A Random Forest Analysis of Remote Sensing Driven Mosquito Habitat Prediction in West Africa

Ruchi Bondre, Ryan Chan, Calista Huang, Andrew Liu, Neil Sangra
NASA STEM Enhancement in Earth Science 2023

Abstract

Mosquito vector-borne diseases, such as Dengue, West Nile Virus, Malaria, and Zika, pose significant global health risks for upwards of 3.9 billion people. An essential component to limiting the spread of mosquito vector-borne disease and assessing disease risk is the prediction of mosquito abundance. Given the severity of mosquito-borne diseases, there is a need for intelligent and automated mosquito abundance forecasting models. Such models would empower government and healthcare authorities to proactively address the mosquito threat and establish long-term disease prevention strategies. This study proposes the implementation of random forest models to predict mosquito larvae abundance in West Africa, suitable for forecasting future mosquito vector-borne disease outbreaks. Our models leverage remote sensing satellite data to extract features including normalized difference vegetation index (NDVI), average rainfall, temperature, humidity, and sporadically recorded GLOBE Mosquito Habitat Mapper (MHM) citizen science data to develop accurate predictions of mosquito population densities. We performed a comparative analysis of random forest classifiers and random forest regressors for the prediction of mosquito larvae counts as categories or numerical values, and determined both models to offer practical benefits for real-world implementation in mosquito habitat forecasting. The outcomes of our research indicate that random forest classifiers exhibit strong viability for predicting mosquito habitats and larvae abundance, achieving an accuracy of over 85%. Whether applied to a classification task or regression task, our work demonstrates the ability of random forest machine learning models to effectively identify correlations between environmental variables and mosquito population characteristics to predict mosquito abundance with high accuracy. In doing so, our research underscores the utility of remote sensing data and machine learning models for real-world mosquito threat management. Moreover, our results provide valuable insights for future research to address mosquito-borne disease prevention by targeting other areas or developing mosquito surveillance systems

Introduction

Mosquitos are the world's deadliest animal, accounting for more than 700,000 annual deaths (Helmer, 2023). Mosquito vector-borne diseases such as Dengue, West Nile Virus, Zika, Yellow Fever, and Malaria are serious public and animal health problems caused by parasites and bacteria transmitted by mosquitoes. Recent research suggests that global trends, modern transportation and globalization, urbanization, and climate change will likely exacerbate the risks of mosquito vector-borne disease, which has plagued living species for generations (Gubler, 2009; Rogers & Randolph, 2006; Ryan et al., 2019). Mosquitos can thrive in a variety of water sites, including fresh water, polluted water, brackish water, and turbid water, where they lay eggs that hatch into larvae (S.N.R et al., 2011; Sutherst, 2004). For outbreaks to occur, local vector levels need to be sufficiently high. Therefore, the ability to predict potential mosquito breeding sites and estimate mosquito abundance is an essential component of assessing disease risk (Kinney et al., 2021; Lega et al., 2017; Ryan et al., 2006).

Citizen science is an increasingly popular form of voluntary public participation in scientific research to expand scientific knowledge (Low et al., 2021). The GLOBE Observer app is a publicly-available application that allows citizen scientists to use their cameras to collect observations of their environment and contribute to the GLOBE database. Notably, the GLOBE Observer app offers a mosquito habitat mapper (MHM) tool for citizen scientists to photograph and classify mosquito larvae to add to the global database (GLOBE, 2022).

Predicting mosquito breeding sites is complicated by global climate change and weather factors. Previous research regarding mosquito risk identification identified specific temperatures as significant contributors to mosquito abundance, with a strong positive relationship between monthly relative humidity and mosquito larvae (Drakou et al., 2020). The Mosquito Landscape Simulation (MoLS), developed by Lega et al. (2017), is a mechanistic stochastic model for estimating Aedes aegypti mosquito abundance based on relative humidity, precipitation, and temperature. Their research demonstrated a model to predict Aedes aegypti abundance in real time using historical climate data coupled with available weather forecasts (Lega et al., 2017). Motivated by the difficulty of scaling such a model up to a large number of locations, Kinney et al. presented a faster Artificial Neural Network (ANN)-based alternative to MoLS using three base-ANN models incorporating recurrent layers, trained on weather time series data. Their research suggests that the use of ANNs trained on weather and surveillance data can effectively "contribute to the development of probabilistic mosquito abundance forecasting models" (Kinney et al., 2021).

