Predicting Mosquito Abundance in Chicago Using Remote Sensing Climate Data and Machine Learning

Sheil Dharan, Daisy Li, Alan Monteiro, and Giovanni Victorio

NASA STEM Enhancement in Earth Sciences: GLOBE Earth System Explorers 2022

Note: 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, SEES, NESEC, CSR, or the Texas Space Grant Consortium

Abstract

In recent years, mosquito-borne diseases such as the Zika virus, West Nile virus, Chikungunya virus, Dengue, and Malaria have become more prevalent in urban areas due to various climate and anthropogenic factors. This led to a greater need for **mosquito abundance** prediction to improve the response to disease outbreaks, especially during the summer when mosquito season peaks and outdoor activities increase significantly. The objective of this study was to evaluate the accuracy of six **machine learning** models for **classifying** extreme mosquito abundance events based on climate data. Data sourced from the Mosquito Habitat Mappers challenge on GLOBE and a City of Chicago dataset were matched to area-averaged time-series climate data for Chicago from GIOVANNI, a NASA open access **remote sensing** database for Earth science. Data was cleaned and then aggregated to a single weekly time-series dataset consisting of **mosquito abundance**, and the past week's three **climate variable** averages. The models were trained and tested on climate data, namely surface humidity, precipitation, and daytime temperature. The mosquito and climate data were recorded from five Chicago summers. The results indicated that the best models for predicting **mosquito abundance** events were the ensemble learning methods of AdaBoost and Random Forest, respectively. Future avenues of research include using other, more-specific factors for prediction such as the chlorophyll from algal blooms (increasingly common due to direct and indirect anthropic activities, such as fertilizer runoff and warming waters due to climate change), more localized predictions, accounting for the microclimates of urban areas, and using regression models to predict precise mosquito numbers.

Introduction

Mosquito-borne diseases account for over 17% of all infectious diseases and cause more than 700,000 annual deaths according to (World Health Organization [WHO], 2020). In fact, (National Association of County & City Health Officials [NACCHO], 2017) found in a survey that 84% of surveyed vector-control operations are lacking in at least one out of the five core competencies of vector-control. Chicago is one city that experiences mosquito-borne diseases, particularly the West Nile Virus. Instances such as the Chicago West Nile Virus outbreak of 2002 as well regular cases of the virus occur in the City of Chicago. The virus affects the urban and suburban areas of Chicago and is a very serious illness. The prevalence of the West Nile Virus has been shown in (Tedesco et al., 2010). . The influence of temperature was shown to be significant and mostly positive, augmenting growth rates of populations (Paz, 2008): warming of the mosquito environment boosted their rates of reproduction and number of blood meals, prolonged their breeding season, excluding the case of extreme temperatures exceeding mosquitos' survivability limits (Drakou et al., 2020). Mosquitoes become inactive to maintain body fluids and reduce energy use in low humidity environments. As a result of the insufficient treatment methods and prevalence of mosquito-borne disease outbreak in urban areas, a method of predicting mosquito abundance is vital to preventing the spread of disease. Much of recent research in this area has applied machine learning to this task because some machine learning algorithms, particularly supervised machine learning algorithms, are efficient at modeling relationships between features. Most past studies that utilized machine learning for disease prevention utilized only the Neural Network and Support Vector Machine (SVM) machine learning models according to (Schaefer et al., 2020). Machine learning has also been used to predict mosquito abundance based on socioeconomic and land cover data such as that of (Chen et al., 2019). The goal of this study is to compare the performance of the predictions of mosquito abundance of six machine learning algorithms learning off of the climate variables of temperature, humidity, and precipitation.

Methodology

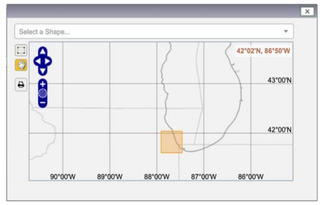

Fig. 1. Bounding box of Chicago used for mosquito habitats and area-averaged remote sensing climate data as shown in the GIOVANNI interface, ("Giovanni").

