Forecasting West Nile Virus Infections: A Machine Learning Approach to Epidemiological Monitoring

Rachel Chen¹, Aidan Schneider¹, Starlika Bauskar¹, and Francisco Rodriguez¹

¹NASA STEM Enhancement in Earth Science

July 3, 2023

Abstract

Mosquitoes are vectors for a number of serious illnesses, such as Dengue, Zika, Malaria, and West Nile Virus. In the United States, West Nile Virus (WNV) is the leading mosquito-borne disease. As there are currently no vaccines to prevent WNV nor medications to cure it, government agencies must sustain financially taxing programs to monitor mosquito populations and WNV infections and share this data across various departments in an effort to prevent WNV outbreaks. In this study, we develop four machine learning models that forecast WNV infections in humans, enabling government and healthcare officials to take proactive action instead of reacting to real-time infection data. Our models take in open-access data describing ecological variables – such as temperature, humidity, wind, air quality index (AQI), and enhanced vegetation index (EVI) — and use that data to predict future WNV infections five weeks in advance. We then perform a comparative analysis of the two types of machine learning architectures – support vector machine (SVM) regressors and random forest (RF) regressors – represented across our four models to evaluate which is best suited for the task. Our results indicate RF regressors are best suited to the task of forecasting WNV infections; however, SVM regressors perform comparably well and even exceed RF regressors when the magnitude of error is unweighted. Additionally, our results contribute a new perspective on the usefulness of AQI and wind speed for predicting mosquito-borne infections. Our RF regressor’s feature importance results indicate that AQI and wind speed were of similar importance as EVI and humidity – ecological variables well-known to influence mosquito population dynamics. Our work provides valuable directions for future research and development of early warning systems for disease prevention efforts as our models’ ability to forecast WNV infections five weeks in advance provides critical lead time for government officials to pursue mosquito containment efforts and healthcare facilities to increase capacity, enabling proactive action in combating WNV.

Link to abstract published at AGU’s Fall 2022 Session
Forecasting West Nile Virus Infections
A Machine Learning Approach to Epidemiological Monitoring

Rachel Chen, Aidan Schneider, Francisco Rodriguez, Starlaka Bauskar
NASA STEM Enhancement in Earth Science 2022

Abstract
Mosquitoes are vectors for a number of serious illnesses, such as Dengue, Zika, Malaria, and West Nile Virus. In the United States, West Nile Virus (WNV) is the leading mosquito-borne disease (2018-2022). As there are currently no vaccines to prevent WNV nor medications to cure it, government agencies must sustain financing tax programs to monitor mosquito populations and WNV infections in an effort to prevent WNV outbreaks. In this study, we develop four machine learning models that forecast WNV infections in humans, enabling government and healthcare officials to take proactive action instead of reacting to real-time infection data. Our models take data on ecological variables – such as humidity, wind, air quality, and vegetation – and use that data to predict future WNV infections five weeks in advance. We then present a comparative analysis of two types of machine learning models – support vector machine regressors and random forest regressors – to evaluate which is best suited for the task. Our results provide a streamlined solution for government agencies as they monitor WNV, enabling effective and low-cost preventative action.

Methodology

Area of Interest (AOI): Our AOI is the Southern California area comprised of Los Angeles Riverside and Orange County. We chose these counties since they have significant GLOBE data (see Fig. 1), open source WNV infection data, and significant changes in environmental variables across each mosquito season (see Fig. 2).

Ecological Variables and WNV: We converted Ecological Data and WNV data from daily data into averaged weekly data based on the CDC’s MMWR Epidemiological weekly format and limited to weeks 24-53 based on WNV data availability and consistency. We then chose our ecological data using means calculated for each year's mosquito season and padded our WNV data with weekly averages.

Data sources: California Department of Water Resources Irrigation Management Information System, United States Environmental Protection Agency, MODIS sensor outputs recorded on the NASA Aqua satellite, CHHS California Department of Public Health.

Results

Table 1 details the performance of our four machine learning models. While overall MAE and overall RMSE are overall error metrics, minimum RMSE and maximum RMSE describe the smallest and largest error values between any two points in the test set, providing another perspective on model performance. When comparing RMSE as a proportion of the desired output range for each model, the RF regresses clearly displays stronger performance than the SVMs. However, when comparing MAE as a proportion of the desired output range for each model, the four models display rather similar performance, with the RF SVM ultimately outperforming all other models. This trend persists in the minimum RMSE value, where all models perform closely but the RF SVM still outperforms its counterparts. This variation is likely a result of the nature of MAE and RMSE. MAE is linear in nature; therefore, it penalizes all errors equally, while RMSE is non-linear in nature and weights errors that are larger in absolute value more heavily (Chat & Draxler, 2014). With this understanding of error, we can conclude that the RF regressor is more likely than the SVMs as it is less likely to produce an error that is large in magnitude. Temperature emerges as the most important feature and precipitation as the least important, while EVI, AQI, wind speed, and humidity are all of similar importance (Table 2).

Discussion

Our results indicate that random forest regressors are the best machine learning architecture for this task; however, support vector machine regressors perform comparably well while exceeding random forest regressors when the magnitude of error is unweighted. Our results are particularly strong given the challenge of capturing varying valid data in a dataset that varies significantly week-to-week, due to delays between infection and reporting and the life cycle of Culex pipiens. The RF regressor’s feature importance reveals noteworthy correlations between our ecological variables and WNV infections. Most notably, EVI, AQI, wind speed, and humidity rank almost equal in importance. This is significant as, as detailed in our literature review, there is a lack of consensus on the importance of AQI and wind speed in mosquito prediction tasks. Our work suggests that AQI and wind speed are almost as important as vegetation and humidity metrics when aiming to predict disease characteristics in the southern California area. These findings reveal new research directions and provide a solid foundation for the continued development of early warning systems for forecasting WNV infections. However, our work also has potential for growth. For example, our models would benefit from more frequent WNV testing, as a more granular dataset with more frequent time steps would likely reveal new patterns that are currently obscured behind the weekly reporting structure and thereby reveal new opportunities to improve our predictions.

Conclusion

In summary, our machine learning models forecast the absolute number of WNV infections five weeks in advance using access ecological variables and remote sensing data. Our methodology and results hold valuable insight for the development of early warning systems that aid healthcare and government officials in preparing for and preventing incoming WNV outbreaks. Our predictions are particularly valuable when assessed from a resource allocation standpoint, as the five-week lead time they provide can aid healthcare providers in predicting when they must prepare to increase capacity. This early notice is critical to avoiding preventable deaths.

References

Visit this link: https://linker.ea/aidanschneider

Acknowledgements
A big thank you to our SEES Earth System Explorer mentors; Dr. Rosanne Lox, Ms. Cassie Steffing, Mr. Peter Nelson, Dr. Erica Podgurski and Andrew Clare. I would have been lost without their support and guidance.

The material contained in this poster is based upon work supported by the National Aeronautics and Space Administration (NASA) through NASA’s Education Opportunities for Global Environmental Strategies (IGES) for the NASA Earth Science Education Collaborative (NESEC) and NASA16ABSPA1 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.