A predictive model for West Nile Virus surveillance by detecting inland water eutrophication using Sentinel-2 imagery

Authors: Sarah Blackett[a], Ishaan Verma[b], Salil Khare[c], Benjamin Kwait-Gonchar[d], and Daisy Li[e]

[a]St. Petersburg High School, St. Petersburg, FL; [b]Bridgewater Raritan Regional High School, Bridgewater, NJ; [c]Irvington High School, Fremont, CA; [d]Brooklyn Technical High School, Brooklyn, NY; [e]Alexander W. Dreyfoos School of the Arts, West Palm Beach, FL;

Mentors: Russanne Low, Ph.D., Peder Vernon Nelson, Cassie Soeffing, Andrew Clark, Erika Podest, Ph.D., and Ria Jain

Abstract:

Monitoring mosquito abundance and its contributing indices are crucial for controlling West Nile Virus (WNV), the most prevalent vector-borne disease in the contiguous United States. As mosquito-borne diseases like West Nile Virus (WNV) are primarily transmitted through infected mosquitoes, it is a good indicator of mosquito prevalence in an area. Empirical data from our field research showed a direct correlation between fertilizer concentration and mosquito larvae population in a small trap experiment. However, detection of the presence of fertilizer in water bodies across counties and states requires site sampling, which is very time-consuming and expensive to perform. This research is based on the well-documented correlation between significant fertilizer presence in a water system and algae blooms. Detection of algae in inland water could provide an early warning signal in controlling vector-borne diseases such as the West Nile Virus (WNV). Remote sensing and satellite imagery provide a cost-effective alternative for monitoring inland water bodies such as rivers, lakes, water reservoirs, ponds, etc. We developed a supervised machine learning model using the Naïve Bayes algorithm to predict WNV breakout by detecting algae from Sentinel-2 MSI images. The model was trained using high spatial resolution products (20m) from Sentinel-2 satellites over Sacramento, California. Methods applied for algae bloom extraction from Sentinel-2 MSI images, with a high spatial resolution, are based on an estimation of Chlorophyll-a (Chla) and the use of the Normalized Difference Chlorophyll Index (NDCI), which is widely used for ocean color data. To suppress the chlorophyll from vegetation in a satellite image, a combination of NDCI and Normalized Difference Water Index (NDWI) was used to measure algae presence in water bodies. A time series dataset was developed using Sentinel-2 images from 2017-2021 for algae bloom information. The training dataset was further enriched with feature sets such as water%, vegetation%, algae observed/reported in the public domain, and the California Department of Public Health's West Nile Virus 2006-Present dataset. The accuracy of the ML predictive model ranges from 0.7 to 0.95, depending on the algorithm and the length of time series used for the training of the model. Our research also validated the time lag between algae bloom and actual detection of the WNV virus reported through public health departments. With additional training data, this model can be extended to predict potential WNV outbreaks for any given county using satellite images.

(Keywords: Machine Learning, Mosquito Abundance, Algae, Eutrophication, Fertilizer Runoff, Remote Sensing, West Nile Virus, Sentinel-2, SVM, Support Vector Machine, Naïve Bayes, Prediction Model)

1. Introduction

The mosquito vector-borne West Nile Virus remains a threat to public health globally. Even in the United States, where many vector-borne diseases have been effectively eliminated, WNV was responsible for the deaths of 66 individuals in the United States, during 2020 alone (CDC). 1 in 150 infected people will develop critical illnesses because of this pathogen (CDC). Because there is currently no vaccine available for WNV in humans, prediction remains one of public health professionals' most potent tools in preventing the spread of the virus. Mosquito populations and density have been shown to be correlated with WNV case rates (Mori et al. 2018) indicating the potential viability of using other variables linked with mosquito density as a means of predicting WNV case rates. A variety of predictive variables including temperature (Mori et al. 2018) as well as rice cultivation (Kovach et al. 2018) with the latter's viability stemming directly from its correlation with mosquito populations.

The excessive use of fertilizer, especially near reservoirs, has also been shown to increase mosquito density in the surrounding regions (Reuben et al., 2008). But beyond simply increasing the number of mosquitos, the investigation has established that aqueous nutrients commonly found in fertilizer are linked to an increased arboviral content in mosquitos found in the area (Yee et al., 2017). However, the impact of algae, nor fertilizer, on mosquitos has not been studied on a scale larger than a small region, let alone a possibly widely applicable model grounded in remote sensing data. To fill this gap, this research aims to detect a practically usable link between algae coverage in inland waters, a proxy for fertilizer runoff, and WNV case rates in the surrounding area. Due to the known effects of fertilizer on mosquito populations, we hypothesized a strong positive relationship. However, due to the larvicidal properties of some algae species, we were aware that a noticeable negative correlation was also a possibility.

