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

Rachel Chen, Aidan Schneider, Francisco Rodriguez, Starlika 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 (CDC 2022). 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 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 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.

Introduction & Literature Review

Prior work has informed our decision to use Random Forest and Support Vector Machine models for this task, as they have consistently proven successful for a variety of mosquito prediction and classification tasks (Genoud et al. 2020, Früh et al. 2018, Wieland et al. 2017). The methodology of Lorenz et al. (2020) and Franklinos et al. (2019) demonstrated learning processes which evaluate mosquito-borne disease, supporting our use of Enhanced Vegetation Index (EVI) data derived the practicality of using remote sensing data in machine from NASA's Aqua and Terra satellites. Previous studies used weather variables such as temperature, precipitation, and humidity to predict mosquito abundance and transmission (Ligot et al. 2021, Buckner et al. 2011, Chuang et al. 2011); therefore, we included these variables as well. While Thiruchelvam et al. (2018) found little effect of AQI on the spread of disease, Gui et al. (2021) observed that extremely poor air quality and high wind speed could reduce the risk of Dengue transmission. Given the lack of scientific consensus and the similar oscillation patterns we observed between AQI and known significant ecological variables, we decided to include AQI in our model to assess its significance. These cases of previous research led to our decision to include humidity, temperature, precipitation, air quality, wind speed, AQI, and EVI in our models. We also included GLOBE data in our preliminary analysis, as several studies discussed the advantages of citizen science data, pointing out that citizen science programs such as the GLOBE Observer app's Mosquito Habitat Mapper and Land Cover facilitate consistency and utility (Carney et al. 2022 and Früh et al. 2018).

Figure 1: ArcGIS map of larvae count in Los Angeles, Orange, and Riverside counties.

Methodology

Area of Interest (AOI): Our AOI is the Southern Californian 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).

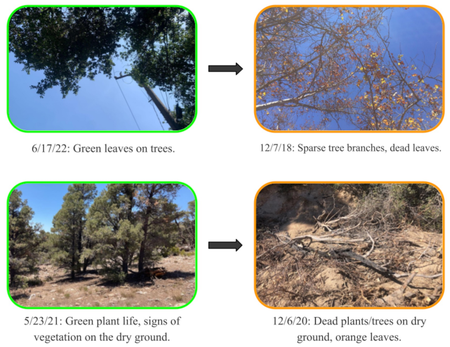

Figure 2: Land Cover Photo in our AOI documententing seasonal changes in vegetation like the process of leaves changing color and falling from trees as a result of cooling temperatures.

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 week format and limited to weeks 24-53 based on WNV data availability and consistency. We then padded our ecological data using means calculated across each year's mosquito season and padded our WNV data with zeros.
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.

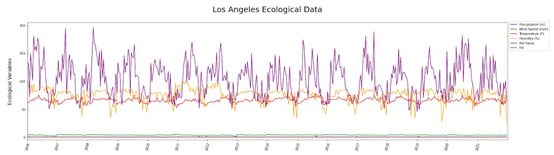

Figure 3: Los Angeles Ecological Variables Graph from data collected from 2006-2021 with a five week lag.

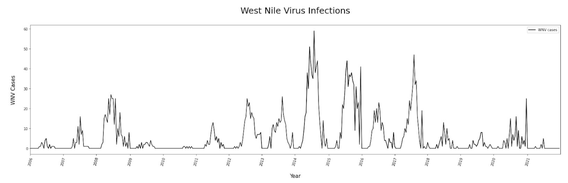

Figure 4: graph of West Nile Virus Infections from 2006-2021.

Time Lag: We tested various time lags for the ecological variables because Lopez et al. (2014), Ligot et al. (2021), and Schneider et al. (2021) emphasized the importance of the incorporation of time lag in order to obtain accurate predictions. We started by testing a three week time lag then evaluated five, six, and eight week lags. We found that a lag of 5 weeks aligned best with our WNV data.

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 regressor 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 RBF SVM ultimately outperforming all other models. This trend persists in the minimum RMSE value, where all models perform closely but the RBF 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 nonlinear in nature and weights errors that are larger in absolute value more heavily (Chai & Draxler, 2014). With this understanding of error, we can conclude that the RF regressor is indeed stronger 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).

Model	Overall MAE	Overall RMSE	Minimum RMSE	Maximum RMSE	Range of Desired Output
RBF SVM	0.514808	0.91283	9.0204E-05	0.37031	5.6596
Linear SVM	0.55336	1.0024	0.000167	0.39182	5.6596
Sigmoid SVM	0.54848	1.0086	0.00012	0.39083	5.6596
RF Regressor	5.74241	8.18072	0.00401	2.9433	59

Table 1: a variety of error metrics used to contextualize our 4 models' performance

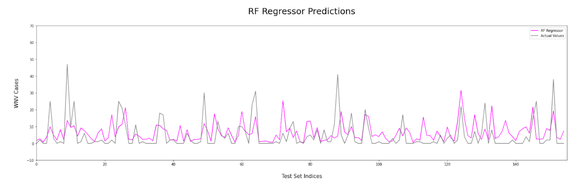

Figure 5: This graph describes our RF regressor's predictions in magenta and the actual WNV cases recorded in gray.

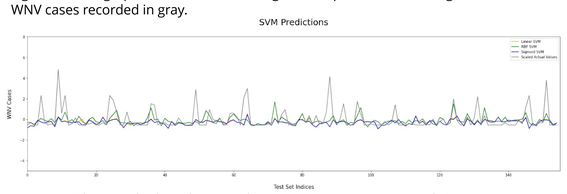

Figure 6: This graph describes our linear SVM regressor's predictions in orange, RBF SVM regressor's predictions in green, sigmoid SVM regressor's predictions in orange, and the actual WNV cases recorded in gray.

Ecological Variable	RF Feature Importance
Average Relative Humidity	0.14559
Average Air Temperature	0.41017
Precipitation	0.02372
Average Wind Speed	0.12840
AQI	0.12990
EVI	0.16222

Table 2: This table details our RF regressor's feature importance

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 and even exceed random forest regressors when the magnitude of error is unweighted. Our results are particularly strong given the challenge of predicting absolute values 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 importances reveal 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 open 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://linktr.ee/aidanschneider

Acknowledgements

A big thank you to our SEES Earth System Explorer mentors; Dr. Rusanne Low, Ms. Cassie Soeffing, Mr. Peder Nelson, Dr. Erika Podest, Andrew Clark, and Julianna Schneider!
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.