Predicting West Nile Virus Positivity Rates and Abundance: A Comparative Evaluation of Machine Learning Methods for Epidemiological Applications

Julianna E. Schneider, Alessandro W. Greco, Jillian Chang, Maria Molchanova, Luke Y. Shao

NASA STEM Enhancement in the Earth Sciences 2021

Author Note

Correspondence concerning this article should be addressed to Julianna E. Schneider. E-mail: julianna.e.schneider@gmail.com

Abstract

Mosquitoes are major vectors of disease and thus a key public health concern. Some cities have programs to track mosquito abundance and vector competence, but such fieldwork is expensive, time-consuming, and retrospective. We present a comparative analysis of two machine-learning-based regression techniques for forecasting the rate at which mosquito abundance changes and the rate at which mosquitoes test positive for West Nile Virus (WNV) in our AOI, the City of Chicago, three weeks in advance. We selected an initial pool of climatic inputs based on the findings of prior work. Ordinary least squares regression was run on each input individually and then in various groups. A p-value cutoff of 0.05 was used to determine which were best suited for predicting the derivatives of mosquito abundance and WNV positivity rate. Using these inputs, we trained four machine learning models using two types of regression: a Random Forest Regressor (RFR) and Backward Elimination Linear Regression (BELR). We optimized our RFR's hyperparameters using Randomized Search Cross Validation and further reduced our BELR inputs using a p-value of 0.05. The enhanced vegetation index and temperature, described in various metrics, emerged as common inputs across the four models. In three of the four models, the respective temperature metric was the most important feature, while EVI varied between second and last place. Our root mean square error largely resided within the hundredths place or less, but spiked at novel, week-to-week extremes in the testing data. Our methodology and results indicate valuable directions for future research into forecasting mosquito population abundance and vector competence. This work is particularly applicable to public health programs – our models' use of open-source, remote sensing data to predict, three weeks in advance, how quickly the mosquito population and their vector competence will change streamlines disease monitoring and prevention.

Keywords: mosquito-borne disease, mosquito abundance, remote sensing, machine learning, West Nile virus

Predicting West Nile Virus Positivity Rates and Abundance: A Comparative Evaluation of Machine Learning Methods for Epidemiological Applications

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 and Review of Literature

Mosquitoes are vectors for several highly infectious diseases, including dengue, malaria, and West Nile Virus (WNV). Therefore, tracking mosquitos is crucial for preventing and combating outbreaks of mosquito-borne disease. Currently, tracking mosquito abundance and vector competence relies on government-funded fieldwork in which mosquito traps are continually tended to and monitored as a means of estimating these metrics. This process is expensive, time-consuming, and retrospective. Machine learning models thus hold valuable potential for streamlining and expanding this process through their predictive abilities. We present a comparative analysis of two machine-learning-based regression techniques for forecasting the rate at which mosquito abundance changes and the rate at which mosquitoes test positive for WNV.

Machine learning models are powerful predictive tools, particularly for regression-based tasks such as ours. The Random Forest (RF) Classifier is popular within the remote sensing community due to its ability to handle high data dimensionality and its insensitivity to overfitting (Belgiu & Drăguţ, 2016). When stimulating spatial distribution of arbovirus vectors, RF models obtained the highest accuracy (Ding et al., 2018). Given the high functionality and success of the RF Classifier, we looked into its regressor counterpart to fit our prediction goals. 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. In particular, 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. (2013) 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.

Research Methods

Our Area of Interest

Chicago, Cook County, is located in the midwestern United States along Lake Michigan. It has a distinct, four season climate, with hot, humid summers and cold, windy

winters, and an average annual precipitation of 34 inches (US Department of Commerce, 2021). The city's natural topography is flat, though there are steep bluffs and ravines along Lake Michigan in the north, and sand dunes to the south. Land cover is made up almost entirely of high-density urban areas, with few spots of forested land (Luman et al., 2004). Chicago's mosquito infestation has also long been a serious public health concern. For the past five years, Orkin has ranked Chicago in the top five cities with the most mosquitoes; a recent examination of Chicago mosquito surveillance data also revealed high West Nile Virus incidence in the city (Orkins, 2021; Roberts, 2021). Cook County experienced major outbreaks of West Nile Virus in recent years — 98 human cases in 2016, 52 in 2017, and 104 in 2018 were reported to CDC's ArboNet. Ultimately, our decision to focus on Chicago as our area of interest (AOI) was largely motivated by both the prevalence of mosquitoes and West Nile virus and the availability of comprehensive, open-source mosquito data.

Data Retrieval and Pre-processing

The following procedures were employed to translate climatic inputs and mosquito abundance and WNV positivity outputs of various resolutions to a common time scale and scope. The time scale utilized was the CDC's epidemiological weeks and our scope was the entirety of the City of Chicago. Our final dataset consisted of our climatic inputs and the derivatives of mosquito abundance and WNV positivity recorded from weeks 22 to 40 of each year from 2007-2020.

Obtaining Ecological Variables. We used Google Earth Engine to export satellite and weather data in weekly timesteps. We retrieved hourly land surface and near-surface temperatures and precipitation from ERA5 provided by the European Centre for Medium-Range Weather Forecasts (ECMWF); hourly specific humidity from NLDAS-2:North American Land Data Assimilation System Forcing Fields provided by National Centers for Environmental Prediction (NCEP), Goddard Earth Sciences Data and Information Services Center (GES DISC), Princeton University, and the University of

Washington; and day and night land surface temperature from the MODIS sensor on Aqua provided by Land Processes Distributed Active Archive Center (LP DACC). The daily EVI (Enhanced Vegetation Index) from the MODIS sensor on Aqua was provided by Google. We obtained most of our ecological data from ERA5 as it aggregated the independent variables we needed in one place at a uniform, high level of 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. We also obtained our EVI measurements from Aqua, as it provided the only dataset with small enough timesteps to fit within the weekly timestep of our mosquito data and provided uniformity across the night and day temperatures and EVI measurements. We separated these ecological variables into 731 week-long timesteps ranging from December 31, 2006 to January 2, 2021: these weeks align with the CDC's epidemiological year. To calculate the climatic variables' means, their respective data points within each epidemiological week were averaged. The weekly averages for all sample points within the city limits of Chicago were then averaged once more to produce a single value describing the City of Chicago for each metric, week by week. To calculate the maximum and minimum values, the maximum and minimum data points within each week were selected. Then, the weekly maximums and minimums for all sample points within the city limits of Chicago were averaged to produce a single value for each week. To calculate weekly total precipitation, the data points within each week were summed, then the weekly sums for all sample points within the city limits of Chicago were averaged to produce a single value for each week. The city limits of Chicago used for filtering our data points were obtained from the official City of Chicago website.