Detection of algae in inland water could provide an early warning signal in controlling vector-borne diseases such as the West Nile Virus (WNV). Remote sensing and satellite imagery provide a cost-effective alternative for monitoring inland water bodies such as rivers, lakes, water reservoirs, ponds, etc. We developed a supervised machine learning model using the Naïve Bayes algorithm to predict WNV breakout by detecting algae from Sentinel-2 MSI images. The model was trained using high spatial resolution products (20m) from Sentinel-2 satellites over Sacramento, California.

Machine learning algorithms are widely used for monitoring soil, water quality, crop classification, and algae blooming in ocean water using satellite images. [SVM & Crop]. 'Deriving Water Quality Parameters Using Sentinel-2 Imagery: A Case Study in the Sado Estuary, Portugal' is a good reference for measuring Chlorophyll-a to indicate the blooming of algae using machine learning.

Empirical data from our experimental field research showed a direct correlation between fertilizer concentration and mosquito larvae population in a small trap experiment.

In view of this, we developed a supervised machine learning model using the Naïve Bayes algorithm to predict WNV breakout by detecting algae from Sentinel-2 MSI images. The model was trained using high spatial resolution products (20m) from Sentinel-2 satellites over Sacramento, California. Methods applied for algal bloom extraction from Sentinel-2 MSI images, with a high spatial resolution, are based on an estimation of Chlorophyll-a (Chla) and the use of the Normalized Difference

Chlorophyll Index (NDCI), which is widely used for ocean color data. The effectiveness of the model was tested with 2 different sentinel tiles (10SFH and 11SKA) for data from 2020 and 2021. The accuracy of the ML predictive model ranges from 0.7 to 0.95, depending on the algorithm and the length of time series used for the training of the model.

2. Study Area and Data

2.1 Study Areas

2.1.1. Experimental Research

For our earth science field research to study the effect of fertilizer on mosquito breeding, St. Petersburg, Florida was chosen as our study area for mosquito traps due to its proximity to one of our researchers and its known mosquito presence. The experimental component involved four mosquito traps in close proximity to one another in the St. Petersburg, Florida area. Each trap consisted of an open five-gallon bucket filled with 10 liters of tap water and a varying amount of "Miracle-Gro All Purpose Plant Food" powder. The quantity of fertilizer served as our four experimental treatments were 0 tsp (0 ml), 1 tsp (4.93 ml), 2 tsp (9.86 ml), and 3 tsp (14.79 ml) per 10 L of water or a concentration of 0 ppt, 0.493 ppt, 0.986 ppt, and 1.479 ppt. The experiment took place over the course of 3 weeks during the time interval June 28 - July 19, 2022. Besides the varying treatments, all other outside factors including trap design, water quality, environmental conditions, and weather were effectively identical between traps and had a minimal effect on variation in the data as a result.

2.1.2. Remote Sensing

To expand the research to study algae blooming and its effect on mosquito breeding (more specifically West Nile Viruses), California was chosen as our area of interest (AOI) for research using remote sensing satellite images. In the recent past, California's Fresno, Sacramento, Yolo, Alameda, and San Joaquin areas have reported both harmful algae blooming and West Nile Virus cases. California offers a broad range of environments- temperate coniferous forests, deserts, mountains, wetlands, lakes, rivers, grasslands, woodlands, urban areas, chaparrals, and agricultural areas. The temperature in California ranges from 10°F to 97°F. California has many open access public data that provides data on inland algal blooms from 2016 to 2022 and WNV data from 2006 to 2022.

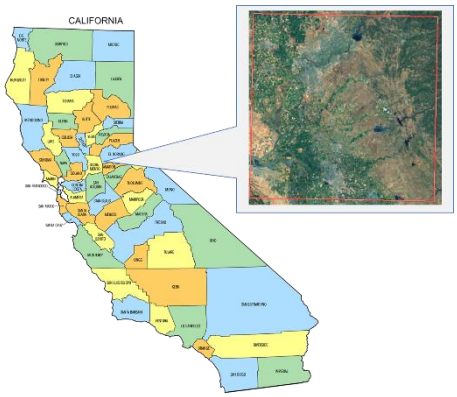

Figure 1. 10 SFH tile for the Sacramento county courtesy of ESA

For Satellite images, the region around Sacramento, CA with a latitude range from 37.96N to 38.59N and a longitude range from -121.76W to -121.19W was selected as the primary region due to the diversity of landscape with inland water bodies, as well as the availability of West Nile Virus data from 2006 onwards. Additionally, freshwater harmful algae bloom incident response data was gathered for the counties within the specified AOI around Sacramento, CA. At a later stage in this research, Fresno, CA was added to expand the dataset to improve the reliability of the ML algorithm for WNV prediction. For this region, Latitude ranges

from 36.04N to 37N and Longitude ranges from -10.37W to -119.14W.