Both climate and mosquito data used in this study were obtained for the Chicago area with the bounding box of -87.9110W, 41.60581N, -87.4606W, 42.0417N. The climate data obtained from GIOVANNI was Area-Averaged for the bounding box

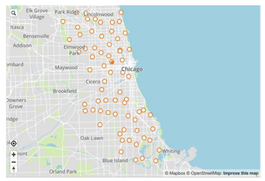

Fig. 2. City of Chicago Data Portal of mosquito observation locations for the five Chicago summers from 2017 to 2021, (City of Chicago, 2022).

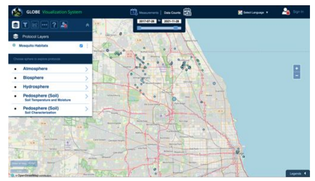

Fig. 3. GLOBE Data Visualization of the Mosquito Habitat protocol locations for the five Chicago summers from 2017 to 2021, ("GLOBE Program").

Mosquito data counts were then obtained via regular (weekly) mosquito trap measurements in the Chicago area recorded using the GLOBE Habitat Mapper protocol as well as regular (weekly) mosquito trap measurement measured and recorded in the City of Chicago data portal.

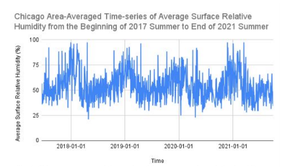

Fig. 4. Chicago area-averaged time-series of the average surface relative humidity from the beginning of 2017 summer to the end of 2021 summer. Graph created by authors; data retrieved from GIOVANNI, (AIRS Science Team & Teixeira, 2013).

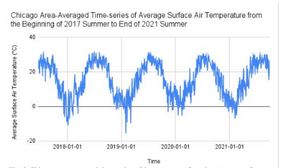

Fig. 5. Chicago area-averaged time-series of the average surface air temperature from the beginning of 2017 summer to the end of 2021 summer. Graph created by authors; data retrieved from GIOVANNI, (AIRS Science Team & Teixeira, 2013).

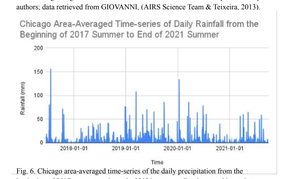

Fig. 6. Chicago area-averaged time-series of the daily precipitation from the beginning of 2017 summer to the end of 2021 summer. Graph created by authors; data retrieved from GIOVANNI, (Huffman et al., 2019).

Remote Sensing climate data for temperature, humidity, and precipitation were obtained, cleaned, and finally aggregated into weekly averages. The result of this process was a weekly mosquito count, temperature, humidity, and precipitation.

A mosquito abundance boolean is added to the dataset where over 1386.743 mosquitoes (the average, assuming mosquito frequency for a timeseries is a Poisson distribution).

Using the Orange Data Mining software, data were split into 70% training on data and 30% for training data, normalized for increased effectiveness, and put through the machine learning pipeline for 6 different supervised machine learning classifiers.

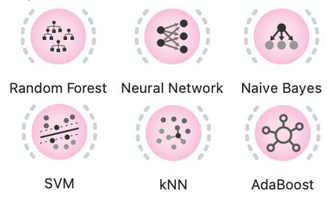

Fig. 7. The six machine learning models used in Orange Data Mining. Image from (Demsar, 2013).

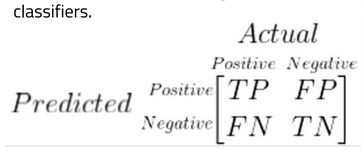

Fig. 8. Confusion matrix (a matrix generated for the testing data of each of the machine learning models) . Image generated using (Codecogs).

The confusion matrix values are used to calculate the performance scores: AUC, Classification Accuracy, Precision, Recall, and F1 Score

For more details describing calculations of these scores, see research report.

