Predicting West Nile Virus Mosquito Positivity Rates and Abundance

A Comparative Evaluation of Machine Learning Methods for Epidemiological Applications

Julianna Schneider, Alexander Greco, Jillian Chang, Maria Molchanova, Luke Shao

NASA STEM Enhancement in the Earth Sciences 2021

Abstract

Mosquitoes are major vectors of disease and thus a key public health concern. Some cities have programs to track them, but such fieldwork is expensive, time-consuming, and retrospective. We present a comparative analysis of two machine-learning-based regression techniques for forecasting the derivatives of mosquito abundance and mosquito West Nile Virus (WNV) positivity in our AOI three weeks in advance. We used OLS regression to determine which of the climatic inputs utilized by prior work were statistically significant in predicting our desired outputs. We then trained four machine learning models, two Random Forest Regressors and two Backward Elimination Linear Regressions, and achieved RMSEs largely in the hundredths place or less. Our results indicate valuable directions for future research into forecasting mosquito population abundance and vector competence. This work is particularly applicable to public health programs, as our models' use of open-source, remote sensing data to predict how quickly the mosquito population and their vector competence will change three weeks in advance, streamlining disease monitoring and prevention.

Research Question

Which machine learning models and climatic inputs are most effective for predicting the derivatives of mosquito abundance and mosquito West Nile virus positivity?

Introduction & Literature Review

Machine learning models are powerful predictive tools, especially for regression-based tasks such as ours. The Random Forest Regressor is particularly applicable to our work, as our desired output consists of numerical metrics across a continuous time series. Prior work, such as Lee et al. (2016), also found success with multiple linear regression. Similarly, Project AEDES implemented backward elimination linear regression to predict the number of Dengue cases per month in a specified location, based on weather variables (temperature and rainfall) and google search trends (Ligot et. al, 2021). Hence, we evaluated the performance of a Random Forest Regressor (RFR) and Backward Elimination Linear Regression (BELR) for each prediction task.

Our selection of precipitation, temperature, humidity, and vegetation metrics as model inputs was informed by the success of prior work. Francisco et al. (2021) used monthly average precipitation, average land surface temperature, and flood susceptibility data to prove a significant correlation between precipitation and dengue outbreaks at a one-month lag in Manila, Philippines. Hassan et al. (2012) derived environmental variables such as urbanization level, Land Use Land Cover, Normalized Difference Vegetation Index (NDVI) from Landsat TM5 and Ikonos imageries to characterize landscape features likely associated with mosquito breeding habitats in Cairo, Egypt; land cover type and vegetation proved important indicators of potential mosquito habitats. Früh et al. (2018) trained a variety of machine learning models on citizen science data to predict the occurrence of *Aedes japonicus japonicus,* an invasive mosquito species in Germany. Their work indicated that mean precipitation, mean temperature, and drought index were the most accurate predictors of mosquito occurrence. Chen et al. (2019) indicated that landscape factors alone yield equal or more accurate modeling when compared to or paired with socioeconomic factors. Consequently, we pursued a hybrid citizen-science and government data approach where we evaluated the performance of a variety of machine learning regressors powered by the aforementioned ecological factors.

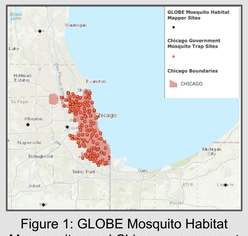

Figure 1: GLOBE Mosquito Habitat Mapper sites and Chicago government mosquito trap sites plotted on ARCGis

Figure 2: Our team records data through the GLOBE Observer mobile application

Materials

- West Nile Virus Mosquito Test Results obtained through the Chicago Data Portal
- Ecological datasets collected from Google Earth Engine: Land surface temperatures, near surface temperatures and precipitation from ERA5 provided by the European Centre for Medium-Range Weather Forecasts (ECMWF); Specific humidity from NLDAS-2 provided by NOAA/NCEP, NASA's Goddard Space and Flight Center, Princeton University, and the University of Washington; Day and night land surface temperature from the MODIS sensor on Aqua provided by NASA's EOSDIS; EVI (Enhanced Vegetation Index) from the MODIS sensor on Aqua provided by Google and NASA's EOSDIS.