Obtaining GLOBE Mosquito Habitat Mapper Data. With the objective of conducting a comparative analysis of machine learning models using citizen science data, similar to Früh et al. (2018), our team initially analyzed the Mosquito Habitat Mapper (MHM) dataset from GLOBE Observer, which records user egg and larvae submissions from natural and artificial mosquito habitats. As interns in the NASA SEES program, we

contributed to this global dataset with our own mosquito observations. Our data contribution consisted of two parts: 3-5 artificial mosquito traps maintained for 6 weeks as well as mosquito habitat and land cover data recorded in an evenly spaced, 3 kilometers square, 36-point grid. All observations were taken near the interns' residences across the West Coast, East Coast, and Western Canada. The traps were made of artificial containers and baited with dog food, fermented grass, fish food, timothy hay, or pond water. The traps were checked weekly and observations were recorded using the GLOBE Observer app's MHM function; MHM and Land Cover observations at each point of the 36-point grid were recorded as well. We processed mosquito data uploaded by citizen scientists through the GLOBE Observer app using the go_utils Python library provided by GLOBE and the go_utils API. To visualize the data and determine possible AOIs, we collected all United States MHM recordings starting from the inception of MHM in 2017 to the present day. We then cleaned the data, organizing the submissions by state, county, and city, and removing any logs that did not submit egg/larvae count. The remaining points were mapped on ArcGIS, and the following regions were pinpointed as areas with the most data: Los Angeles, California and Harris County, Texas. Although the two counties had the greatest number of MHM GLOBE submissions relative to other areas, the frequency of recordings was inconsistent. To supplement the MHM data, we reached out to the governments of Los Angeles and Harris for access to the mosquito data collected through local government initiatives. Due to time and legal constraints, our request could not be accommodated. In our search for an open-source dataset, we found the highly comprehensive City of Chicago West Nile Virus Mosquito Test Results dataset, which contained consistent data on the number of mosquitoes captured through Gravid and CDC traps from 2007 to 2021 and how many of those mosquitoes tested positive for WNV. In an attempt to mesh this dataset with MHM data, both the City of Chicago data and Cook County MHM data were mapped against the boundaries of the City of Chicago in ArcGIS (see Figure 2). No MHM points fell within city limits; as our largest dataset resided solely

within the City of Chicago, we could not mesh it with MHM data. Land cover data from GLOBE Land Cover in Chicago was similarly cleaned, but also yielded insufficient quantity within our AOI for analysis. 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.

Pre-processing the City of Chicago's West Nile Virus Mosquito Test Results. The City of Chicago's open access West Nile Virus Mosquito Test Results data was downloaded through the Chicago Data Portal. This data 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. This data was cleaned in Python using Pandas, Numpy, SciKit Learn, and Epi Weeks. First, the CDC trap data was removed, as our study focuses on the results of Gravid trap data alone. Then, the weekly measurements were aligned with the epidemiological year using Epi Weeks and the date on which each record was logged. We quantified weekly mosquito abundance as the number of mosquitoes divided by the number of total traps in the area, and West Nile Virus positivity rate as the number of mosquitoes testing positive for the disease divided by the mosquito abundance. 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: seven out of the total 13 years contained records for either week 22, week 40, or both. We filled in the missing data points for weeks 22 through 40 every season using SciKit Learn's imputer. We graphed the original dataset against the dataset when filled with the Median and Most Frequent methods. Filling in weeks of missing data for the number of mosquitoes observed (Figure 7) and for the number of positively tested mosquitoes (Figure 8) was

most accurate when using the most frequent value for each respective season.

Introducing Lag. 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. We graphed the climatic variables vs. mosquito abundance and the climatic variables vs. West Nile Virus positivity using Plotly, an open-sourced Python graphing library. In doing so, we aimed to observe the extent and necessity of shifting our weather and land cover variables to account for any delayed effects on mosquito abundance or WNV positivity. Although the two line plots suggested positive correlations between the environmental variables and mosquito prevalence/disease positivity rates, time lags seemed to occur between the inputs and outputs. We therefore shifted 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. As shown in Figure 9, shifting Land Surface Temp, Specific Humidity, and EVI three weeks forward yielded more similar peaks in relation to the mosquito abundance output. Figure 10 also demonstrates that similar peaks in total precipitation and West Nile Virus positivity rates were achieved after shifting precipitation three weeks forward. Thus, after correcting for the time lags, the peaks in both graphs were better aligned, producing results with higher correlations - as later solidifed through OLS regressions seen in Table 1 - between ecological inputs and mosquito and vector competence outputs.

Data Characteristics. Having identified the optimal filling method and lags, we then padded our data using the Most Frequent filling method, extending the weeks in each summer to 21-41. This enabled us 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. While predicting the raw mosquito

abundance and WNV positivity numbers would describe what quantities public health officials could expect to see in their traps, knowing the derivatives of these metrics provides a holistic view into how the mosquito population in the AOI is changing and how quickly they are becoming more or less potent vectors of disease. The mean, standard deviation, and range of the ecological inputs and mosquito population characteristics outputs are provided in 2. The derivatives of mosquito abundance and mosquito WNV positivity display high variability, as evidenced by their standard deviations that both vary by 10^3 from their means at times.

Developing the Machine Learning Models

The following procedures were employed to select statistically significant climatic inputs for predicting the derivatives of mosquito abundance and WNV mosquito positivity. The inputs that proved statistically significant for each prediction task were then used to train a Random Forest Regressor (RFR) and Backward Elimination Linear Regression (BELR), totaling four machine learning models.

Narrowing Down the Pool of Independent Variables. Having assembled our initial pool of independent climatic variables (see 3) 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 (see Table 1). The results from the mosquito abundance OLS regressions revealed EVI and land surface temperature as a statistically significant pairing on their own and when incorporated into most groups; hence, we selected these two inputs. Based on the established relationships between water and mosquito abundance, we looked into which of our many water metrics were the best indicators. Total precipitation proved to be