Results

Classification Metrics

Model	*AUC*	*CA*	*Precision*	*Recall*	*F1*
Random Forest	**0.988**	**0.944**	**0.945**	**0.944**	**0.944**
Neural Network	0.871	0.748	0.748	0.748	0.748
Naïve Bayes	0.920	0.868	0.854	0.856	0.855
SVM	0.885	0.772	0.771	0.772	0.771
k-NN	0.822	0.748	0.746	0.748	0.747
AdaBoost	**0.996**	**0.973**	**0.963**	**0.962**	**0.962**

Table. 1. Comparison of the performance off six different classification machine learning models as measured by the five standard classification metrics. The highest performing models are AdaBoost, followed closely by Random Forest.

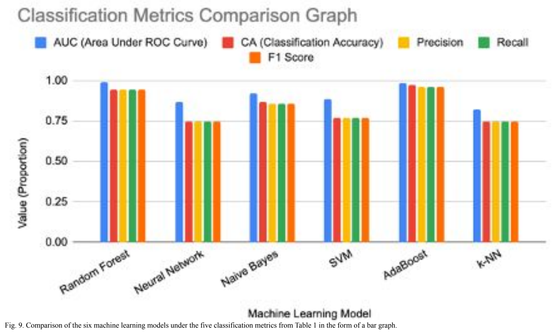

Fig. 9. Comparison of the six machine learning models under the five classification metrics from Table 1 in the form of a bar graph.

Discussion

All six models performed well with classification accuracies above 75%. Algorithms that performed the best were the ensemble learning algorithms of the Random Forest classifier (with a classification accuracy of 94.4%) and the AdaBoost classifier (with a classification accuracy of 99.6%). AdaBoost is known to have low generalization error, which means that the algorithm is much less prone to overfitting and performs better classification on previously unseen (testing) data. This phenomenon was noted in (Vezhnevets & Vezhnevets, 2005). The Random Forest classifier may have also performed well because of its ability to balance data, as stated in (More & Rana, 2017), when the amount of data in one class outnumbers the amount of data in another class, as it it did in this study. The Neural Network, SVM, and k-NN performed the worst. (Oleinik, 2019) states that because neural networks are good at identifying patterns in structured data, they are limited when they must combine pattern combination and recognition. According to (Yadav, 2018), SVM is less effective in low-dimensional spaces, so the use of only three climate variables may have limited the model's effectiveness. Finally, the k-NN classifier's performance may be limited to redundant features in the data as all features contribute similarly as noted in (Imandoust & Bolandraftar, 2013). Some possible sources of error include missing climate data values and confounding variables on mosquito traps (like urban microclimates, river eutrophication, and human activity). Future studies should collect mosquito data for the study itself. Similar studies to this one include (Chen et al., 2019), which predicted abundance with machine learning based on socioeconomic and land cover factors in Charlotte, NC and (Gardner et al., 2013), which correlated vegetation and focused on spatial distribution in basins in Chicago, IL.

Conclusion

Predicting mosquito abundance in urban areas is important as preventative measures can be taken to stop the abundance, and thus potentially stop an outbreak of a mosquito-borne disease. The findings show the need for high-quality, public citizen science datasets that can be used for data analysis to increase scientific knowledge and solve local problems. In the future, using sensor data at traps for recording climate data, regression analysis to predict precise numbers, and using satellite imagery to identify potential mosquito habitats. Furthermore, other factors could be used to predict mosquito abundance such as chlorophyll, an indicator of algal blooms and eutrophication, and socioeconomic factors such as housing prices as they are extremely variable factors in Chicago (Chen et al., 2019; Schelske & Stoermer, 1971). Future GLOBE protocols that could be used include land cover for mosquito habitat prediction, pedosphere protocols for soil characterization, and other hydrosphere protocols utilizing fluids variables such as nitrates and pH

Acknowledgements

We would like to acknowledge our research mentors for their continued technical support throughout our research: Dr. Rusty Low, from the Institute of Global Environmental Strategies; Dr. Cassie Soeffing from the Institute for Global Environmental Strategies; Andrew Clark from the Institute for Global Environmental Strategies; Peder Nelson from Oregon State University; Dr. Erika Podest from NASA's Jet Propulsion Laboratory; and our peer mentor, Alessandro Greco, from Western Canada High School. In addition, we would like to thank everyone from the SEES STEM Enhancement in Earth Science 2022 cohort and especially the GLOBE Earth System Explorers 2022 cohort including all our mentors, administrative members, guest speakers, and peers.