Methodology

Data Retrieval and Cleaning: We used Google Earth Engine to export satellite and weather data in weekly timesteps. We obtained most of our ecological data from ERA5 as it aggregated the variables we needed in one place at a uniform, high level of measurement precision. We chose the Aqua satellite data for the day and night temperatures as it provided a full dataset across our time series of interest, 2007-2020 and complete EVI data within weekly timesteps. We separated these ecological variables into 731 week-long timesteps ranging from December 31, 2006 and January 2, 2021 to align with the CDC's epidemiological year.

The City of Chicago's open access West Nile Virus Mosquito Test Results dataset provided the output we aimed to predict. It contains the results from Gravid and CDC mosquito traps located across the City of Chicago measured on a weekly basis throughout summer from 2007 - 2021. This data provided two crucial metrics for our project: the number of Culex mosquitoes captured at each trap and the number of Culex mosquitoes captured that tested positive for West Nile Virus, meaning they were capable of transmitting it. During data cleaning, the Gravid trap data was isolated, the weekly measurements were aligned with the epidemiological year, weekly mosquito abundance was quantified as the number of mosquitoes divided by the number of total traps in the area, and weekly West Nile Virus positivity rate was quantified as the number of mosquitoes testing positive for the disease divided by the mosquito abundance (figure 3). Points of discontinuity across the summer months were identified and analyzed: 2009 and 2011 displayed the least continuity. Weeks 22 through 40 emerged as the widest common range across the data, so we filled in the missing data points for weeks 22 through 40 every season using SciKit Learn's imputer's Most Frequent filling method. Aiming to mesh this dataset with citizen science GLOBE Mosquito Habitat Mapper (MHM) data as Früh et al. (2018) did, both the City of Chicago data and Cook County MHM data were mapped against the boundaries of the City of Chicago in ArcGIS – no MHM points fell within city limits (Figure 1). Although the GLOBE Observer project's data was not applicable to our study, our search revealed that GLOBE MHM data was remarkably comparable to official, government-collected mosquito trap data — we hope future improvements on our models will utilize that potential.

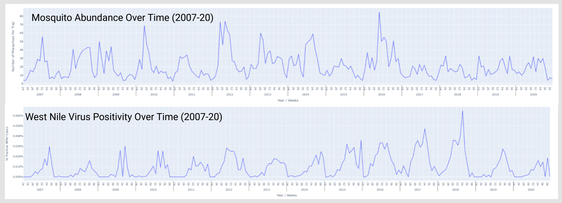

Figure 3: Mosquito abundance and mosquito WNV positivity data graphed over time.

Pre-processing: Lopez et al. (2014) observed higher correlations between dengue outbreaks and environmental factors when time lags were introduced. Inspired by this work, we examined the relationship between the various climatic variables collected from our literature review and the mosquito abundance and positivity outputs by graphing them. Shifting EVI, Land Surface Temp, and Specific Humidity (as it relates to mosquito abundance), and total precipitation (as it relates to West Nile Virus positivity rates) three weeks forward in time better aligned the input and output peaks, as seen in Figure 4. We then padded our dataset, extending the weeks in each summer to 21-41, so as to calculate the derivative of mosquito abundance and WNV positivity for our weeks of interest, 22-40. We elected to predict the derivatives of mosquito abundance and WNV positivity as it enables our models to act as predictors for the state of the mosquito population in our AOI, as opposed to predicting what quantities would be observed in government traps.

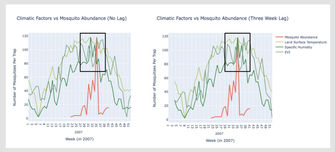

Figure 4: Introducing three-week lag into our climatic variables to help their peaks align more closely with the peaks in mosquito abundance - same technique used for positivity.

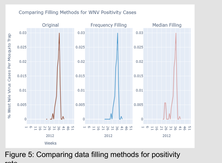

Figure 5: Comparing data filling methods for positivity rate.