the most statistically significant of our precipitation metrics. Specific humidity also displayed statistical significance on its own and when paired with various other metrics, such as EVI. However, both these water inputs were rendered statistically insignificant in the OLS regression when paired on their own or together with EVI and land surface temperature. We hypothesized this behavior was a result of OLS regression's simplicity paired with the non-linearities in the EVI, land surface temperature, total precipitation, and specific humidity inputs, as evident throughout our graphs. Consequently, these four indicators were used as the inputs to our mosquito abundance derivative RFR and BELR models. The results from the WNV mosquito positivity OLS regressions indicated that temperature was a statistically significant indicator; however, multiple temperature metrics were statistically significant in different circumstances. Near-surface temperature, surface temperature, and the near-surface temperature range were statistically significant on their own, but not when grouped with other temperature metrics. Meanwhile, night-time temperature was statistically significant on its own and when paired with day-time temperature, but became statistically insignificant when paired with near-surface temperature range. near-surface temperature range and average near-surface temperature were statistically significant when paired with EVI and specific humidity, but they rendered EVI and specific humidity notably insignificant with p-values of 0.194 and 0.296 and 0.446 and 0.762, respectively. Similarly, night-time temperature was statistically significant when paired with EVI and specific humidity, but rendered EVI and specific humidity statistically insignificant with p-values of 0.059 and 0.337, respectively. As night-time temperature rendered EVI less statistically insignificant and near-surface temperature range rendered specific humidity less statistically insignificant than the other temperature metric pairings, we selected EVI, specific humidity, near-surface temperature range, and night-time temperature as the inputs to our mosquito WNV positivity derivative RFR and BELR models.

Test-Train Split. For each prediction task, data from epidemiological weeks 22-40 of each year from 2007 - 2014 were used as training and data from 2015-2020 was used as testing data. We did not include 2021 as, at the time of development, we had not yet reached the 40th epidemiological week in 2021 and did not want to introduce an unknown effect into our models' predictions by providing it an unfinished season of data. Consequently, our test-train split was 57.14% to 42.86%.

Training the Models. We 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 — Table 4 details the possible settings for each feature. On running 100 iterations with three cross folds each, the RFRs with the parameters in Table 5 emerged as the best performing for each prediction task. Figure 11 displays the feature importance for each RFR model. 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 6 and 7 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 12, 13, 14, and 15 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 8 compares the overall RMSE of our RFR and BELR models to that of a similar study by Lee et al. (2016) 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. (2016). Given the high variability of the desired mosquito population characteristic output as seen in Table 2 and the extreme outliers evident at points such as week 29 in 2016 in Figures 12 and 14, our models' errors are comparatively low and demonstrate strong overall performance.

Discussion

In this study, we present a comparative evaluation of four machine learning models for two mosquito population and vector competence 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 accurately predict the derivatives of mosquito abundance and mosquito WNV positivity in our AOI. 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. Our work builds on recent predictive models, such as that of Koolhof et al. (2020), who created a predictive early warning forecast model for the transmission of Ross River Virus, a mosquito-borne disease in Australia, by time-lagging several environmental predictors of the disease. Their model does not use remote sensing data; instead, it uses monthly averages from measurements taken on the ground. In contrast, our model provides predictions on a weekly basis, which ultimately allows for a more precise prediction. 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. In the data science sector, additional machine-learning or deep-learning based regression models' performance in these tasks should be evaluated and further research should be directed into how to optimize the performance and hyperparameter tuning of our RFR and BELR models. In the public health sector, there remains room for further collaboration between data scientists and public health officials to turn our results into actionable metrics that align with and enhance the efficacy of current public health programs dealing with mosquito-borne diseases.

Acknowledgements

A big thank you to our NASA STEM Enhancement in the Earth Sciences (SEES) 2021 mentors Dr. Rusanne Low, Ms. Cassie Soeffing, Mr. Peder Nelson, Dr. Erika Podest, and Dr. Becky Boger!

The Team

Julianna E. Schneider

School: The Davidson Academy of Nevada

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

Julianna led the project team — designing and coordinating the various components of our research — and developed our machine learning models. She also contributed to general data cleaning, writing our research report, editing our research report, and conducting the research that powered our project.

Alexander W. Greco

School: Western Canada High School

Roles: Google Earth Engine and Remote Sensing Data Specialist, Data Management, Writer, Researcher

Alexander led the acquisition, analysis, and preprocessing of the ecological inputs obtained through Google Earth Engine's datasets. He also contributed to general data cleaning, writing our research report, and conducting the research that powered our project.

Jillian Chang

School: Great Neck South High School

Roles: Data Visualization Specialist, Editor, Writer, Researcher

Jillian led the graphing and associated analyses of our ecological inputs and mosquito outputs to uncover patterns within our data and aid visual comprehension of these patterns within our documentation. She also contributed to writing our research report, editing our research report, and conducting the research that powered our project.

Maria Molchanova

School: Thomas Jefferson High School for Science and Technology

Roles: Citizen Science Data Specialist, Data Management, Writer, Researcher

Maria led the acquisition, analysis, and preprocessing of the citizen science datasets we worked with and the Chicago mosquito dataset. She also contributed to general data cleaning, writing our research report, and conducting the research that powered our project.

Luke Y. Shao

School: West Windsor Plainsboro High School North

Roles: Documentation and Presentation Specialist, Data Visualization, Writer, Researcher

Luke led the documentation and presentation of our findings. He also contributed to data visualization tasks, writing our report, and conducting the research that powered our project.

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.

References

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

Table 1

P-values of Ecological Variables, including single and multi-inputs, after performing OLS Regression (mosquito abundance). An asterisk (*) denotes p value <0.05, indicating statistical significance.

Indicator(s) for Mosquito Abundance	P Value >|t|
Minimum Precipitation	0.122
Maximum Precipitation	0.356
Maximum Full-Day Land Surface Temperature	0.223
Maximum Full-Day Near-Surface Temperature	0.356
Minimum Full-Day Land Surface Temperature	0.223
Minimum Full-Day Near-Surface Temperature	0.737
Full-Day Land Surface Temperature	0.737
Full-Day Near-Surface Temperature	0.188
Total Precipitation	0.255
Average Specific Humidity	0.191
Average Enhanced Vegetation Index	0.657
Average Day Land Surface Temperature	0.892
Average Night Land Surface Temperature	0.348
Full-Day Land Surface Temperature Range	0.133
Full-Day Near-Surface Temperature Range	0.784
Difference between Near-Surface Temperature and Land Surface Temperature	0.002*
Minimum Precipitation	0.047*
Maximum Precipitation	0.000*
Minimum Precipitation	0.219
Maximum Precipitation	0.928
Total Precipitation	0.000*

Table 1

P-values of Ecological Variables, including single and multi-inputs, after performing OLS Regression (mosquito abundance). An asterisk (*) denotes p value <0.05, indicating statistical significance.