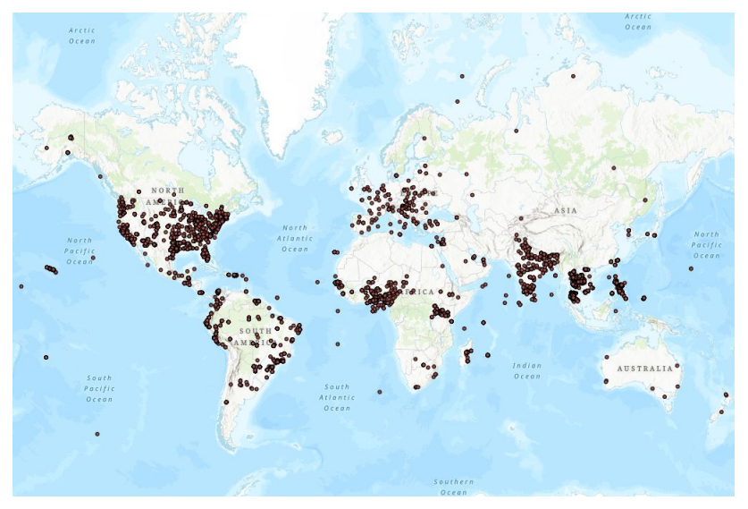
While recent research has focused on the use of AI and modeling to develop models to predict mosquito abundance and breeding sites, little research has combined land cover satellite data, remote sensing environmental data, and mosquito habitat citizen science data as a means to forecast mosquito abundance. In this paper, we present a random forest analysis of classification and regression algorithms that predict mosquito larvae abundance in Benin, Africa.

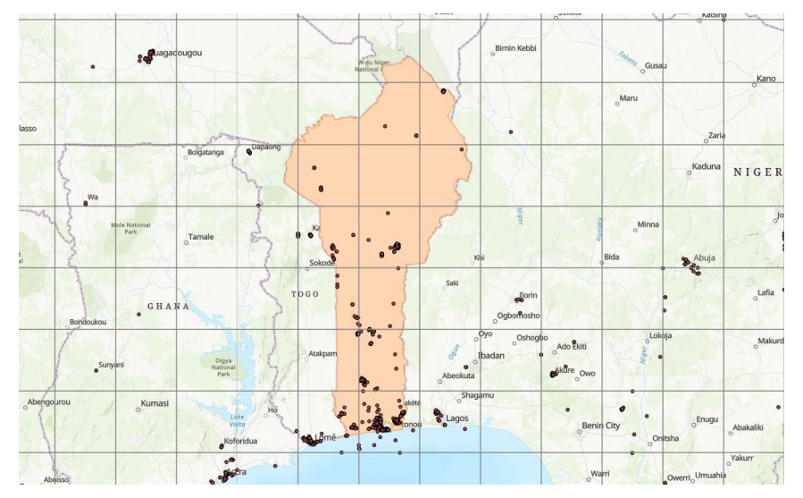
Benin was selected as an area of interest due to the higher density of data available from this location. The significant amount of mosquito habitat observations from Benin demonstrates the impact of the mosquito population in this area and the need for effective methods for mosquito abundance forecasting. This work seeks to compare random forest regression and classification for mosquito habitat prediction using remote sensing data in Benin.