Feature Selection and Model Training: Having assembled our initial pool of independent climatic variables based on the findings of prior work, we narrowed down our pool of inputs using ordinary least squares (OLS) regression. First, we ran OLS regression on each input individually to establish which had statistically significant correlations with mosquito abundance and which had statistically significant correlations with mosquito WNV positivity using a p-value of 0.05. Then, we grouped the promising climatic inputs for each prediction task into various sets and ran OLS regression on each set, revealing EVI, land surface temperature, total precipitation, and specific humidity as the optimal inputs for predicting mosquito abundance and EVI, specific humidity, near-surface temperature range, and night-time temperature as the optimal inputs for predicting mosquito WNV positivity. We then trained an RFR and BELR to predict the derivative of mosquito abundance and an RFR and BELR to predict the derivative of mosquito WNV positivity. The RFRs were built using SciKit-Learn's RFR model and trained using its Randomized Search Cross Validation tool. On running 100 iterations with three cross folds each, the optimal hyperparameters for each RFR emerged. The BELRs were built using Sci-Kit Learn's Linear Regression model and inputs were eliminated using OLS regression to determine which inputs were statistically insignificant to the BELR's predictions.

Results

Tables 1 and 2 detail the performance of the RFR and BELR models for each prediction task using overall MAE, overall RMSE, maximum RMSE, and minimum RMSE. Overall MAE and RMSE provide a single value describing the prediction error for the entire testing set, while Max RMSE and Min RMSE provide the maximum and minimum values from the RMSE calculated at each time step in our testing set. We opted to provide our general error metric in both MAE and RMSE as each provides a different view into model performance: while MAE's linear nature results in equal weight given to all errors, RMSE's nonlinear nature further penalizes errors that are larger in absolute values (Chai & Draxler, 2014). Figures 6-9 provide a graphical representation of the RMSE calculated at each time step.In comparing the overall MAE and RMSE values for the RFR and the BELR models used for each task, the BELRs outperforms the RFRs. However, the RFR model for predicting the mosquito abundance derivatives has a minimum RMSE almost 3 times smaller than the BELR's. Similarly, the RFR model for predicting the mosquito WNV positivity derivatives has a minimum RMSE almost 2.5 times smaller than its BELR counterpart. This indicates that the RFR models are capable of predicting the desired output more closely than the BELR models: a result supported by RFR's ability to fit nonlinear data, as opposed to BELRs which can only fit linearly.

Table 3 compares the overall RMSE of our RFR and BELR models to that of a similar study by Lee et. al that aimed to predict mosquito abundance using a multiple linear regression (MLR) and an artificial neural network (ANN). Our mosquito abundance derivative models display a lower overall RMSE for larger ranges in the desired output. Additionally, our mosquito WNV positivity derivative models' overall RMSE comprises a smaller fraction of the desired output's range than that of Lee et. al. Given the high variability of the desired mosquito population characteristic output and the extreme outliers evident at points such as week 29 in 2016 in Figures 6-9, our models' errors are comparatively low and demonstrate strong overall performance.

Model	Overall MAE	Overall RMSE	Max RMSE	Min RMSE
Random Forest Regressor	5.31120471 1842105	7.47669616 537873	3.69971047 29924346	0.00092115 9306679115 9
Backward Elimination Linear Regression	4.45476933 2456141	6.69620884 3026129	3.31894896 83901288	0.0027213 9 5246351403

Table 1: Error results for RFR and BELR models used to predict mosquito abundance derivatives.

Model	Overall MAE	Overall RMSE	Max RMSE	Min RMSE
Random Forest Regressor	0.004442247 8389793	0.006522116 317871164	0.002433148 7872215904	1.378654314 6482505e-06
Backward Elimination Linear Regression	0.004278856 9042500006	0.006450763 508888683	0.002520300 9067467017	3.539567595 920732e-06

Table 2: Error results for RFR and BELR models used to predict mosquito WNV positivity derivatives.

Model	Overall RMSE	Range of Desired Output
RFR for Mosquito Abundance Derivative	7.47669616537873	98
BELR for Mosquito Abundance Derivative	6.696208843026129	98
RFR for Mosquito WNV Positivity	0.006522116317871164	0.0435
BELR for Mosquito WNV Positivity	0.006450763508888683	0.0435
MLR for Mosquito Abundance	17.53	78
ANN for Mosquito Abundance	14.38	78

Table 3: Comparison of overall RMSEs for our RFR and BELR models and Lee et. al's MLR and ANN.