Indicator(s) for Mosquito Abundance	P Value >|t|
Maximum Full-Day Land Surface Temperature	0.191
Minimum Full-Day Land Surface Temperature	0.241
Maximum Full-Day Land Surface Temperature	0.686
Minimum Full-Day Land Surface Temperature	0.866
Full-Day Land Surface Temperature	0.679
Full-Day Land Surface Temperature	0.002*
Full-Day Near-Surface Temperature	0.005*
Full-Day Land Surface Temperature	0.172
Full-Day Near-Surface Temperature	0.438
Full-Day Land Surface Temperature	0.805
Full-Day Land Surface Temperature Range	0.004*
Difference between Near-Surface Temperature and Land Surface Temperature	0.203
Full-Day Land Surface Temperature Range	0.388
Full-Day Land Surface Temperature	0.438
Difference between Near-Surface Temperature and Land Surface Temperature	0.005*
Average Day Land Surface Temperature	0.352
Average Day Land Surface Temperature	0.773
Average Night Land Surface Temperature	0.030*

Table 1

P-values of Ecological Variables, including single and multi-inputs, after performing OLS Regression (mosquito abundance). An asterisk (*) denotes p value <0.05, indicating statistical significance.

Indicator(s) for Mosquito Abundance	P Value >|t|
Average Night Land Surface Temperature	0.016*
Average Night Land Surface Temperature	0.684
Full-Day Land Surface Temperature	0.062
Average Enhanced Vegetation Index	0.394
Average Enhanced Vegetation Index	0.023*
Full-Day Land Surface Temperature	0.000*
Average Enhanced Vegetation Index	0.032*
Full-Day Land Surface Temperature	0.001*
Total Precipitation	0.403
Average Specific Humidity	0.042*
Average Specific Humidity	0.117
Total Precipitation	0.374
Average Specific Humidity	0.022*
Average Enhanced Vegetation Index	0.152
Average Specific Humidity	0.988
Average Enhanced Vegetation Index	0.028*
Full-Day Land Surface Temperature	0.006*

Table 1

P-values of Ecological Variables, including single and multi-inputs, after performing OLS Regression (mosquito abundance). An asterisk (*) denotes p value <0.05, indicating statistical significance.

Indicator(s) for Mosquito Abundance	P Value >|t|
Average Specific Humidity	0.796
Average Enhanced Vegetation Index	0.040*
Full-Day Land Surface Temperature	0.007*
Total Precipitation	0.396

Table 2

Mean, Standard Deviation, and Range for Ecological Inputs and Mosquito Population Characteristics Outputs

Input	Mean ± Standard Deviation	Range
EVI	1.718630e-01 ± 4.1438e-2	0.027678 — 0.256264
Land Surface Temperature	2.953583e+02 ± 3.825502e0	283.137074 — 305.097237
Specific Humidity	1.210294e-02 ± 2.731e-3	0.0052 — 0.019949
Total Precipitation	2.793383e-01 ± 2.80093e-1	0.000042 — 1.353915
Near-Surface Temperature Range	1.430573e+01 ± 2.868203e0	7.779097 - 22.887282
Landsat Night Temperature	2.894261e+02 ± 4.599292e0	275.508599 - 298.45227
Derivative of Mosquito Abundance	-6.203008e-02 ± 1.0006173e1	-55.0 - 43.0
Derivative of Mosquito WNV Positivity	9.398496e-06 ± 5.202e-3	-0.0265 - 0.017

Table 3

Final pool of ecological variables considered in OLS regression testing.

No.	Input Type
1	Average Full-Day Near-Surface Temperature
2	Minimum Full-Day Near-Surface Temperature
3	Maximum Full-Day Near-Surface Temperature
4	Maximum Full-Day Land Surface Temperature
5	Minimum Full-Day Land Surface Temperature
6	Average Full-Day Land Surface Temperature
7	Average Day Land Surface Temperature
8	Average Night Land Surface Temperature
9	Total Precipitation
10	Average Enhanced Vegetation Index
11	Average Specific Humidity
12	Maximum Precipitation
13	Minimum Precipitation
14	Difference between Near-Surface Temperature and Land Surface Temperature
15	Full-Day Near-Surface Temperature Range
16	Full-Day Land Surface Temperature Range

Table 4

Possible values provided to the Randomized Search Cross Validation model selection tool when developing RFR models.

Hyperparameter	Tested Values
'bootstrap'	[True, False]
'max-depth'	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None]
'max_features'	['auto', 'sqrt']
'min_samples_split'	[2, 5, 10]
'min_samples_leaf'	[1, 2, 4]
'n_estimators'	[200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]

Table 5

Optimal hyperparameters for each RFR model.

Hyperparameter	RFR for predicting mosquito abundance derivative	RFR for predicting mosquito WNV positivity derivative
'bootstrap'	True	True
'max-depth'	50	20
'max_features'	'sqrt'	'sqrt'
'min_samples_split'	4	4
'min_samples_leaf'	10	10
'n_estimators'	1200	2000

Table 6

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

Model	Overall MAE	Overall RMSE	Max RMSE	Min RMSE
RFR	5.311205	7.476696	3.699710	0.000921
BELR	4.454769	6.696208	3.318949	0.002721

Table 7

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

Model	Overall MAE	Overall RMSE	Max RMSE	Min RMSE
RFR	0.004442	0.006522	0.002433	1.378654e-06
BELR	0.004279	0.006451	0.002520	3.539568e-06

Table 8

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

Model	Overall RMSE	Range of Desired Output
RFR for Mosquito Abundance Derivative	7.476696	98
BELR for Mosquito Abundance Derivative	6.696209	98
RFR for Mosquito WNV Positivity	0.006522	0.0435
BELR for Mosquito WNV Positivity	0.006451	0.0435
MLR for Mosquito Abundance	17.53	78
ANN for Mosquito Abundance	14.38	78

Figure 1

Map of Enhanced Vegetation Index for Chicago and Surrounding Area from August 2018.

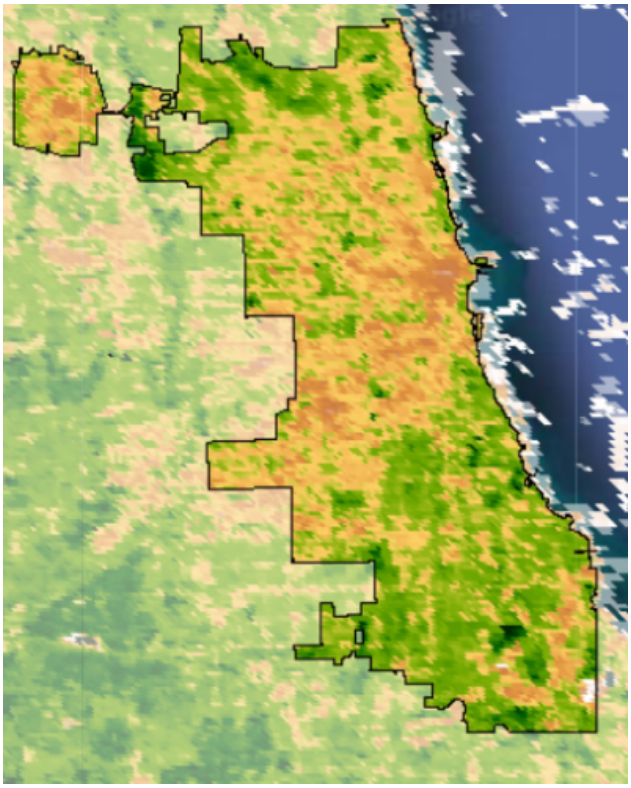

Figure 2

GLOBE MHM points and City of Chicago points in relation to Chicago boundaries.