Data

Mosquito Larvae Dataset

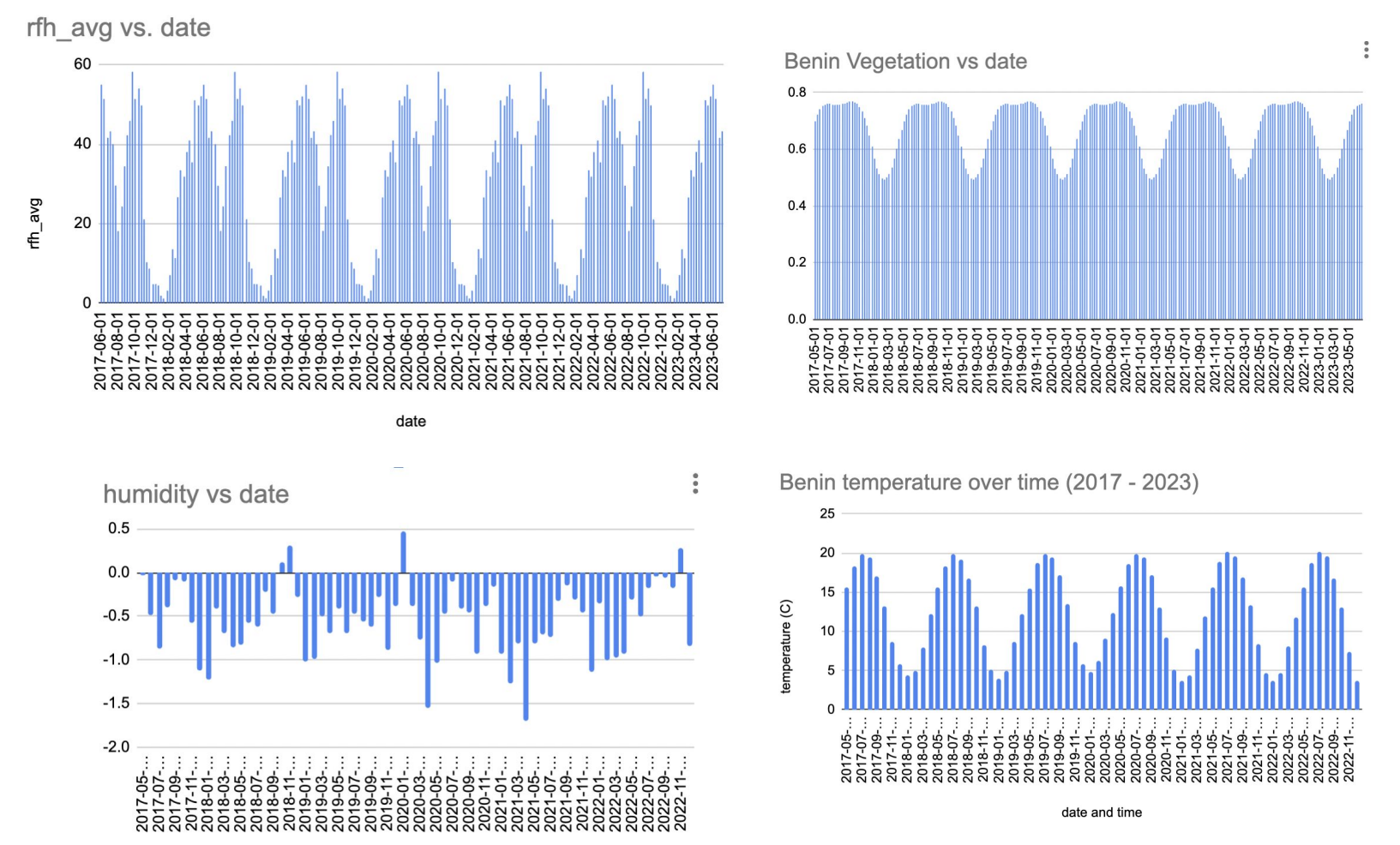
Mosquito larvae abundance data was obtained through the GLOBE Mosquito Habitat Mapper program (shown in figure 2 above), an app based tool where citizen scientists worldwide can submit data on mosquito habitats. The observations used in this study spanned from June 14, 2018 to July 5, 2022 , in Benin.





Remote Sensing Data

The remote sensing data used in this project included measurements of the 10 day Normalized Difference Vegetation Index (NDVI), 10 day rainfall (RFH), relative humidity, and temperature. NDVI and RFH data was obtained from the *Benin: NDVI at Subnational Level* and *Benin: Rainfall Indicators at Subnational Level* datasets through the Humanitarian Data Exchange, a humanitarian open data sharing platform run by the United Nations Office for the Coordination of Humanitarian Affairs. The data was contributed by the World Food Programme, a humanitarian organization dedicated to fighting hunger worldwide. The relative humidity and temperature data was obtained from the *HadISDH* dataset, a global gridded monthly mean surface humidity dataset maintained by the Met Office Hadley Centre.

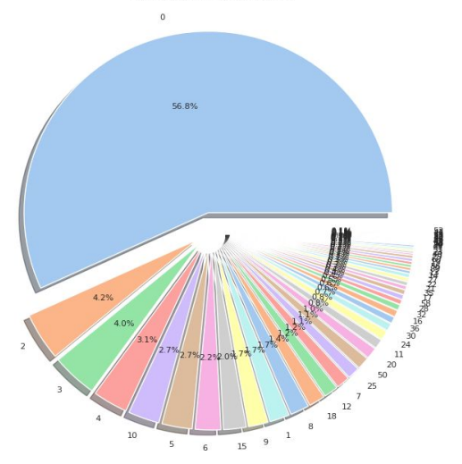


Graphs of vegetation, rainfall, humidity, and temperature over date from June 14th, 2018 to July 5, 2022 in Benin, Africa.

