Beyond sporadic outbreaks: Classifying and explaining dengue endemicity over scenarios of global change

Hailey Robertson[1]

1. Department of Epidemiology of Microbial Diseases, Yale School of Public Health

Background and motivation

Dengue virus (DENV), a member of the Flaviviridae family, is a vector-borne RNA virus with four serotypes carried by *Aedes aegypti* and *Aedes albopictus* mosquitoes [1]. All four DENV serotypes can cause disease in humans, ranging from asymptomatic to mild fever to dengue hemorrhagic fever and dengue shock syndrome [1]. In 2024, there were over 13 million reported dengue cases, making it the largest dengue season on record [2]. While the majority of the global burden of dengue is reported in Latin America, the Caribbean, and Southeast Asia, it is estimated that over 50% of the world's population lives in areas suitable for DENV transmission [3–5]. These suitable areas are characterized by warm temperatures, high humidity, and other conditions that support the survival and breeding of *Aedes sp.* mosquitoes [6,7]. Climate change is projected to further expand the suitable range for DENV transmission by increasing temperatures and altering precipitation patterns, allowing *Aedes sp.* mosquitoes to thrive in new regions that were previously too cool or dry for sustained transmission [8–10].

One challenge associated with this change in vector range is the possibility for new areas to not just become suitable for dengue occurrence, but to become endemic for DENV. Understanding these dynamics requires novel methods at the intersection of ecology and epidemiology (eco-epidemiology) that go beyond classic species distribution mapping of the spread of mosquito vectors. For example, while *Aedes aegypti* mosquitoes are common in much of the United States and Europe, DENV has only caused brief outbreaks in Florida, Italy, and southern France, where it cannot persist without reintroduction. However, the minimum conditions that determine whether DENV transitions from hypo-endemic to endemic transmission following an urban introduction remain unclear, particularly regarding how environmental, socioeconomic, and ecological factors interact. These factors include temperature, land use, urbanization, and living conditions like the use of air conditioning or quality of water systems [9,11]. **Understanding which conditions modulate risk and their relative importance (e.g., how much does the climate need to change to overcome protective socioeconomic factors?) is essential for predicting when locales should prepare to make the transition to endemicity.** Accurately forecasting this transition has enormous public health consequences – estimates from Florida and Europe reveal that sporadic cases of dengue may only cause infections in < 0.05% of the population [12–14]. Large epidemics where DENV is stably maintained in an endemic cycle, however, may cause annual cases in >5% of the population . Therefore, the current peaks of 100-200 DENV cases reported in Miami, Florida could rise to ~25,000 per year if the virus were to become endemic.

Thus, I propose to use machine learning (ML) models, specifically multiclass logistic regression and *k*-means clustering to predict endemicity status (Aim 1) and identify the key contributing factors of endemicity status over time using explainable AI (xAI) (Aim 2). I hypothesize that climate change, human population size, and semi-urban landscapes will be significant drivers of stable DENV maintenance, though increased socioeconomic conditions in otherwise high-risk areas may act as protective factors [9,15,16]. By exploring spatial and temporal variation in these thresholds, we can move towards more precisely characterizing the DENV risk landscape with future climate scenarios and identifying areas with potential for large-scale outbreaks to ultimately improve prevention and control efforts in at-risk regions. This project is novel both in its exploration of DENV endemicity, which has not been quantified nor examined as an

outcome in epidemiological studies, as well as its integration of ML and xAI approaches with eco-epidemiology to predict and identify drivers of DENV transmission potential.

Aim 1: Classify countries by endemicity status.

The objective of Aim 1 is to predict if countries are **non-endemic** (absent), **hypo-endemic** (sporadic outbreaks, reintroductions), or **endemic** (trans-seasonal maintenance without reintroductions) for DENV. To make these predictions, I will compile existing global datasets on dengue cases, climate (temperature and precipitation), land use, and socioeconomic factors as features for endemicity status, shown in **Table 1.**

Table 1. Data dictionary for select* model predictors

*Note that more datasets may be included for the project but will include those below at a minimum.