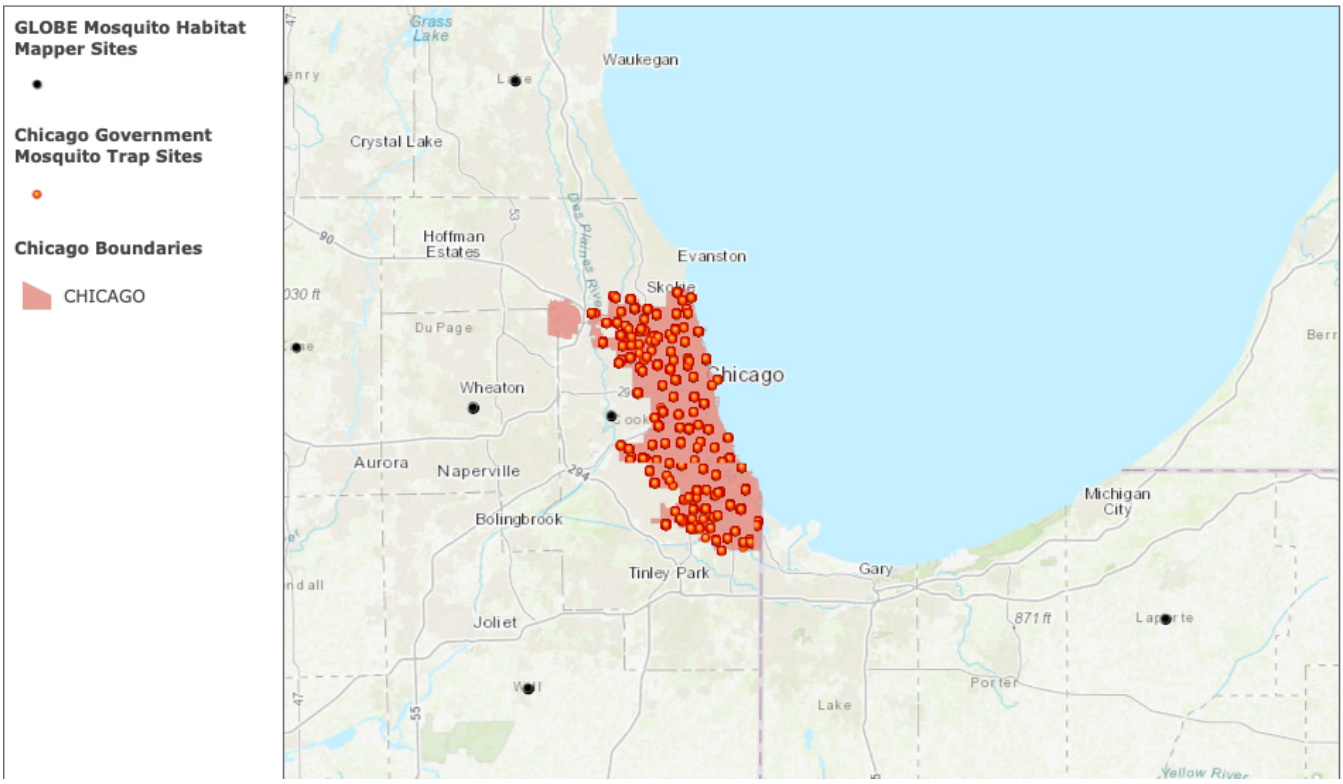

Figure 3

Mosquito Abundance in Chicago Over Time.

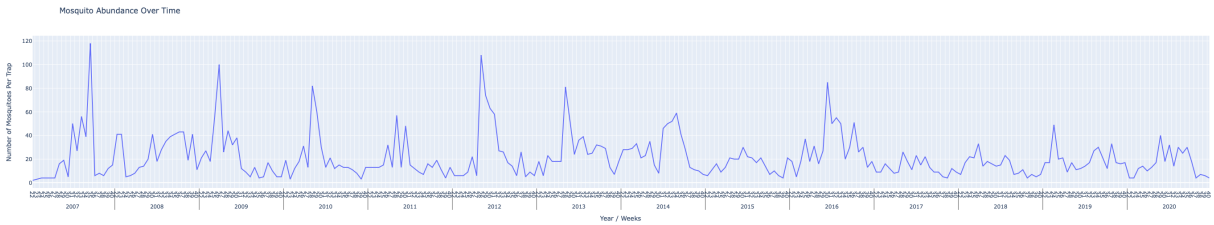

Figure 4

Derivatives For Mosquito Abundance in Chicago Over Time .

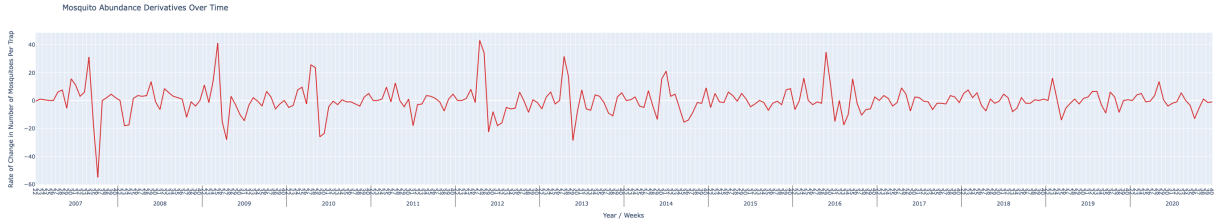

Figure 5

Percent of Mosquitoes Tested for West Nile Virus Per Trap in Chicago Over Time.