Data Preparation

Preprocessing Data

All observations with null and undefined (-9999) values were set to 0 for amount of larvae. The NDVI dataset had 59 observations each day every 10 days with no geocoding so we averaged the values of each day to create an accurate representation of the NDVI on that day.



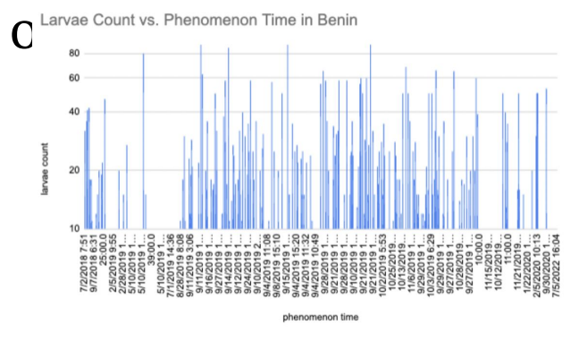
Combining 5 Datasets + Making Tier Levels

Once all the datasets were cleaned and normalized we merged them using the mosquito habitat mapper dataset as a base and attached the corresponding measurements from each remote sensing dataset by matching the closest in date and latitude and longitude.

After merging all the datasets we then created another column for larvae abundance tiers:

low (0): 0 - 9 larvae
medium (1): 9 - 22 larvae
high (2): 22+ larvae

Our quantity of larvae abundance for tiers was determined based on the distribution of c



Larvae Count Distribution				
Min	*Min non-zero*	*Max*	*Median*	*Average*
0	10	89	0	6.83297

Methodology

Random Forest Regressor

Random Forest Regressor is a powerful machine-learning algorithm used for regression tasks. It is an ensemble method that combines multiple decision tree models to make accurate predictions on continuous numeric data. The algorithm works by creating a multitude of decision trees during the training process and then aggregating their predictions to obtain a more robust and stable final output.

After splitting our dataset into a training set and testing set, we trained a baseline Random Forest Regressor. To optimize our evaluation metrics, we used a random grid to search for the optimal hyperparameters after creating RandomForestRegressor() mode. This optimization method conducts a random search of parameters using 3-fold cross-validation and searches across 100 different combinations, and uses all available cores. The optimal parameters were the following:

```
{'criterion': 'log_loss', 'max_depth': 118, 'min_samples_leaf': 337,
'min_samples_split': 124, 'n_estimators': 238}
```

Random Forest Classifier

Random Forest classifiers make use of a labeled dataset where each data point is associated with a target class label. Bootstrap Aggregating, or Bagging, is then utilized to create multiple random subsets within the dataset, some data points may be repeated, and others may be left out. For each subset of the data, a decision tree is built by recursively splitting the data based on features to make decisions and predict the target class.

Random forest models include many hyperparameters such as param_distributions or n_iterations, that can be altered to produce a model with a higher accuracy. Utilizing a RandomizedSearchCV function allows us to train many models with different parameters for each model, and make use of the "best_estimator" and "best_params_" attributes to respectively print out the best accuracy achieved, along with the parameters used to achieve it. The parameters used were:

```
{max_depth = 80, max_features = 'auto', min_samples_leaf = 4,
min_samples_split = 5, n_estimators = 200}
```

Results

The results are shown in Table 2 below. The random forest classifier performed generally better, which was expected as the classification task at hand involved simpler decision boundaries, and the regression tasks required the model to capture complex relationships between variables to predict continuous values accurately.

Algorithm Results				
Model	*MSE*	*RMSE*	*R^2*	*MAE*
RF Regressor	114.381	10.695	0.079	7.281
	Accuracy			
RF Classifier	0.8512			

TABLE 2. Results from random forest regression and classification.

One potential issue faced during the training of the models was the overrepresentation of 0 larvae counts. This overrepresentation as well as the lack of constant data taken by MHM plays into the sources of error which we suspect is the primary cause of insufficient results by the regression model. While under-sampling only seemed to decrease the accuracy, the use of a random_state to ensure the reproducibility of the training and testing data split allowed us to split the data in such a way that the imbalance was solved to some extent.

