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

Iona Xia[1], Neha Singirikonda[2], Landon Hellman[3], Jasmine Watson[4], Marvel Hanna[5]
[1]Monta Vista High School, [2]CHIREC International School [3]Dos Pueblos High School, [4]Brewer High School, [5]Huntington Beach High School

The material contained in this poster 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.

Abstract

Mosquitoes pose a large threat to humans and other species due to their ability to carry deadly viruses, including the West Nile and Zika Viruses. Thus, tracking mosquito habitats and their breeding patterns is vital towards addressing public health concerns. Although fieldwork techniques have improved over the years, tracking and analyzing mosquitos is difficult, dangerous, and time-consuming. We plan to address this issue by creating a *Culex* mosquito predictor using machine learning techniques. We hope that by creating this, we will be better equipped to determine under which conditions the *Culex* mosquitoes thrive and reproduce. We hypothesized that precipitation levels have the most significant effect on mosquito populations. We used machine learning techniques in our area of interest (AOI), Washington D.C., to further predict mosquito breeding patterns. We used four environmental variables to conduct this experiment: precipitation, humidity, enhanced vegetation index (EVI), and temperature. We measured these variables in Washington D.C. using the NASA Giovanni Earth science data website. We determined the p-values of each ecological variable using an Ordinary Least Squares model. Using this data, we created various machine learning regression models to determine each ecological variable's significance in mosquito breeding patterns. Although there are infinitely many factors that can affect mosquito breeding patterns, we present our project data to health programs that can use our models to further benefit their disease observations and predictions. This research will shed light on the outcomes of our initial analysis and identify the steps we took throughout the process.

Introduction

- *Culex* mosquitoes are some of the most common species of mosquitoes in the world and can carry many of the infamous diseases that we know of today, including the West Nile (WNV) and malaria Viruses
- There is no established system that predicts mosquito-borne disease outbreaks, makes several societies prone to these deadly diseases
- Machine learning models have proven to be useful tools in predicting trends and operational patterns in various types of fields
- We tested four models: Random Forest, Decision Tree, Support Vector, and Multilayer regression algorithms which all allow users to train models that identify trends, predictions, and solutions in their own ways
- Washington DC, the capital of the United States, is known to have a humid subtropical climate which tends to be ideal for mosquito breeding habitats
- With 13 reported human cases in 2018 and 11 reported human cases in 2019, it is evident that not only does Washington D.C. have a major public health mosquito threat

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.

Research Questions

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

Methodology

- Data Collection
  - We obtained data that spans from April of 2016 to October of 2018 in Washington D.C.
  - From the NASA Giovanni Earth data collection website, we gathered the daily data regarding our environmental variables: Average Surface Skin Temperature, Specific Humidity, Precipitation and EVI
  - The Washington D.C. government provided open-sourced quantitative mosquito data in our respective AOI
  - We used the GLOBE data to analyze certain land cover observations in Washington D.C.
- Data Preprocessing
  - We narrowed the scope when needed and filled some holes in the environmental factor data by using SciPy's interpolation method interp1d, which we found to be the most accurate.

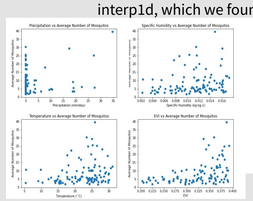

Figure 3: The correlation between average number of mosquitoes and our environmental factors.

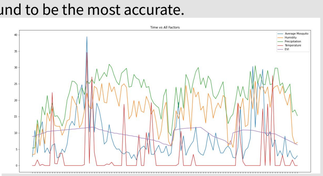

Figure 4: A graph of the number of average mosquitos and environmental factor for each of our data points. We found that peaks match up better with no lag.

- Analyzing Trends and Lag
  - We found the p-values of each of the environmental factors through the Ordinary Least Square Regression model.
  - 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.
  - Furthermore, we received a p-value < 0.05 for every ecological variable, proving that they all have statistically significant effects on mosquito populations.
- Training the 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.

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

Table 1: The p-values for our environmental factors with one week lag and no lag.

Hyperparameter	Values Tested	Chosen
'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 2: The hyperparameters for Random Forest

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)
  - MAE measures the average magnitude of the difference of the error, without caring about the direction
  - RMSE measures the square root of the average of the squared differences

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

Table 3: The Mean Absolute Error and the Root Mean Square Error of each of the four different models

- When comparing all four, we found that Random Forest performed better in both MAE and RMSE, and Support Vector Machine performed the worst in RMSE, and Multi-Layer Perceptron performed the worst in MAE

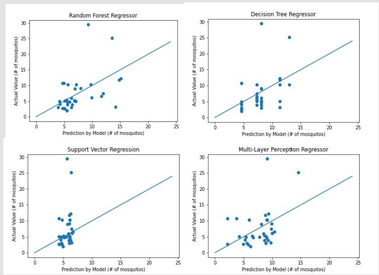

Figure 5: A graph of the predicted versus actual number of mosquitoes for each of the four models.

- Random Forest had a clump near 5 mosquitoes, but each point was relatively close to the line
- Decision Tree had a bunch of predictions in a line, which is an example of not a great fit, although because many points were close together, it still had a high MAE
- Support Vector had all its predictions less than around 7, meaning that it was unable to predict any high values
- Multi-Layer Perceptron was more similar to Random Forest, but its predictions were farther off

Results: Outliers

- 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
- When these two points were removed, the Random Forest Regressor was able to significantly improve (MAE by 0.81 and RMSE by 1.81)
- The change in environmental factors is not always comparatively to the chance in the average number of mosquitos, especially when the average number of mosquitos is extremely high

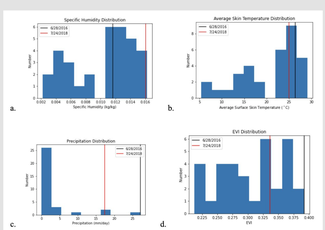

Figure 6: Comparison of Outliers and Environmental Factors
While these two data points did have higher values in most of the environmental factors, especially for 6/28/16, which had the highest value for two of the environmental factors. This is most likely why while the model predicted this point wrong, it still predicted it higher than the other points. For, 7/24/18, it only had the highest specific humidity level, and was in more of the medium range for the others, which is why it was not scored very high

Discussion

- Our research presents an analytical and consistent viewpoint on the relationships between mosquito abundance and environmental factors
- Found that the most influential factor was EVI
  - 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
- Even though our machine learning model is fairly accurate, there are still many sources of errors
  - 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

Conclusion

- We concluded that the random forest regression model provided the best model, even though all models provided relatively similar results
- While the R.F.R model worked slightly better, all models are able to be of use in predicting and analyzing patterns, providing us easy to understand correlations between environmental factors and mosquito breeding patterns
- In the future, we plan on including more land cover data, using citizen science and using the same models on other areas across the world
- We hope that this machine learning model will aid various public health organizations in more successfully predicting mosquito behaviors across numerous areas of interests (including by but not limited to washington D.C.) so that we can be better equipped to making strides towards greatly decreasing the number of mosquito-borne mosquitoes eventually world wide

References

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

Acknowledgements

We would like to acknowledge NASA, the SEES program, and UT Austin, as well as our SEES Earth System Explorer mentors: Dr. Rusanne Low, Ms. Cassie Soeffing, Mr. Peder Nelson, Dr. Erika Podest, Andrew Clark, and Alexander Greco.