References

[1] Vector-borne diseases. (2020). *World Health Organization*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases

[2] Mosquito Control Capabilities in the U.S. (2017). *National Association of County & City Health Officials*. Retrieved from https://www.naccho.org/uploads/downloadable-resources/Mosquito-control-in-the-U.S.-Report.pdf

[3] Tedesco, C., Ruiz, M., & McLafferty, S. (2010). Mosquito politics: Local vector control policies and the spread of West Nile Virus in the Chicago region. *Health & Place* (Vol. 16, Issue 6, pp. 1188–1195). https://doi.org/10.1016/j.healthplace.2010.08.003

[4] Paz, S., Albersheim, I. (2008). Influence of Warming Tendency on *Culex pipiens* Population Abundance and on the Probability of West Nile Fever Outbreaks (Israel: Case Study: 2001–2005). *EcoHealth* **5**, 40–48 (2008). https://doi.org/10.1007/s10393-007-0156-0

[5] Drakou, K., Nikolaou, T., Vasquez, M., Petric, D., Michaelakis, A., Kapranas, A., Papathedoulou, A., & Kolios, M. (2020). The Effect of Weather Variables on Mosquito Activity: A Snapshot of the Main Point of Entry of Cyprus. *International journal of environmental research and public health, 17*(4), 1403.

[6] Schaefer, J., Lehne, M., Schepers, J., Prasser, F., & Thun, S. (2020). The use of machine learning in rare diseases: a scoping review. *Orphanet Journal of Rare Diseases* (Vol. 15, Issue 1). https://doi.org/10.1186/s13023-020-01424-6

[7] Chen, S., Whiteman, A., Li, A., Rago, T., Delmelle, E., Chen, G., Brown, C. L., Robinson, P., Coffman, M. J., Janies, D., & Dulin, M. (2019). An operational machine learning approach to predict mosquito abundance based on socioeconomic and landscape patterns. *Landscape Ecology* (Vol. 34, Issue 8, pp. 1295–1311). https://doi.org/10.1007/s10980-019-00839-2

[8] Giovanni. (n.d.). The Bridge Between Data and Science. NASA GIOVANNI. Retrieved from https://giovanni.gsfc.nasa.gov/giovanni

[9] City of Chicago (2022). West Nile Virus (WNV) Mosquito Test Results: City of Chicago: Data Portal. Retrieved from https://data.cityofchicago.org/Health- Human- Services/West- Nile- Virus- WNV- Mosquito- Test-Results/jqe8-fnfu

[10] Global Learning and Observations to Benefit the Environment (GLOBE) Program, *Data Accessed: 2022, July 10*, Retrieved from globe.gov

[11] AIRS Science Team/Joao Teixeira (2013), AIRS/Aqua L3 Monthly Standard Physical Retrieval (AIRS-only) 1 degree x 1 degree V006, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), https://doi.org/10.5067/Aqua/AIRS/DATA321

[12] Huffman, G.J., E.F. Stocker, D.T. Bolvin, E.J. Nelkin, Jackson Tan (2019), GPM IMERG Final Precipitation L3 1 day 0.1 degree x 0.1 degree V06, Edited by Andrey Savtchenko, Greenbelt, MD, *Goddard Earth Sciences Data and Information Services Center (GES DISC)*,

[13] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, *Journal of Machine Learning Research* 14(Aug): 2349−2353.

