higher values in most of the environmental factors, especially for June 28th, 2016, which had the highest value for two of the environmental factors. This is most likely why while the model predicted this point wrong, it still predicted it higher than the other points. However, for July 24th, 2018, it only had the highest specific humidity level, and was in the medium range for the others, which is why it most likely was not scored very high. Nevertheless, this represents how the change in environmental factors is not always comparable to the chance in the average number of mosquitos, especially when the average number of mosquitos is extremely high. It is thus better to have a greater amount of data.

Figure 9: The distributions of environmental factors for our test set versus those of our outliers. *Top Left:* The distribution for specific humidity. 7/24/2018 had one of the highest specific humidities. *Top Right:* The distribution for average surface skin temperature, both 7/24/2018 and 6/28/2016 had high temperatures. *Bottom Left:* The distribution for precipitation. 6/28/2016 had the highest precipitation. *Bottom Right:* The distribution for EVI. 6/28/2016 had the highest EVI.









Discussion

Our research presents an analytical and consistent viewpoint on the relationships between mosquito abundance and environmental factors. Our machine learning models were able to produce highly accurate predictions by creating comparisons between our factors and increases or decreases in mosquito abundance. Based on our results, we found that EVI has the biggest effect on mosquito abundance populations in our respective AOI. This idea can be applied to virtually anywhere across the globe, however we acknowledge that there are infinitely many anthropogenic and non-anthropogenic factors, including the variables that we used, that could affect mosquito abundance in any specific area. Several examples of literature have claimed that temperature and EVI have a significant effect on mosquito population growth, which partially supports our results taken from our machine learning models. Our research could have been improved in several ways. All of our mosquito data was taken from Washington D.C's public database, dated from 2016 to 2018. Due to climate change, environmental factors have changed drastically over the years, and our machine learning predictors are more suitable for analyzing 2016 trends. Due to the opportunistic nature of recording mosquito abundance in an AOI, we did not have complete mosquito abundance data when we ran our machine learning models, leaving us to only have 108 data points - which is relatively low compared to most training and testing sets. With even more consistent mosquito data and additional ecological variables, our models would have been able to create more accurate predictions about isolated areas around Washington D.C.

Conclusion

Similar to how we predicted, random forest regressor did provide the best model. However, we did find that all models did provide similar results. This shows that all machine learning models have proven to be useful predictors in analyzing patterns. Our model gave us a thorough, easily examinable way to understand the correlations between environmental factors and mosquito populations. Our machine learning models can be used for future use in predicting mosquito behavioral breeding patterns across various areas of interests, including but not limited to Washington D.C. Thus, this model can be utilized in various public health organizations to predict mosquito patterns and the diseases they can potentially carry. We hope future research projects will expand their areas of interest to geographical locations outside of Washington D.C and the United States of America, as performing machine model tasks in diverse areas can yield better predictors. Moreover, we hope more ecological variables are used in future research projects, as more variables can lead to more accurate predictions.

Acknowledgements

The authors would like to acknowledge the SEES Earth System Explorer mentors; Dr. Rusanne Low, Ms. Cassie Soeffing, Mr. Peder Nelson, Dr. Erika Podest, Andrew Clark, and Alexander Greco. The material contained in this paper is based upon work supported by the National Aeronautics and Space Administration (NASA) cooperative agreements NNX16AE28A to the Institute for Global Environmental Strategies (IGES) for the NASA Earth Science Education Collaborative (NESEC) and NNX16AB89A to the University of Texas Austin for the STEM Enhancement in Earth Science (SEES). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA.

The Team

Iona Xia

Monta Vista High School

Roles: Machine Learning Specialist, Data Management, Editor, Writer, Researcher

Jasmine Watson

Brewer High School

Roles: Editor, Writer, Researcher, Data Management

Neha Singirikonda

CHIREC International School, India

Roles: Data Management, Researcher, Writer, Editor

Landon Hellman

Dos Pueblos High School

Roles: Writer, Editor, Researcher, Data Management

Marvel Hanna

Huntington Beach High School

Roles: Editor, Writer, Researcher

IVSS Badges

I am a Collaborator

We are applying for this badge because we were a diverse team that created a comfortable atmosphere in which everyone's opinions were considered. Because we were from different schools and backgrounds, each of us had unique ideas, perspectives and skills that we took advantage of to achieve our common goal. By playing to our strengths and dividing the work optimally, we ensured that our research paper would be a product of our collective best effort. With qualities such as forethought, determination and logical thinking, we strengthened each other when encountering obstacles such as the lack of our desired GLOBE data or an unsatisfactory machine learning result. Working together enabled us to create better results for a more complex problem than would have been possible alone.