Figure 6: Random Forest Regressor for Predicting the Derivative of Mosquito Abundance

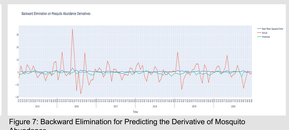

Figure 7: Backward Elimination for Predicting the Derivative of Mosquito Abundance

Figure 8: Random Forest Regressor for Predicting the Derivative of Mosquito West Nile Positivity

Figure 9: Backward Elimination Linear Regression for Predicting the Derivative of Mosquito West Nile Positivity

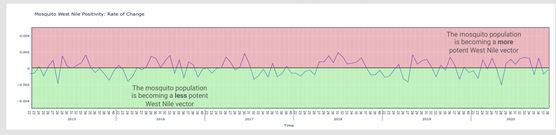

Figure 10: Annotated graph of Random Forest Regressor's predictions of the derivatives of mosquito WNV positivity rate illustrating the public health program applicability of our finds.

IVSS Badges

I am a Collaborator: We are applying for this badge because we were a strong team that collaborated in an environment built on teamwork where each member used their skills to contribute to the project in a meaningful way. Working with students from different schools and backgrounds improved our research by providing a broader pool of knowledge and skills from which we could pull in developing our project. By combining our diverse perspectives with these resources, we developed new approaches to solve the challenge of effectively tracking and monitoring mosquito abundance and vector competence. Working together enabled us to tackle a more complex problem than we could have on our own.

I am a Data Scientist: We analyzed a multitude of datasets, including the GLOBE Mosquito Habitat Mapper data (which contains our contributions as NASA SEES interns), Chicago Public Health mosquito trap data, and remote sensing datasets from various NASA satellites accessed through the Google Earth Engine. We analyzed and discussed the issues with and limitations of our data within our report and selected the best datasets for our tasks based on these analyses. We identified patterns in our data and made successful predictions based on these patterns using RFRs and BELRs.

I am a STEM Professional: We collaborated with several STEM professionals through the NASA SEES program at which this research was conducted. We worked closely with Dr. Erika Podest who provided invaluable guidance throughout our development process. Dr. Podest shared her team's paper on identifying statistically significant correlations between Dengue outbreaks and ecological factors in Brazil and suggested that we implement a similar time lag into our environmental variables; she provided feedback as we worked to select the best scope and timestep for our predictions using the data available to us; and advised us on how to interpret our findings in our report. In short, Dr. Podest's guidance helped us expand our viewpoints and consider new ways of approaching the problem we wanted to tackle.

Discussion

In this study, we present a comparative evaluation of four machine learning models for two mosquito abundance and vector competence derivative prediction tasks and assess the statistical significance of a variety of climatic inputs for doing so. Our results show that these models improve on prior work's ability to predict how quickly the mosquito population is growing or declining and how quickly mosquitoes are becoming disease vectors for West Nile in an AOI.

Particularly noteworthy is how temperature was a crucial input in all of our models, but each model performed better with a different temperature metric or combination of metrics. The RFR for predicting the mosquito abundance derivative preferred full day surface temperature; the RFR for predicting the WNV mosquito positivity derivative preferred a combination of full day near surface temperature and full day land surface night temperatures; the backward elimination model for the abundance derivative preferred full day surface temperature; and the backward elimination model for positivity preferred land surface night temperatures alone.

Unlike much of the literature that informed our research, precipitation did not prove a significant factor across our machine learning models. However, indirect measurements of water quantity, such as EVI, did prove crucial and common across all models. This may be the result of differences between our AOI and that of other studies or the OLS regression we used to narrow down our climatic inputs, which only fits — and therefore deems significant — linear correlations.