Predictor	Data source	Rationale
Dengue cases	Open Dengue	Higher incidence of dengue is the primary measure of endemicity based on reporting of cases
Population density	World Bank	Higher population density leads to increased contact rates and longer transmission chains
Annual and monthly mean temperature	ERA5-Land	Geographical limits on vector abundance and dengue transmission; warmer temperatures may lead to year-round transmission
Precipitation	WFDE5 v2.1	Increases availability of mosquito breeding sites but also causes flushing; may be non-linear
Urban expansion rate	Landsat urban dynamics	Early stages of urbanization (e.g informal settlements, construction) may increase availability of suitable mosquito habitat
GDP per capita, PPP	World Bank	Higher GDP based on purchasing power parity (PPP) may be associated with better health infrastructure, vector control, and decreased transmission risk
International tourism	World Bank	Greater volume of international arrivals and departures create more opportunities for new virus introductions

I will train and compare two different ML models: 1) **supervised multiclass logistic regression** using softmax and categorical cross-entropy loss to predict endemicity status across countries, and 2) **unsupervised *k*-means clustering** to identify patterns across countries (and whether they align with endemicity status) without predefined labels. I will explore how these different models perform, with and without optimization techniques like (momentum-based) stochastic gradient descent for logistic regression. I will evaluate performance based on mean accuracy (≥ 0.70 for "good" accuracy, with subject-expert assessment) for logistic regression and silhouette scores (≥ 0.5 for "good" result) for *k*-means clustering. Hyperparameters will be tuned manually, but if time, I may experiment with cross-validation for optimization of hyperparameters. For logistic regression, I will account for overfitting by applying regularization, but if overfitting remains an issue, I will remove highly correlated predictors or use principal component analysis (PCA) prior to training to reduce dimensionality / complexity (with the goal of retaining components explaining ~95% of variance). For *k*-means clustering, I will account for overfitting by testing different numbers of clusters using the elbow method and using *k*-fold cross validation.

I anticipate that the multiclass nature of this problem may pose challenges, particularly in distinguishing hypo-endemic regions, where transmission patterns are less well-defined (both in the literature and in observed case data). If the models struggle to achieve reliable classification performance – such as low accuracy, poor class separation, or difficulty in optimizing decision boundaries – I may simplify the problem by converting it into a binary classification task. In this case, I would use the sigmoid function instead of softmax, and logistic loss instead of classification cross entropy. This adjustment would still capture the critical distinction between areas with sustained transmission and those without.

Aim 2: Use explainable AI to interpret predictions

In Aim 2, predictions will be explained with feature contributions by calculating SHapley Additive exPlanations (SHAP) values, which give the average marginal contribution of each feature across all possible combinations, allowing us to determine which factors are most influential for each prediction, i.e., the endemicity status of each country over time [17]:

$$\phi_i = \sum_{S \subseteq \{1, \ldots, n\} \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} [f(S \cup \{i\} - f(S)]$$

where *S* is the set of input features, *f(S)* is the model prediction using only features in *S,* and $f(S \cup \{i\}) - f(S)$ quantifies the marginal contribution of each feature *i.* By ranking features according to mean absolute SHAP values, I will identify the set of global factors most important for determining endemicity status as follows:

$$Feature\ Importance = \frac{1}{N} \sum_{j=1}^{N} |\phi_i^{(j)}|$$

where *N* is the number of instances and $\phi_i^{(j)}$ is the SHAP value of feature *i* for an individual data point *j* (e.g., a country). This can also be done for country-level factors. If time allows, I may remove features with consistently low SHAP values and re-run the model to simplify it and compare performance. Additionally, SHAP values capture feature interactions which I will use to reveal how variables work in tandem to drive endemicity status. If there are meaningful interactions, I will visualize these using dependence plots.