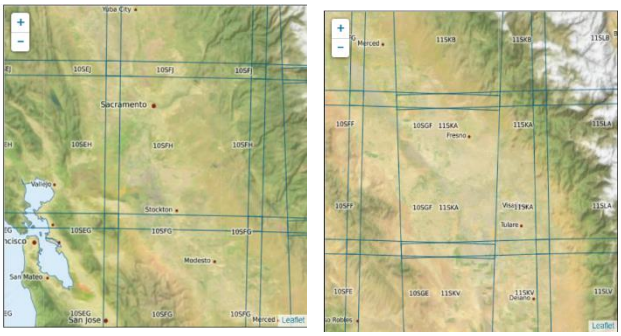

Figure 2.. Sentinel-2 tile images via ESA

During the initial stage, images from Landsat, Sentinel-3, and Sentinel 2 were evaluated for easy retrieval and classification of water bodies. While both Landsat and Sentinel-2 provide good spatial resolution (10m-60m) for water classification algorithm and detection of inland water, we finally settled on accessing Sentinel-2 image to take advantage of visible and Near Infrared (VNIR) and availability of level 2A products. Based on the earlier research, Reflectance-classification methods can be sufficient for mapping algal blooms with the spectral bands located in visible and NIR wavelength regions [1]. Spectral bands in red-edge and NIR regions show much better results for the discrimination of algal blooms in inland waters than visible bands.

Level 2A provides tiles of 100km x 100km images for each band. While no single tile covered the entire area of interest, tiles 10SFH and 11SKA were selected for Sacramento and Fresno regions respectively,

2.2. Data

2.2.1 Experimental Data

Larvae was tallied manually and weekly across the three-week experimentation period using a miniature clip on microscope attachment. The contents in each five-gallon bucket were purged and refilled after each week. Although the magnifying power fell short of identifying the specific species of larvae, the broad genus, either culex or aedes, was specified. Overall, culex was more abundant in number than aedes, as found in the trap experiment. This can likely be attributed to the regional variability of mosquito species in St. Petersburg, Florida. Mosquitoes had a preference to lay their larvae in fertilized water but only up to a certain concentration, peaking in the mid-ranges of 2 tsp (9.86 ml).

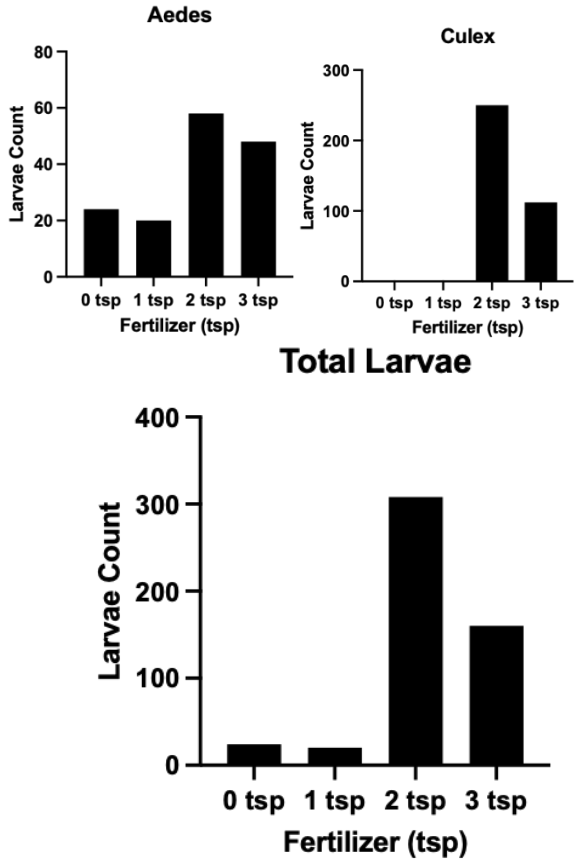

Figure 3. Aedes, Culex, and Total larvae counts for fertilizer treatments

2.2.2 Satellite Imagery Data

Retrieval of satellite data from California was done by utilizing a GeoJSON file to define a specific California county's area. The file was then used as an input to the Sentinel-2 product database along with restrictions on time range and cloud cover (kept to between 0% and 5% for training). Products were additionally filtered to be Level-2A and with a low unidentified pixel percentage. Products were attempted to download using direct access (Copernicus Open Access Hub) as well as python script using API for Sentinelsat. Due to the flexibility to programmatically download images from long-term archival (i.e. old data), python API was preferred.