[14] Codecogs. (n.d.). Code Cogs Equation Editor. *CODECOGS*. Retrieved from https://latex.codecogs.com/eqneditor/editor.php

[15] Vezhnevets, Alexander & Vezhnevets, Vladimir. (2005). 'Modest AdaBoost' - Teaching AdaBoost to Generalize Better. *Graphicon*. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.2346&rep=rep1&type=pdf

[16] More, A.S., & Rana, D.P. (2017). Review of random forest classification techniques to resolve data imbalance. *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, 72–78. https://doi.org/10.1109/ICISIM.2017.8122151

[17] Oleinik, A. (2019). What are neural networks not good at? On artificial creativity. Big Data & Society. https://doi.org/10.1177/2053951719839433

[18] Imandoust, S.B., & Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: *Theoretical Background*. Retrieved from https://www.ijera.com/papers/Vol3_issue5/DI35605610.pdf

[19] Gardner, A. M., Anderson, T. K., Harner, G. L., Johnson, D. E., Varela, K. E., Walker, E. D., & Ruiz, M. O. (2013). Terrestrial vegetation and aquatic chemistry influence larval mosquito abundance in catch basins, Chicago, USA. *Parasites & Vectors* (Vol. 6, Issue 1). https://doi.org/10.1186/1756-3305-6-9

[20] Schelske, C. L., & Stoermer, E. F. (1971). Eutrophication, Silica Depletion, and Predicted Changes in Algal Quality in Lake Michigan. In Science (Vol. 173, Issue 3995, pp. 423–424). *American Association for the Advancement of Science (AAAS)*. https://doi.org/10.1126/science.173.3995.423

IVSS Badges

I am a Data Scientist: The machine learning models in this study utilized a dataset created by us (the authors) which was compiled from a variety of data sources including the City of Chicago Data Portal, GLOBE, and GIOVANNI. We discuss the limitations of the data in our discussion (IV) as the GIOVANNI data was missing values and did not include errors and mosquito measurements might be influenced by confounding variables that are not controlled for as a result of using public data instead of collecting our own data. The data was also limited in terms of availability and reliability. Many data options were incompatible with our study such as climate data being recorded every 8-days instead of 7-days, which led us to averaging the daily values of a climate variable each week. In addition, this study uses the data with machine learning models to make inferences (predictions) about mosquito abundance events in the future. These inferences are made at a high accuracy and various standard classification metrics and statistical concepts are used to evaluate each model's performance in predicting mosquito abundance in Chicago. Our data analysis aimed to solve the problem of mosquito-borne disease outbreaks in urban areas as predicting mosquito abundance and thus enacting preventative measures can stop the spread of mosquito-borne diseases.

I am a Collaborator: As we (the authors) come from completely different backgrounds and parts of the world spanning three different time zones, we each brought our own skills which were vital to completing this project. This project required the integration of many skills including machine learning, mosquitoes, data analysis, literature review, scientific writing, and climatology. S.D. has a background in computer science and machine learning, so he used his knowledge to clean the data into a usable format and develop the machine learning models and their pipelines. D.L. has a background in Earth science data, so she was able to obtain and clean the mosquito data as well as climate data. A.M. has a background in scientific writing and was able to research the background and aide in the formatting of the report. G.V. has a background in graphic design and created graphics for the report and designed the presentation. All authors had experience with writing in general and wrote and gave feedback on the writing of the report. Without collaborative effort, this study would not have been possible as no one person had the background to complete this entire project. Furthermore, working as a team allowed us to get crucial feedback to improve all aspects of the study and report. With diverse backgrounds from schools across the Americas, as a team, we were able to develop creative solutions that we encountered throughout the research process.

I Make an Impact: All of us (the authors) come from humid climates near urban areas. This means that mosquito-borne diseases are a local issue in all four of our communities. There is a need for prediction of mosquito abundance to enact preventative measures to inhibit the spread of disease. However, not all communities have the resources nor the infrastructure to enact enhanced methods of mosquito prevention without prediction. Our results utilize readily available remote sensing climate data to predict mosquito abundance. This can act as a cost-effective way to predict mosquito abundance and precent disease spread for communities who cannot afford them. Though our results focused on Chicago, our study could be applied to our local communities once mosquito data is available. We plan to share our research with communities who can utilize them and take effective preventative measures to stop the spread of mosquito-borne diseases.