One possible challenge with this aim is that calculating SHAP values can be computationally expensive, which could be problematic in terms of memory and processing time. If SHAP computation becomes infeasible, I may use Local Interpretable Model-Agnostic Explanations (LIME) which is faster than SHAP but does not give exact attributions of features to model output or interactions between variables.

By leveraging xAI, we can ensure the model is not only accurate but also interpretable rather than a black box, which is of increasing concern in ecology and epidemiology [18]. Note that while this is not causal analysis, SHAP values provide insight into features consistently associated with changes in endemicity status. Additional subject-matter expertise will be required to confirm the underlying mechanistic relationships.

Conclusion

This project is ambitious, and while completing all aims as described within the timeframe may not be feasible (specifically Aim 2, as it is not material covered in class), any progress made will be valuable as it forms the foundation of my dissertation as a 1st-year PhD student in the School of Public Health. **More broadly, this research project sits at the cutting edge of methods in eco-epidemiology and is valuable to the field as it will improve our ability to predict and explain the transition from sporadic outbreaks**

to endemic dengue transmission. By comparing multiple ML approaches and integrating xAI, this project will help determine the endemicity status of a country to make informed decisions about interventions and generate precise risk assessments. In the future, these outputs can be used as covariates to project the future range of dengue endemicity over different climate scenarios and improve preparedness and resilience to global changes.

References

1. Vasilakis, N. & Weaver, S. C. The history and evolution of human dengue emergence. *Adv. Virus Res.* **72**, 1–76 (2008).
2. Dengue worldwide overview. *European Centre for Disease Prevention and Control* https://www.ecdc.europa.eu/en/dengue-monthly (2024).
3. Brady, O. J. *et al.* Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl. Trop. Dis.* **6**, e1760 (2012).
4. Bhatt, S. *et al.* The global distribution and burden of dengue. *Nature* **496**, 504–507 (2013).
5. Messina, J. P. *et al.* The current and future global distribution and population at risk of dengue. *Nat. Microbiol.* **4**, 1508–1515 (2019).
6. Mordecai, E. A. *et al.* Thermal biology of mosquito-borne disease. *Ecol. Lett.* **22**, 1690–1708 (2019).
7. Kraemer, M. U. G. *et al.* The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *Elife* **4**, e08347 (2015).
8. Ebi, K. L. & Nealon, J. Dengue in a changing climate. *Environ. Res.* **151**, 115–123 (2016).
9. Gibb, R. *et al.* Interactions between climate change, urban infrastructure and mobility are driving dengue emergence in Vietnam. *bioRxiv* 2023.07.25.23293110 (2023) doi:10.1101/2023.07.25.23293110.
10. Mordecai, E. A. *et al.* Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models. *PLoS Negl. Trop. Dis.* **11**, e0005568 (2017).
11. Franklinos, L. H. V., Jones, K. E., Redding, D. W. & Abubakar, I. The effect of global change on mosquito-borne disease. *Lancet Infect. Dis.* **19**, e302–e312 (2019).
12. Branda, F. *et al.* Dengue virus transmission in Italy: historical trends up to 2023 and a data repository into the future. *Sci. Data* **11**, 1325 (2024).
13. Rey, J. R. Dengue in Florida (USA). *Insects* **5**, 991–1000 (2014).
14. Kada, S., Paz-Bailey, G., Adams, L. E. & Johansson, M. A. Age-specific case data reveal varying dengue transmission intensity in US states and territories. *PLoS Negl. Trop. Dis.* **18**, e0011143 (2024).
15. Reiter, P. *et al.* Texas lifestyle limits transmission of dengue virus. *Emerg. Infect. Dis.* **9**, 86–89 (2003).
16. Nakase, T., Giovanetti, M., Obolski, U. & Lourenço, J. Population at risk of dengue virus transmission has increased due to coupled climate factors and population growth. *Commun. Earth Environ.* **5**, 1–11 (2024).
17. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv [cs.AI]* (2017).
18. Ryo, M. *et al.* Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography (Cop.)* **44**, 199–205 (2021).