I am a STEM Professional

We collaborated and received guidance from STEM professionals such as Andrew Clark for inquiries and assistance regarding the utilization of GLOBE data and Dr. Rusty Low's help in our project ideation. This enabled us to provide a more professional analysis and interpretation of our machine learning results. Andrew Clark and Dr. Rusty Low's guidance helped us expand our sophistication and varying methods to consider approaching our data.

I make an impact

Our research provides a machine learning model that aids the community of Washington D.C. in analyzing and predicting mosquito habitats to prevent mosquito-borne disease outbreaks such as West Nile virus. We collected data from NASA Giovanni Earth science data website, GLOBE Observer, and Washington D.C. government data.

References

Basak, D., & Pal, S. (2007). Support Vector Regression. *Statistics and Computing*, *11*(10), 203–224.

Celestin, M. N., & Musteata, F. M. (2021). Impact of Changes in Free Concentrations and Drug-Protein Binding on Drug Dosing Regimens in Special Populations and Disease States. *Journal of pharmaceutical sciences*, *110*(10), 3331–3344. https://doi.org/10.1016/j.xphs.2021.05.018

Centers for Disease Control and Prevention. (2021, December 17). *Final cumulative maps and data*. Centers for Disease Control and Prevention. Retrieved July 22, 2022, from https://www.cdc.gov/westnile/statsmaps/cumMapsData.html

Centers for Disease Control and Prevention. (2020, December 7). *Prevention*. Centers for Disease Control and Prevention. Retrieved July 22, 2022, from https://www.cdc.gov/westnile/prevention/index.html

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (mae)? – arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Loh, W. Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

Environmental Protection Agency. (n.d.). EPA. Retrieved July 22, 2022, from https://www.epa.gov/mosquitocontrol/mosquito-life-cycle

Francisco, M. E., Carvajal, T. M., Ryo, M., Nukazawa, K., Amalin, D. M., & Watanabe, K. (2021). Dengue disease dynamics are modulated by the combined influences of precipitation and landscape: A machine learning approach. *Science of The Total Environment*, *792*, 148406. https://doi.org/10.1016/j.scitotenv.2021.148406

GLOBE, *Globe Data User Guide*. (n.d.). Retrieved July 23, 2022, from https://www.globe.gov/documents/10157/2592674/GLOBE+Data+User+Guide_v1_final.pdf/863a971d-95c5-4dd9-b75c-46713f019088

Open Data DC.(2021, December 8), *Mosquito trap sites*. Retrieved July 22, 2022, from https://opendata.dc.gov/datasets/DCGIS::mosquito-trap-sites/about

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, *2*(5-6), 183–197. https://doi.org/10.1016/0925-2312(91)90023-5

NASA. (n.d.). *Giovanni*. NASA. Retrieved July 22, 2022, from https://giovanni.gsfc.nasa.gov/giovanni/

National Centers for Environmental Information. (2022, May 11). *Washington D.C. Precipitation*. Retrieved July 27, 2022, from https://www.weather.gov/media/lwx/climate/dcaprecip.pdf

The GLOBE Program. (n.d.). Retrieved July 22, 2022, from https://www.globe.gov/

Schneider, J. E., Greco, A. W., Chang, J., Molchanova, M., & Shao, L. Y. (n.d.). *Predicting West Nile virus positivity rates and abundance: A ...* A Comparative Evaluation of Machine Learning Methods for Epidemiological Applications. Retrieved

July 23, 2022, from https://globe.gov/documents/10157/0/10754848/8715b9a3-de6e-47e6-53ee-9fb6fd5f2be4

Soh, S., & Aik, J. (2021). The abundance of Culex mosquito vectors for West Nile virus and other flaviviruses: A Time-series analysis of rainfall and temperature dependence in Singapore. *Science of The Total Environment*, *754*, 142420. https://doi.org/10.1016/j.scitotenv.2020.142420

Schonlau, M., & Zou, R. Y. (2020). The Random Forest Algorithm for Statistical Learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867x20909688

Washington, D.C. Topographic map, elevation, relief. topographic. (n.d.). Retrieved July 22, 2022, from https://en-nz.topographic-map.com/maps/sqll/Washington-D-C/

West Nile virus. West Nile Virus. (n.d.). Retrieved July 22, 2022, from https://dchealth.dc.gov/service/west-nile-virus