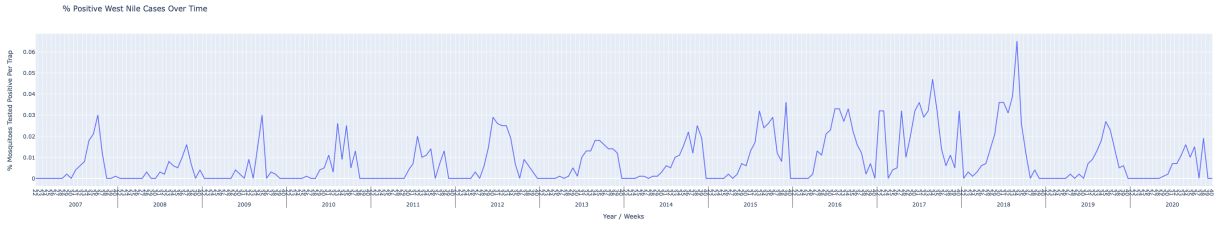

Figure 6

Derivatives for Percent of Mosquitoes Tested for West Nile Virus Per Trap in Chicago Over Time.

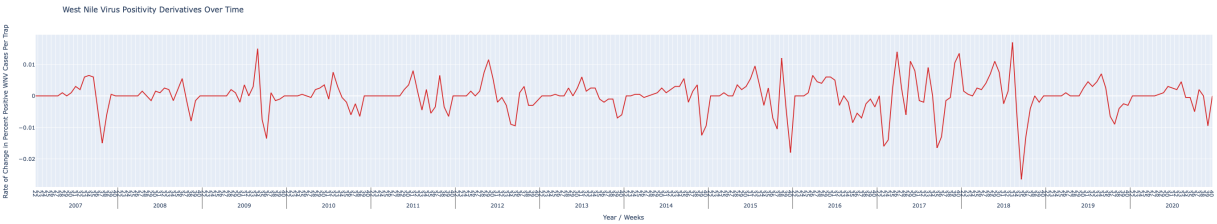

Figure 7

Left: Time series of mosquito abundance in Chicago for 2009. Middle: Missing data points filled with the Most Frequent method. Right: Missing data points filled with the Median method.

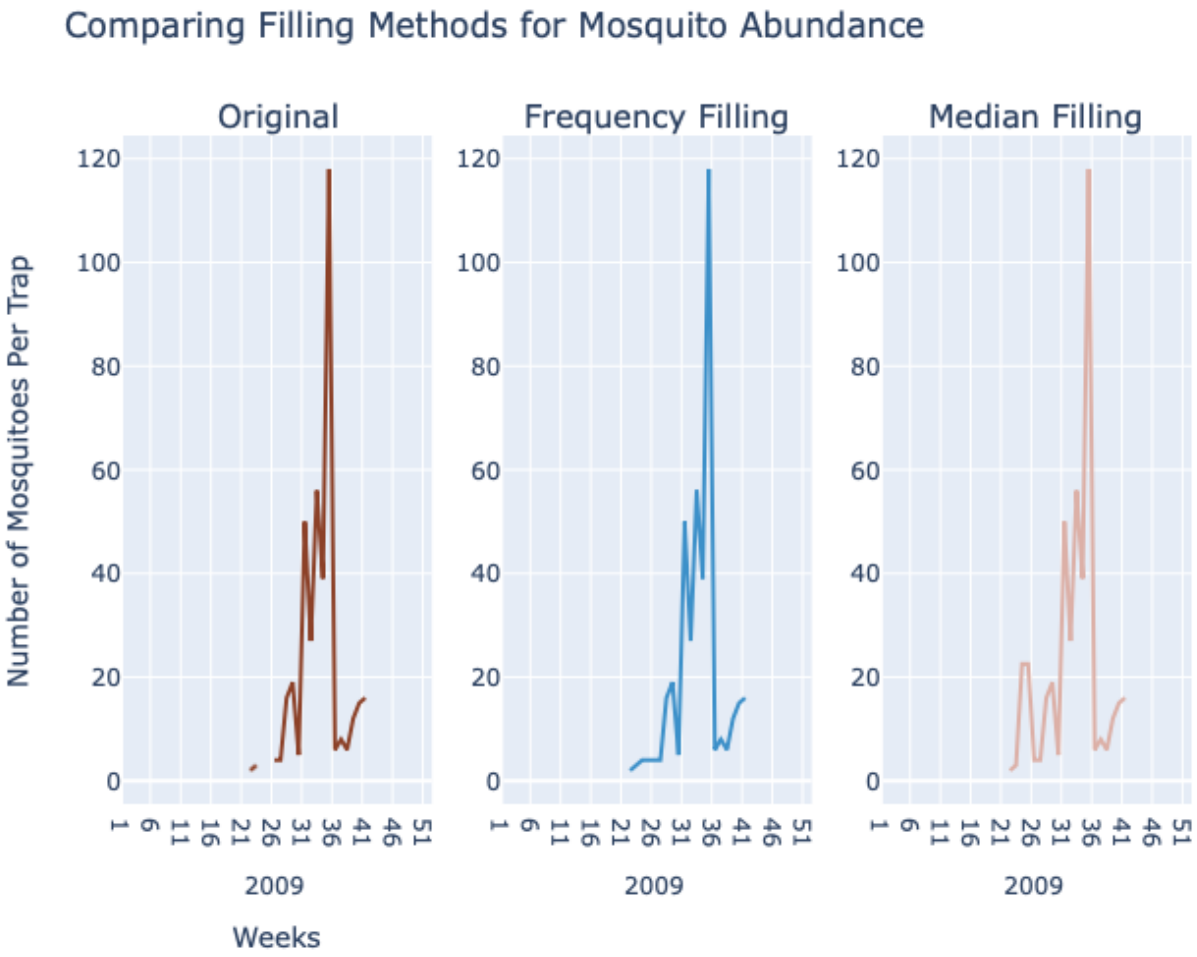

Figure 8

Left: Time series of West Nile Virus positivity rates in Chicago for 2012. Middle: Missing data points filled with the Most Frequent method. Right: Missing data points filled with the Median method.