A total of 27 Sentinel-2 MSI products from 2009 to 2022 were downloaded for spatial resolution of 10M, 20m, and 60m across California (Sacramento, Orange County, and Fresno). All Sentinel-2 MSI products and images were validated using SNAP software provided by the European Space Agency. Finally, 15 Sentinel-2 MSI products were processed for generating a Normalized Difference Water Index (NDWI) and NDCI.

3. Methodology

3.1 Experimental Research

For the experiment, we split the experimental results into two groups for analysis: One for Culex mosquitoes and one for Aedes mosquitoes. The weekly totals for each treatment were then summed to create an overall observed larvae total for each treatment for both the Culex and Aedes group. We then ran a Chi-Squared Goodness of Fit test on each group's four treatments and their respective larvae counts.

Throughout each week, algae grew in all the buckets. Though the algae were not quantitatively measured, the water with more fertilizer was consistently greener than the water with less fertilizer. Larvae did not appear in the buckets until after the algae had a few days to grow.

Figure 4. Weekly Progress

Figure 5. Larvae growth

At the end of each week, the larvae in each bucket were counted and classified by genus. The larvae were extracted from the bucket using a smaller container, from which they were then placed onto a white surface coated with isopropyl alcohol via pipette. This killed the larvae, which prevented them from moving so they could be easily photographed, identified, and counted. After all larvae were counted and identified, each bucket was reset; they were dumped and refilled with new water and fertilizer. This way, each week was a different trial, and we could record the total number of larvae from each bucket accumulated from the three different weeks

Only two genera of mosquitoes, Aedes and Culex, were identified from the traps after all three weeks of trials, and species-specific identification was no possible with the level of magnification we could achieve.

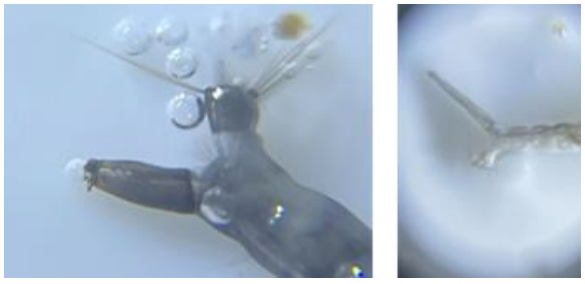

Figure 6. Genera of mosquitoes

In order to conduct a proper Chi-Squared Test, our data must be derived from a random sample, consist of only mutually exclusive categorical variables, and result in expected counts each greater than five. Nothing to our knowledge is unique enough about St. Petersburg Florida mosquitos that would make them unrepresentative of the national mosquito population as a whole. Furthermore, while our categorical variables are represented by numerical fertilizer concentrations, these quantitative values are not being used for regressive or predictive analysis, and the Chi-Squared GOF test is only analyzing differences between these 4 groups. Finally, each of our expected counts is well over five for both our Culex and Aedes samples.

3.2 Remote Sensing and Algae Detection

3.2.1 Sentinel-2 Product Pre-processing

Since we downloaded Sentinel-2 MSI Level 2A products, it's already been processed for atmospheric corrections by PDGS using the Sen2Cor processor [2]. Level 2A also includes a scene classification map (SCM) and quality indicators (QI) that allow easy segmentation to identify cloud/snow coverage, vegetation, water, and soil. This study utilized this algorithm to obtain data on vegetation and water in individual products. The algorithm is based on a series of threshold tests that use input TOA reflectance as input from the Sentinel-2 spectral bands. The products provided by Sentinel-2's satellites cover an area of 100km x 100km. This large size leads to the products covering multiple counties. However, data from the globe or on land observations are available by county within a state. So, it's necessary to slice satellite images to map with county boundary before it can be used for further analysis.

This resulted in the need to splice the test images into 5 individual smaller columns (labeled as zones in the dataset) to be analyzed and processed separately. Doing so provides the algorithm the ability to better establish a correlation between a county's water body eutrophication and WNV cases. County-level segmentation created 5x data from a single satellite image for training the prediction model.

Using the latitude and longitude data, harmful algae observation and WNV dataset were manually labeled for zones. Further occurrence of algae observation was converted as a numerical value for any given year and month for a specific zone. WNV data for the same period were manually labeled as "yes" or "no" for a period of study (2020 and 2021) and all counties are covered within each tile of the sentinel-2 MSI image.

3.2.2 Inland Water and Algae Classification

Reflectance-classification methods can be sufficient for mapping algal blooms with the spectral bands located in visible and NIR wavelength regions [3]. The Sentinel-2 products additionally come in three resolutions: 10m, 20m, and 60m. Each resolution offers a different set of 13 bands, with 10m offering the least number of bands

and 20m and 60m offering the same number of bands. In this study, we aimed to observe algae presence over inland water bodies. During this research, we focused on extracting Chlorophyll-a (Chla) information as an indicator of algae presence.

Normalized Difference Chlorophyll Index (NDCI) was proposed as a novel index in this research [4] due to its sensitivity to CHL-a in turbid water, a common occurrence in inland water.

The NDCI results from the following equation:

$$NDWI = \frac{band\ 2 - band\ 11}{band\ 2 + band\ 11}$$

A NumPy array was created to plot using rasterio plot function. Based on the plot, it was observed that chlorophyll reflectance from vegetation was making it difficult to distinguish algae indicators in the inland water. To overcome this, NDWI (Normalized Difference Water Index) algorithm was used to mask out vegetation from the image. NDWI indicated a change in liquid water content and it's less sensitive to atmospheric effects.

In order to best observe algae present in the water bodies, an algorithm was developed further to combine both NDCI and NDWI to generate a heatmap view and index: Relative Normalized Difference Chlorophyll on Water Index (R-NDCWI).

This index allows for the direct observation of chlorophyll content over inland water bodies. By using the NDWI as a mask to mask out non-water bodies out of the NDCI, the R-NDCWI provides a reliable index for the observation of chlorophyll on water bodies. A higher presence of chlorophyll in these water bodies would typically indicate eutrophication of these water bodies.

The raster was further sliced into 5 individual slices of 50,000 x 10,000 pixels each. Since the data was too large to be analyzed for each pixel, each image was scaled down (0.1) using the "resampling" function available with Rasterio. A threshold was determined through trial and error to suppress noise (i.e. replacing all pixel values below a threshold value by the threshold itself). A heatmap was generated to view the output before the threshold was finalized. Based on the final threshold value, R-NDCWI was generated for each slice of the raster. A scaled factor of 1000 was used to normalize the data further. Final r-NDCWI was manually mapped to the zone identified during the pre-processing step.

Using Sentinel-2 data products from 2020 to 2021, a time series view of NDCI, NDWI, and R-NDCWI was developed for both Sacramento and Fresno areas as a training dataset. The data was saved to a CSV file for use as input to train the ML prediction model.

3.2.3 Feature Enrichment

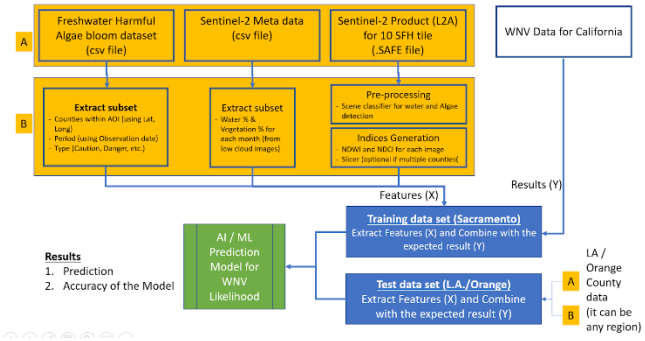

Figure 7. Data model for WNV prediction based on Sentinel-2 products and additional information

The indices generated for each image were further enriched with additional features available from both satellite data as well as each observation data. Since various satellite images for different locations may have

different landscapes (vegetation, water bodies, etc.), it was decided to add additional dimensions to make the prediction model multidimensional. A python utility using geopandas was developed to extract water% and vegetation% from the scene classification details available in the metadata for each sentinel-2 MSI image. In addition GLOBE landcover data was also extracted.

Earth Inland Observation data, reported/observed blooming of harmful algae for the same period (2021-22), was used to further enhance the dataset(X-train). WNV data available for California was further sliced to map with the county zone and period (2020-21) as the Y-train dataset.

Chlorophyll, GLOBE landcover), we decided to use a Naïve Bayes algorithm as a machine learning classifier for supervised learning. Another reason for selecting the Naïve Bayes machine learning algorithm was the availability of a limited dataset for training. Naïve Bayes is known to perform better even with small training datasets.

3.3.1 Gaussian Naïve Bayes Classifier

The Naïve Bayes algorithm operates on a probabilistic model that decides whether a piece of data belongs to a specific class based on the probability of it belonging to each specific class. The model in the study utilized a feature matrix of the county, water percentage, vegetation percentage, landcover data from GLOBE, algae presence, R-NDCWI, and NDWI from the CSV file

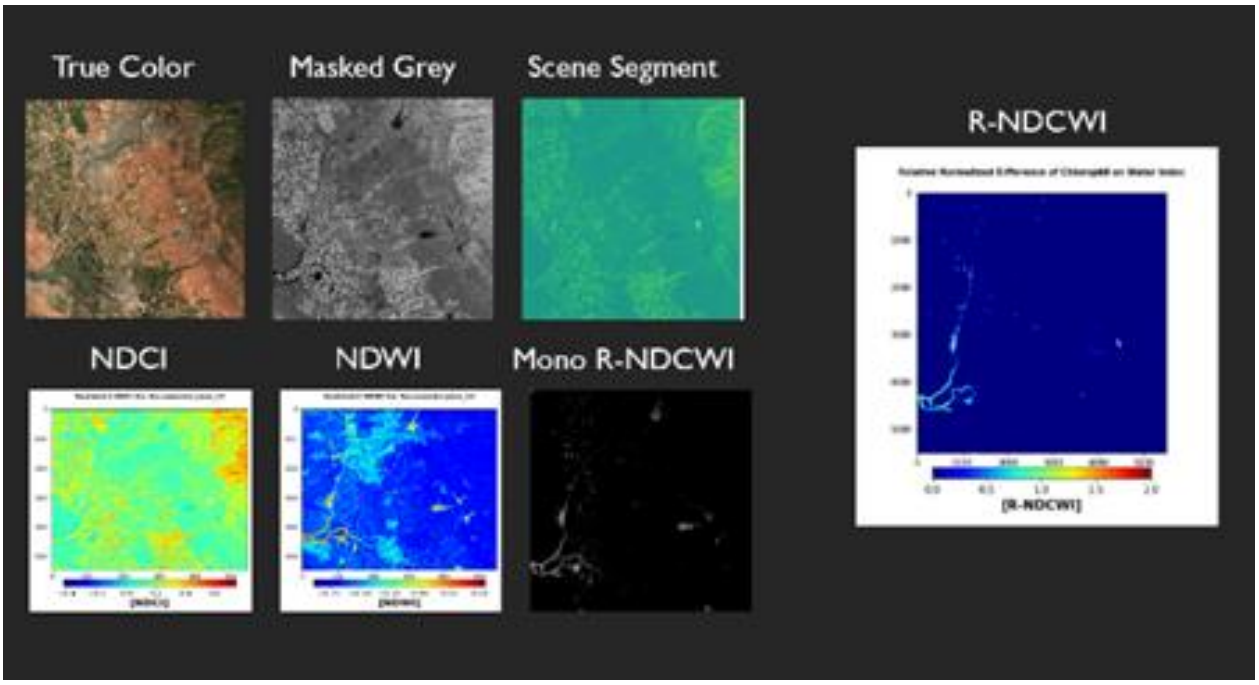

Figure 8 NDCI, NDWI and R-NDCWI Index generation and scene classifiation

3.3. Machine Learning Predictive Model

Since the dataset has features which were independent of each other (Vegetation%, Algae Observed, Water%, created during feature enrichment. The response vector is the WNV case presence also taken from the feature enrichment CSV file as output or prediction. The predictive model is based on training data retrieved

through the Sentinel-2 products for the Sacramento and Fresno regions.

Since feature values are expected to follow gaussian distribution for a given landscape/area of interest and the likelihood of West Nile Virus near water bodies is high, a Gaussian Naïve Bayes classifier seems to be the right model for our objective. Also, this data was used in a Gaussian Naïve Bayes model with a linear classifier and was tested using Sentinel-2 products for the Los Angeles and Orange County regions.

In addition to exploring a Gaussian Naïve Bayes algorithm, a Support Vector Machine (SVM) and Decision tree predictive model were also explored.

3.3.2 Support Vector Machine Predictive Model

SVM is a supervised learning method based on statistical learning theory and structural risk minimization principle. It has unique advantages in handling small sample data, solving nonlinear problems, and identifying high dimensional patterns.

Utilizing an SVM algorithm, the predictive model took the same input of the county, water percentage, vegetation percentage, landcover data from GLOBE, algae presence, R-NDCWI, and NDWI as a training input and WNV cases as the training output. An SVM model operates by creating a plane with the largest margins between two separate sets of data which are labeled for respective classes. For example, the classes that would be used in our model would be WNV presence or no WNV presence. What is unique about this model is that the space in which the plane is drawn and the data is plotted can be interpreted non-linearly by the algorithm to better separate the classes through a plane.

3.3.3 Decision Tree Predictive Model

A Decision Tree model operates on a set number of rules to ascertain whether a piece of data belongs to a specific class. These strict rules are operated in a "tree" fashion that processes the data through various conditionals to predict the class the data belongs to. The study utilized this model by using the same training input of county, water percentage, vegetation percentage, landcover data from GLOBE, algae presence, R-NDCWI, and NDWI and training output of WNV cases of the county.

4. Results

4.1 Experimental Research Results

4.1.1 Statistical Analysis and Results

The resulting P-Values for the Culex and Aedes mosquito count distributions were $6.068412355065628e^{-101}$ for Culex mosquito count and $5.414747992821513e^{-06}$ for Aedes mosquito count, both less than our alpha. These results indicated a statistically significant difference in the proportion of the total mosquito larvae population in each of the four fertilizer concentrations.

A Chi-Squared test cannot be used to indicate a causal relationship, or that the different larvae counts were a direct result of the varying fertilizer concentrations. However, the significance of our results indicates that there may be some relationship, and furthermore, that that relationship may be useful in predicting the number of mosquitos. Due to the limited time frame of our experiment we were unable to gather enough distinct samples to conduct an ANOVA test to experimentally analyze this relationship. But, we were able to assess the strength of this correlation with remote sensing and public health data, both input to a Naïve Bayes Machine learning predictive model.

4.2 ML Predictive Analysis in Python and Results

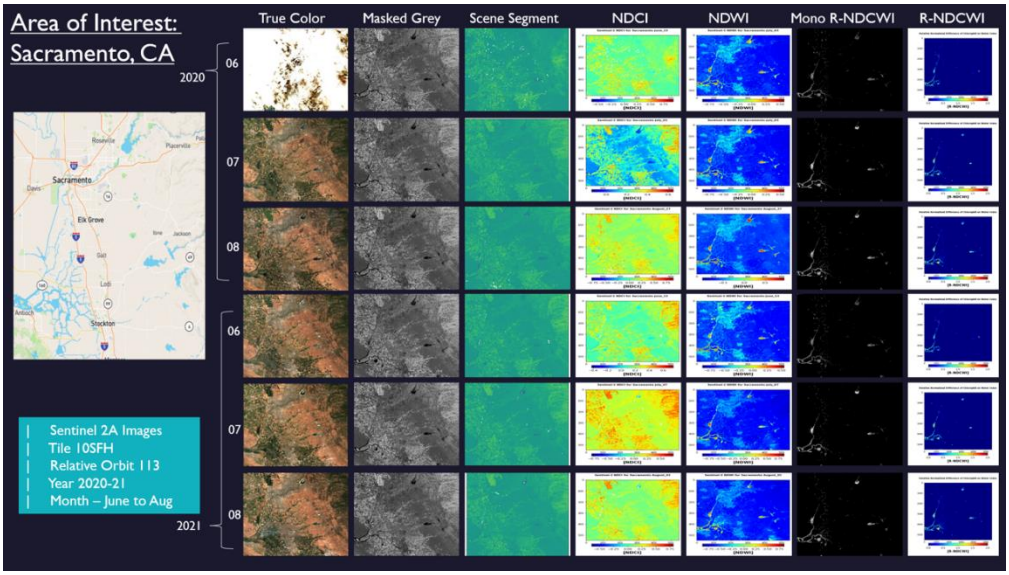

Figure 9. Sentinel-2 product of Sacramento area of interest region through preprocessing step

Different satellite images were used for the model with Water % ranging from 0.5% to 10% and Vegetation percentage ranging from 31% to 43% due to season variance during the months of June to August in 2020 and 2021. The table below shows the average value for results extracted from the satellite image using machine learning algorithms in python.

Table: 1 Prediction Results

ML	Water %	Vegetation %	R-NDCWI	Accuracy %
Naïve Bayes	1%	37%	17	67%
SVM	1%	37%	17	69%

Results from the decision tree algorithm have been excluded as it didn't show any variance with different training datasets.

While water % is relatively low in 100Km x 100Km satellite image, with a raster slice to match the image with a county boundary, % water was observed to be higher in counties with rivers, lakes, and other water bodies.

1. *4.2.1 Use of Satellite Images as a Viable Option for Monitoring Algae*

With the combination of NDCI and NDWI, the presence of algae was easily detected across 3 different tiles from Sentinel-2 MSI Level 2A products. With tiles having a low percentage of water, it was difficult to determine algae presence in inland water. However, algae detection using R-NDCWI showed a noticeable result with images having relatively large inland water bodies (lake, river, canals). These inland water bodies in turn showed high levels of chlorophyll during months in which algae coverage warnings were issued. Among all sentinel images analyzed, the presence of algae, as indicated with R-NDCWI was observed to be the highest during July and August months.

4.2.2 Validation of Algae Extract from Satellite Image with Hfab Data

Relatively high R-NDCWI values showed a strong correlation with Hfab observed data available for incidents reported in the state of California. This indicated that the model of using high R-NDCWI to assume algae presence had merit due to a large

number of data points used and a lack of outliers that conflicted with the R-NDCWI.

4.2.3 Time Delay in WNV Cases

Based on the data analyzed, algae bloom that was detected from satellite imaging spiked during June to August period. West Nile Virus cases reported for California also saw a peak during the October to December period. The consistency of this trend indicated that these 2-3 months are crucial in the surveillance of vector-borne diseases. While this research was limited in scope and time, this could be significant in the overall monitoring of relatively large-scale areas' inland water bodies and to detect and prevent mosquito breeding in the inland water bodies.