These findings, among others elaborated in our report, provide avenues for further research and a deeper understanding of how mosquito populations thrive and become more potent disease vectors in response to climatic variation. Similarly, there remain areas for improvement upon our research. First, we applied our methodology to a single area of interest — to test its robustness, future work should see how well the development procedure adjusts to different areas of interest. Second, we averaged data over the entirety of Chicago, making our predictions applicable to the whole of Chicago but not specific to a single area within it. With more consistent and detailed data recorded on more frequent time steps, our model would likely perform better and output predictions further localized to mosquito and West Nile hotspots within the greater City of Chicago.

Conclusion

In summary, our models can accurately predict the derivatives of mosquito abundance and mosquito WNV positivity in our AOI. The BELRs slightly outperformed the RFRs in terms of overall metrics, however, the RFRs proved significantly more capable of fitting the observed mosquito values in weekly RMSE. Our methodology and results hold potential for valuable applications to public health programs and concerns. As our ecological variables are lagged three weeks forward in time for training purposes, our models can be used in real-time as predictors for the derivatives of mosquito abundance and WNV mosquito positivity three weeks in advance — providing public health officials with critical information on the development of mosquito populations in time for appropriate intervention and mitigation as seen in Figure 10. Avenues for future work and development on our results revolve around how to further increase accuracy and best incorporate our methodology into existing public health initiatives.

Bibliography

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, *114*, 24–31.

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.

Chen, S., Whiteman, A., Li, A., Rapp, T., Delmelle, E., Chen, G., Brown, C. L., Robinson, P., Coffman, M. J., Janies, D., et al. (2019). An operational machine learning approach to predict mosquito abundance based on socioeconomic and landscape patterns. *Landscape Ecology*, *34*(6), 1295–1311.

Ding, F., Fu, J., Jiang, D., Hao, M., & Lin, G. (2018). Mapping the spatial distribution of aedes aegypti and aedes albopictus. *Acta tropica*, *178*, 155–162.

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*, 148406.

Früh, L., Kampen, H., Kerkow, A., Schaub, G. A., Walther, D., & Wieland, R. (2018). Modelling the potential distribution of an invasive mosquito species: Comparative evaluation of four machine learning methods and their combinations. *Ecological Modelling*, *388*, 136–144.

GLOBE. (2021). Global learning and observations to benefit the environment (globe) program. globe.gov

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, *202*, 18–27.

Hassan, A. N., El Nogoumy, N., & Kassem, H. A. (2013). Characterization of landscape features associated with mosquito breeding in urban cairo using remote sensing. *The Egyptian Journal of Remote Sensing and Space Science*, *16* (1), 63–69.

Koolhof, I. S., Gibney, K. B., Bettiol, S., Charleston, M., Wiethoelter, A., Arnold, A.-L., Campbell, P. T., Neville, P. J., Aung, P., Shiga, T., et al. (2020). The forecasting of dynamical ross river virus outbreaks: Victoria, australia. *Epidemics*, *30*, 100377.

Lee, K. Y., Chung, N., & Hwang, S. (2016). Application of an artificial neural network (ann) model for predicting mosquito abundances in urban areas. *Ecological Informatics*, *36*, 172–180.

Ligot, D., Toledo, M., & Melendres, R. (2021). Project aedes dpg repository wiki. https://github.com/Cirrolytix/aedes_dpg/wiki

Luman, D., Tweddale, T., Bahnsen, B., & Willis, P. (2004). Illinois land cover: Champaign, il, illinois state geological survey, illinois map 12, scale 1:500,000. https://files.isgs.illinois.edu/sites/default/files/maps/statewide/imap12.pdf

Orkins. (2021). Orkin's 2021 top mosquito cities list. https://www.orkin.com/press-room/orkins-2021-top-mosquito-cities

Roberts, M. (2021). West nile virus (wnv) mosquito test results. https://data.cityofchicago.org/Health-Human-Services/West-Nile-Virus-WNV-Mosquito-Test-Results/jqe8-8r6s

US Department of Commerce, N. (2021). Annual precipitation rankings for chicago, illinois. https://www.weather.gov/lot/Annual_Precip_Rankings_Chicago

Acknowledgements

Our gratitude goes to our NASA SEES 2021 mentors Dr. Rusanne Low, Ms. Cassie Soeffing, Mr. Peder Nelson, Dr. Erika Podest, and Dr. Becky Boger!

The material contained in this poster is based upon work supported by 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.