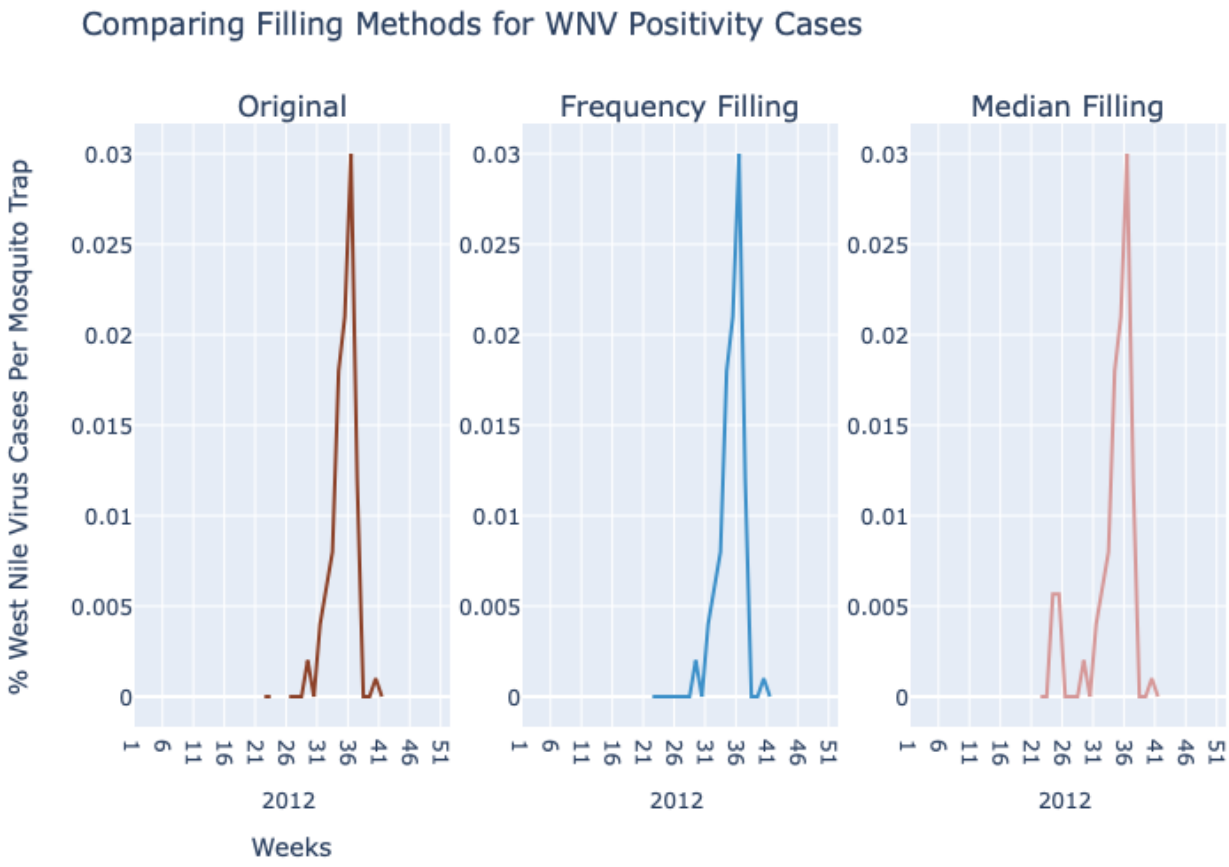

Figure 9

Left: time series of Land Surface Temperature, Specific Humidity, and EVI for 2017 in the City of Chicago. Right: adjustment of Land Surface Temperature, Specific Humidity, and EVI based on a forward time shift of three weeks.

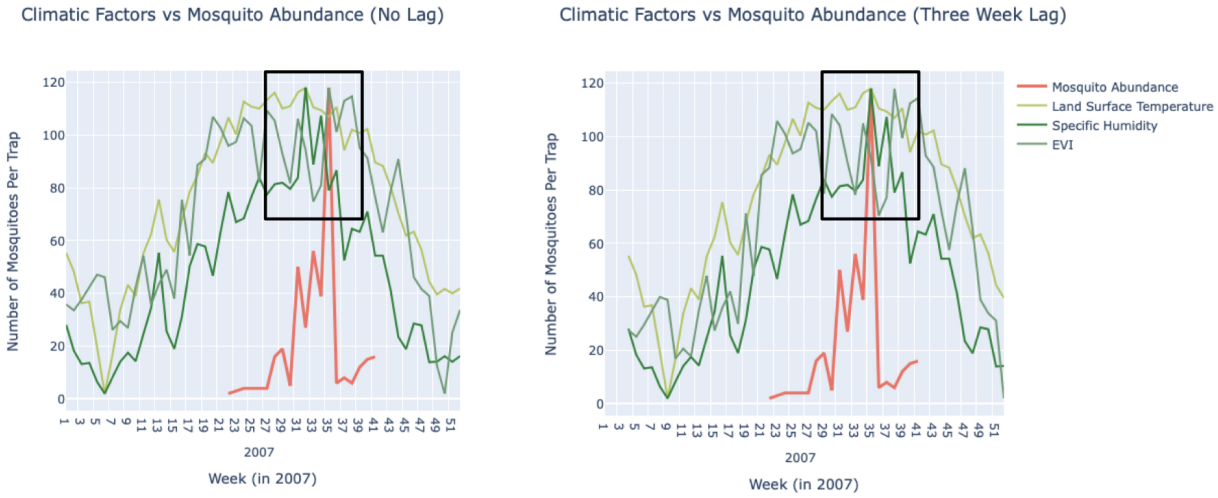

Figure 10

Left: time series of total precipitation and West Nile positivity rates for 2015 in the City of Chicago. Right: adjustment of total precipitation based on a forward time shift of three weeks.

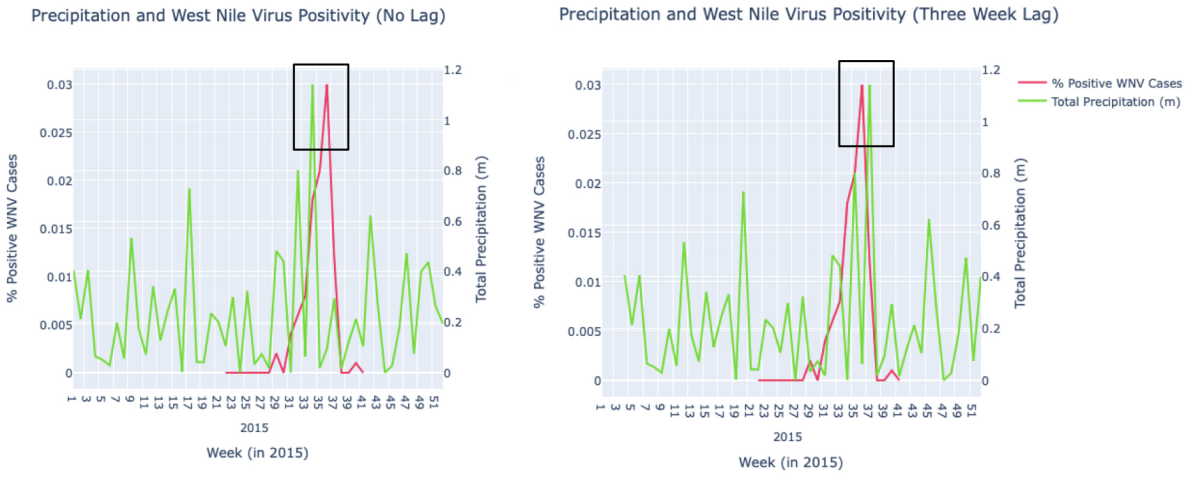

Figure 11

Top: Feature importance for the RFR predicting the derivative of mosquito abundance at each timestep. Bottom: Feature importance for the RFR predicting the derivative of mosquito WNV positivity at each timestep.