4.2.4 Accuracy of the ML Prediction Model

Using Naïve Bayes with the 2020 and 2021 datasets, we achieved 67% accuracy with Sacramento, CA data. Accuracy slightly improved to 70% by combining both Sacramento and Fresno data. While the SVM prediction model gave a similar accuracy result, the Decision Tree machine learning algorithm was able to provide a nearly 100% accurate prediction. Due to time constraints and the limited cloud-free dataset available from Sentinel-2 for the 2020-21 summer, these results are indicative.

5. Conclusion

Both the field experiment and machine learning algorithms have a definite indication of the effect of fertilizer on algae blooms and the breeding of mosquitoes. However, this needs to be further tested with the larger dataset to train the prediction model. As a next step, cloud-free satellite images are to be sourced for an extended period (2015-2020) from multiple satellites (Landsat and Sentinel 2) to expand the dataset for training the prediction model. Early detection of potential breeding of mosquitoes could provide enough time to prevent a catastrophic impact of a West Nile Virus outbreak.

Acknowledgment

This work is part of NASA STEM Enhancement in the Earth Sciences (SEES) summer high school research internship 2022. Special thanks to Dr. Rusty Low and Peder Nelson for their guidance during this research and Ria Jain (Peer Mentor) for review and feedback to help finalize this research paper. We would also like to thank the NASA SEES internship program for giving us the background knowledge and support needed at every stage of our research.

The authors would like to acknowledge the support of the 2022 Earth Explorers Team, NASA STEM Enhancement in the Earth Sciences (SEES) Virtual High School Internship program. The NASA Earth Science Education Collaborative leads Earth Explorers through 8 award to the Institute for Global Environmental Strategies, Arlington, VA (NASA Award NNX6AE28A). The SEES High School Summer Intern Program is led by the Texas Space Grant Consortium at the University of Texas at Austin (NASA Award NNX16AB89A), or The SEES High School Summer Intern Program is in partnership with NASA Cooperative Agreement Notice NNH15ZDA004C Award NNX16AB89A.

Supplementary Material

1. NDCI and NDWI algorithm: python script
2. Feature Extraction from Algae data, West Nile Virus data for training ML prediction model
3. Sentinel2 MSI Images acquisition for Sacramento and Fresno, CA using Sentinelsat API in python (https://scihub.copernicus.eu/dhus/#/home)
4. California West Nile Virus Cases (https://data.chhs.ca.gov/dataset/west-nile-virus-cases-2006-present)
5. California Harmful Algae Bloom Incident Reports Map (https://mywaterquality.ca.gov/habs/where/freshwater_events.html)
6. Rasterio Library in Python (GDAL and NumPy based) (https://automating-gis-processes.github.io/CSC18/lessons/L1/Intro-Python-GIS.html#why-python-for-gis)
7. GIS Libraries in Python:
    1. Rasterio, GDAL and Geopandas
8. Data analysis & visualization in Python:
    1. Numpy, Pandas, Matplotlib and Scikit-learn

References

1. Centers for Disease Control and Prevention. (2021, December 17). *Final cumulative maps and data*. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/westnile/statsmaps/cumMapsData.html
2. Kovach, T. J., & Kilpatrick, A. M. (2018). Increased Human Incidence of West Nile Virus Disease near Rice Fields in California but Not in Southern United States. *The American journal of tropical medicine and hygiene*, *99*(1), 222–228. https://doi.org/10.4269/ajtmh.18-0120
3. Mori, H., Wu, J., Ibaraki, M., & Schwartz, F. W. (2018). Key Factors Influencing the Incidence of West Nile Virus in Burleigh County, North Dakota. *International journal of environmental research and public health*, *15*(9), 1928. https://doi.org/10.3390/ijerph15091928
4. Taylor, C. A., & Heal, G. (2021). Using satellite data to detect algae and link it to fertilizer. *National Bureau of Economic Research* https://www.nber.org/system/files/chapters/c14615/c14615.pdf
5. Victor, T.J. and Reuben, R. (2000), Effects of organic and inorganic fertilisers on mosquito populations in rice fields of southern India. Medical and Veterinary Entomology, 14: 361-368. https://doi.org/10.1046/j.1365-2915.2000.00255.x
6. Yee, S. H., Yee, D. A., de Jesus Crespo, R., Oczkowski, A., Bai, F., & Friedman, S. (2019). Linking Water Quality to Aedes aegypti and Zika in Flood-Prone Neighborhoods. *EcoHealth*, *16*(2), 191–209. https://doi.org/10.1007/s10393-019-01406-6