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

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

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

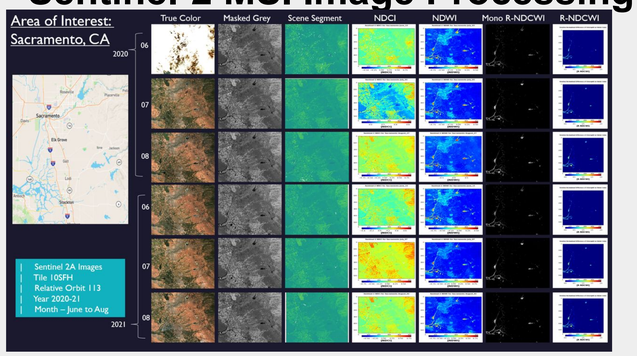
ABSTRACT

Monitoring mosquito abundance and its contributing indices are crucial for controlling the West Nile Virus (WNV), the most prevalent vector-borne disease in the contiguous United States. As mosquito-borne diseases like the 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.

RESEARCH QUESTIONS

1. **What is the impact of eutrophication on West Nile Virus (WNV) cases?**
2. **Can algae detection using satellite image be used for the surveillance and prediction of vector-borne diseases?**

Sentinel-2 MSI Image Processing



Methods

Methodology adopted two complementary approaches:

1. Experimental research using mosquito traps with fertilizers
2. Machine Learning (ML) model for algae detection and prediction of West Nile Viruses using Sentinel-2 images

For experience setup, 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.

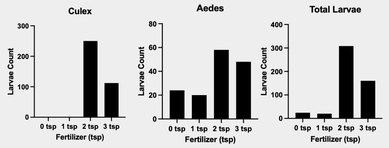
For ML prediction model using satellite images, 27 Sentinel-2 MSI images for Sacramento, CA Fresno, CA area was used.

Experimental research using mosquito traps with fertilizers

For experience setup, 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.

For the experiment, we split the experimental results into two groups for analysis: One for Culex mosquitoes and the other 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.

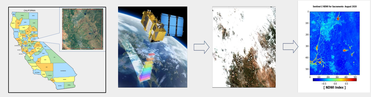
Results



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.

Machine Learning Predictive Model Using Sentinel-2 MSI

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.



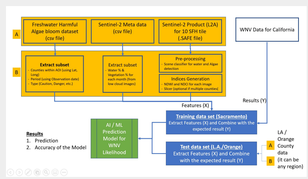
Pre-processing :

1. Exclusion of Images with Cloud Coverages - All images were processed using python script for water detection. Based on this 15 satellite images were selected for further processing.
2. Tile Slicing - Since data from the globe or on land observations are available by county within a state, it's necessary to slice satellite images to map with county boundary before it can be used for further analysis.
3. Scene Segmentation - Images were further analyzed for inland water (rivers, lakes, ponds, etc.) and vegetation to find the right "fit" images for algae detection.

Inland water and algae classification:

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.

The indices generated for each image were further enriched with additional features available from both satellite data as well as each observation data.

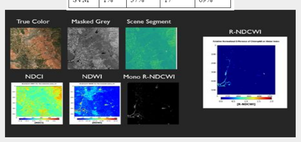


Machine Learning Predictive Model

Feature matrix (Vegetation%, Water %, NDCI, NDWI, Algae Observed, land cover data from GLOBE) were used to train supervised prediction model using Gaussian Naive Bayes and Support Vector Classifier algorithm.

Results and Visu

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



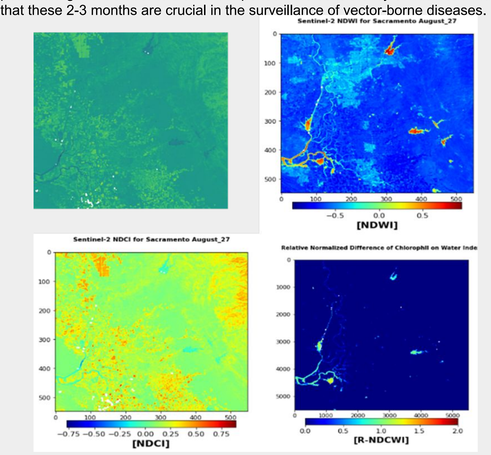
Discussion

Experimental research using mosquito traps with fertilizers discussion

- The 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.
- The significance of the Chi-Squared test does indicate that there may be some relationship and that this relationship could prove useful in predicting the number of mosquitos.

Machine Learning Predictive Model using Sentinel-2 MSI discussion

- The initial model created, showed an accuracy between 67% for the Naive Bayes model and 69% for the Support Vector Machine model
- Relatively high R-NDCWI values showed a strong correlation with Hfab observed data available for incidents reported in the state of California.
- The 67% accuracy we were able to achieve increased after revising the initial model to include both Fresno and Sacramento as training areas as opposed to just using Sacramento as a training area.
- Based on the data analyzed, algae blooms detected from satellite imaging spiked during June to August period and West Nile Virus cases reported for California 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.



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.
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 Naive Bayes Machine learning predictive model.

Acknowledgements

Rusty Low, Peder Nelson, Cassie Soeffing, Ria Jain

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.