Figure 12

Top: Time series of observed values, predicted values, and root mean square error for Random Forest Regressor's performance on predicting mosquito abundance. Bottom: Observed values and predicted values made less transparent to highlight root mean square error.

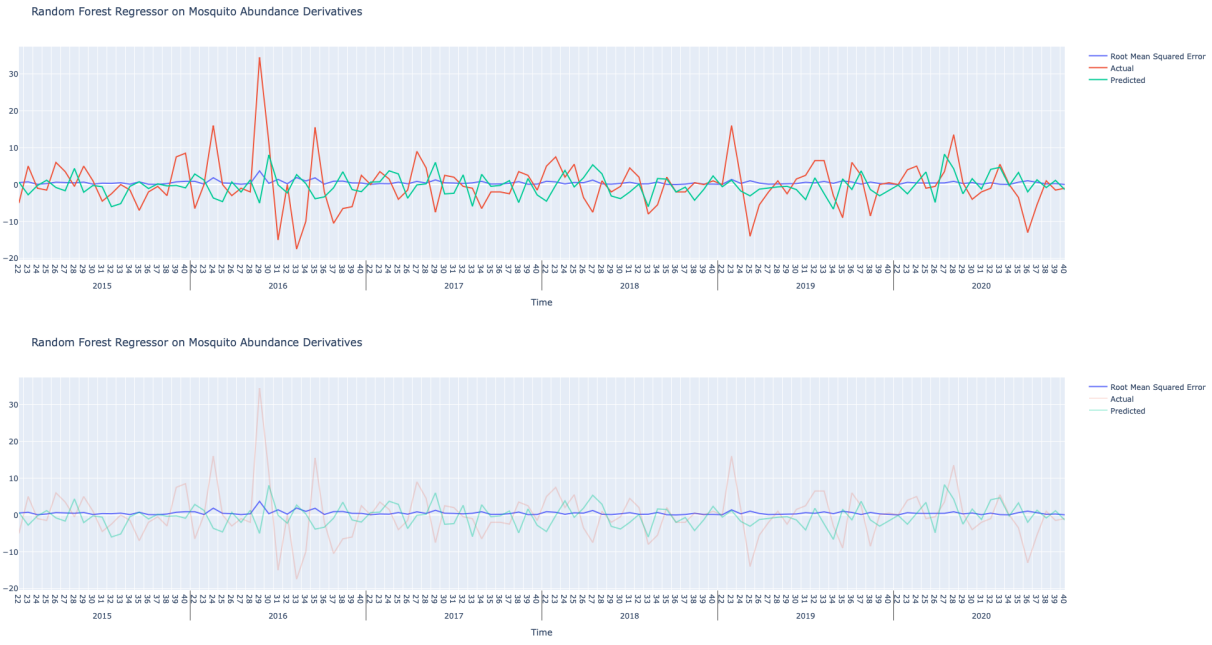

Figure 13

Top: Time series of observed values, predicted values, and root mean square error for Random Forest Regressor's performance on predicting mosquito West Nile Virus positivity rate. Bottom: Observed values and predicted values made less transparent to highlight root mean square error.

Figure 14

Top: Time series of observed values, predicted values, and root mean square error for Backward Elimination Linear Regressor's performance on predicting mosquito abundance. Bottom: Observed values and predicted values made less transparent to highlight root mean square error.

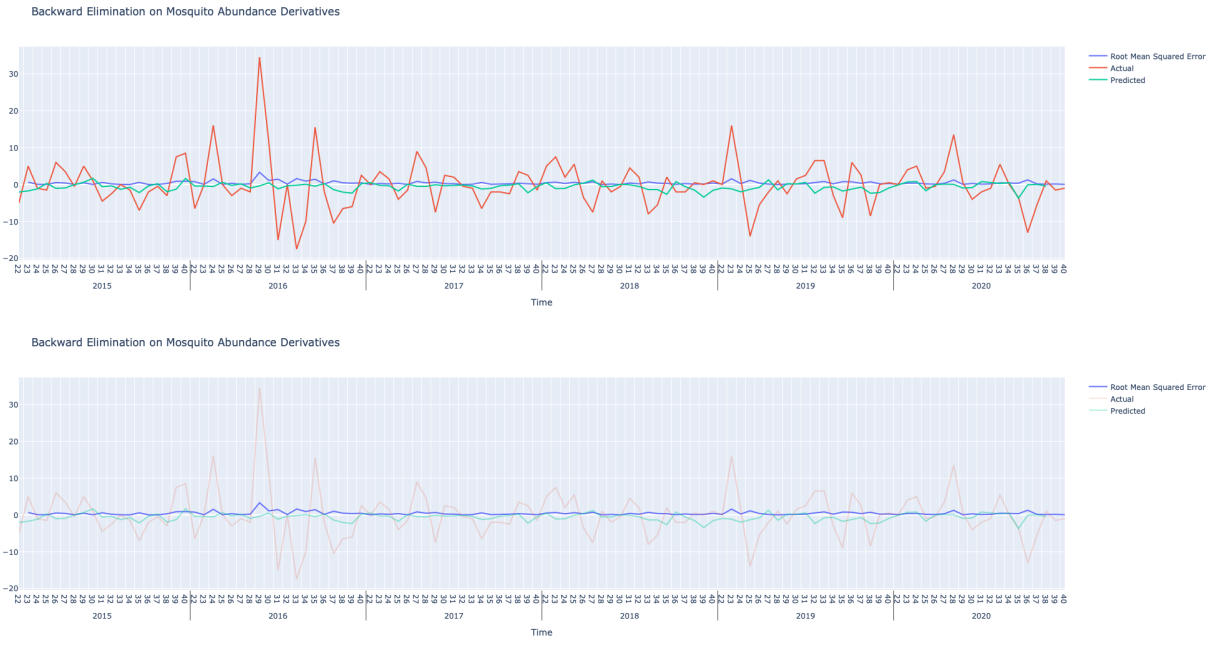

Figure 15

Top: Time series of observed values, predicted values, and root mean square error for Backward Elimination Linear Reggresor's performance on predicting mosquito West Nile Virus positivity rate. Bottom: Observed values and predicted values made less transparent to highlight root mean square error.