The main difference between regression and classification is that regression predicts numerical larvae count values whereas classification predicts the category of larvae count. Because precise numbers may not be required when assessing disease risk, an understanding of the concentration of mosquitos, as provided by classification, leaves room for a range of larvae count values to be classified and considered accurate, making this method robust and suitable for real-world mosquito threat management. Regression could serve to illustrate trends in mosquito larvae counts over time, allowing for the forecasting of potential spikes or high-risk periods in mosquito populations.

Discussion

Considering the threat of vector-borne diseases, machine learning models like the ones we implemented in our research are vital to forecasting disease outbreaks and developing disease management strategies. Notably, our results demonstrate that whether applied to a classification or regression task, random forest models show promise for real-world mosquito abundance prediction, providing further evidence that this type of model is well-suited to identifying correlations between ecological variables and mosquito population characteristics. The high accuracy achieved by our models, despite lacking consistently recorded data, demonstrates our random forest models' ability to handle multi-dimensional data and effectively predict mosquito habitats.

Although our models achieved high accuracy, there are several limitations to be aware of when understanding our results. Firstly, since GLOBE Observer Mosquito Habitat Mapper data comes from citizen scientists, the data acquired did not come at regular or consistent time intervals, limiting the amount of representative data points we were able to align with environmental sources to ultimately include in our training dataset. Furthermore, data collection methods may not be accurate or consistent between volunteer observers, and it may be important to acknowledge certain features of observations. For example, the types of tools available to volunteers, such as microscope lenses, may affect the accuracy of observations, and thus have the potential to be influencers in our models.

Similar studies have modeled potential mosquito habitats by examining aerial satellite images and aquatic habitats or calculating habitat suitability to indicate where certain mosquito species are most likely to occur (Mushinzimana et al., 2006; Cleckner et. al, 2011). The integration of remote sensing environmental variables with mosquito trap data or land cover images highlights that our methods of leveraging remote sensing environmental data, land cover data, and mosquito habitat data likely reveal important information about mosquito abundance. While our models did not extend to examining specific mosquito species, with adequate data mapping mosquito species and their habitats, our models could be adapted to predict mosquito habitats and disease risks of specific species.

Conclusion

Our random forest classifier optimized with random search achieved its highest accuracy of 85.12 %. Adopting mosquito larvae prediction models like demonstrated could greatly enhance mosquito vector-borne disease management, by enabling authorities to have advanced knowledge of mosquito risks and develop proactive strategies for preventing disease spread. Our models could be applied to data in other target locations to provide accurate and relevant forecasting models for mosquito larvae trend observation.

Our results provide valuable insights for future research to develop surveillance systems or alert mechanisms to aid in mosquito-borne disease prevention efforts. In particular, future research could experience with other machine learning architectures or ensembling techniques to achieve better results. As a note, when we later tested an AdaBoost classification model, we achieved an improved accuracy of 92.144%. Based on our results, future work could also explore alternative approaches for classifying larvae count into tiers, leverage additional features such as location and date, or integrate more mosquito datasets from other sources such as iNaturalist, VectorSurv, or Mosquito Alert to acquire more consistent data for improved prediction accuracy.

Acknowledgements

We would like to thank the NASA SEES Internship program for giving us the background knowledge and support needed to start this project. We acknowledge our mentors Russanne Low, Peder Nelson, Cassie Soeffing, Andrew Clark, and Erika Podest for their support and guidance in the research and publication process. We would also like to thank our peer mentors Aidan Schneider, Aswin Surya, and Kellen Meymarian, for their guidance through solidifying our project idea and utilizing online resources.

References

CDC. (2020, March 5). *Mosquitoes in the US | CDC*. Centers for Disease Control and Prevention; CDC. https://www.cdc.gov/mosquitoes/about/mosquitoes-in-the-us.html

Cleckner, H. L., Allen, T. R., & Bellows, A. S. (2011). Remote Sensing and Modeling of Mosquito Abundance and Habitats in Coastal Virginia, USA. *Remote Sensing*, *3*(12), 2663–2681. MDPI. https://doi.org/10.3390/rs3122663

For complete references, scan here:

The material contained in this poster is based upon work supported by the National Aeronautics and Space Administration (NASA) cooperative agreements NNX16AE28A to the Institute for Global Environmental Strategies (IGES) for the NASA Earth Science Education Collaborative (NESEC) and NNX16AB89A to the University of Texas Austin for the STEM Enhancement in Earth Science (SEES). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA.