**Abstract**

The three most important viruses transmitted by *Aedes* mosquitoes include Chikungunya (CHIKV), Dengue (DENV), and Zika virus (ZIKV). The spatio-temporal patterns of these *Aedes*-borne viruses (ABVs) throughout history have been dynamic, likely due to a combination of factors related to changes in the human host, viral genetics, mosquito vector, and the environment. Nearly 105 million dengue infections are reported globally per year. The recent spread of CHIKV and ZIKV into the Americas has increased the complexity of transmission and diagnosis due to these ABVs co-circulating in the same vector and human populations. In this study we present an integrated analysis of the reported Aedes-borne disease cases between 2012 and 2019, the local climate, and the socio-demographic profiles of 2,511 municipalities in Mexico. We used SaTScan (v. 9.6.1) to detect spatial clusters separately for CHIK, DEN, and ZIK. For each cluster, we utilized the Pearson correlation coefficient, Randomized Dependence Coefficient(RDC) and SHAP (SHapley Additive exPlanations) for analyzing the influence of the socio-demographic and climatic factors on the incidence of each disease. We also compared six popular models including XGBoost, decision tree, SVM with RBF kernel, KNN (K nearest neighbors), random forest and neural network for predicting the spatial clusters. Our results reveal that socio-demographic and climatic factors have significant and differential influence on ABV transmission in different regions of Mexico.

**Method**

**Spatial analysis**

**I will prefer some advance method**

<https://healthitanalytics.com/news/dod-using-artificial-intelligence-to-detect-coronavirus-hotspots>

COVID-19 Outbreak Prediction with Machine Learning (check from Google)

<https://www.researchgate.net/publication/261204615_Machine-Learning-Based_Hotspot_Detection_Using_Topological_Classification_and_Critical_Feature_Extraction>

<https://arxiv.org/pdf/2007.05879.pdf>

<https://www.nature.com/articles/s41598-017-03405-5>

<https://www.researchgate.net/publication/313353680_Spoton_A_Machine-Learning_Approach_for_Hot-Spot_Determination>

A GIS-Based Artificial Neural Network Model for Spatial Distribution of Tuberculosis across the Continental United States (check from Google)

**Will not use Sat Scan**

SaTScan (v. 9.6.1) was used to detect spatial clusters (adjusted for more likely clusters) separately for CHIK, DEN, and ZIK (settings: spatial analysis; discrete Poisson probability model; latitude/longitude coordinates; no geographical overlap; scanning for clusters with high rates). Spatial clusters were determined by calculating the maximum likelihood ratio. Standardized prevalence ratios were estimated by dividing the number of observed cases by the number of expected cases in each cluster. Simulated p values were obtained using Monte Carlo methods with 9,999 replications.30 Based on clustering information, we ran a different algorithm to show the influence of SES attributes and climate attributes for all clusters and non-clusters.

For analysis we use the Pearson correlation coefficient31, Randomized Dependence Coefficient32 (RDC) and SHAP33 (SHapley Additive exPlanations). We used these approaches to assess the influence of the socio-economic and climate factors for the incidence of each arbovirus. RDCis a measure of nonlinear dependence between random variables of arbitrary dimension based on the Hirschfeld-Gebelein-Renyi Maximum Correlation Coefficient. SHAPis a game theoretic approach to explain the output of any machine learning model. Details on each of these measures are found in the supplementary text. We use arbovirus disease cases as our target variable in our dataset, and socio-economic variables and climate variables as our features to build the model. We compute SHAP values based on XGBoost model. Based on the XGBoost model, we use SHAP to show the importance values for each feature with respect to our model. RDC is another approach in which can reflect the relationship between features and the target variable. For every feature in our dataset, we compute the importance coefficient between this feature and the target variable based on this RDC approach. We compute the correlation coefficient between every feature and the target variable for Pearson coefficients based on covariance of our dataset.

We used SHAP values, RDC and Pearson coefficients to indicate the socio-economic and climate attributes impact. In order to comprehensively summarize the result, we developed two evaluation metrics: majority vote and average. The majority vote metric gives the weighted impact for socio-economic and climate attributes based on majority of SHAP, RDC and Pearson results. In the case of the majority voting metric for a specific cluster, if SHAP values and RDC value indicate socio-economic attributes have more impact than climate attributes, we then took the average of the results of the SHAP values and the RDC values as the majority vote result for this cluster. For the average metric, we take average result of SHAP, RDC and Pearson as the average result.

We also compare six popular predictions, such as XGBoost, decision tree, SVM with RBF kernel, KNN (K nearest neighbors) with 5 neighbors, random forest with 6 estimators and neural network with 100 hidden layers. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

We use 10-fold cross validation for this given data, the details of which is left to the supplementary text.

**Ethical approval**

This study has been approved by the ethical committee of UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN, Mexico.

**Results**

### ***General disease pattern between 2012 - 2019***

Dengue was the most prevalent of the three diseases throughout Mexico. Nearly 60.6% (1,498/2,469) municipalities reported DENV, 29.3% (723/2,469) municipalities reported CHIKV, and 31.2% (771/2469) municipalities reported ZIKV. Of the municipalities, 2.1% (52/2469) reported all three ABVs. However, 39.6% (978/2469) of the municipalities in Mexico never reported any case of disease from these viruses during the period of 2012 - 2019. In total, 2,6211 CHIK, 224,701 DEN, and 12,813 laboratory confirmed ZIK cases were reported throughout the eight year study period. In Mexico, 67 municipalities constantly reported more than 1% DENV prevalence, with the Tomatlán (Jalisco) municipality in the state of Jalisco having reported the highest prevalence (2.48%). A sharp increase in CHIKV, DENV, and ZIKV cases were reported in Veracruz.  There was no fluctuation in the climatic parameters over the years.

***Spatiotemporal clusters***

Identified spatial clusters of CHIK, DEN, and ZIK are shown in Figure 1. Twenty-one statistically significant (p = 0.0001) clusters were observed in Mexico. We analyzed for all clusters and non-clusters, as socio-economic (SES) features and climate features may have different levels of impact for all clusters and non-clusters. Table S1 indicates the majority vote and average results for CHIK, DEN, and ZIK viruses based on different clusters. There were three spatiotemporal clusters of CHIKV (Table S1). SES features have more impact than climatic features on model output in cluster 1, and 3 (Figure S1-S3). There were twelve spatiotemporal clusters of DENV (Table S1). Climatic features have more impact than SES features on model output in cluster 1, 3, 4, 5, 6, 7, and 12 (Figure S4-S15). There were six spatiotemporal clusters of ZIKV (Table S1). Climatic features have more impact in cluster 1, 2, and 3, whereas SES features have more impact than climatic features on model output in clusters 4, 5, and 6 (Figure S16-S21).

Table 1 displays the performance of various ML classification algorithms across all clusters after taking the average. Table 3 shows performance of various ML classification algorithms across non clusters after taking average. The results show that XGBoost is performing the best in terms of precision-measure for, CHIK, DEN, and ZIK (Table 1, 3). The F1 scores of XGBoost for CHIK, DEN, and ZIK are larger than other baseline approaches, which means XGBoost has better performance than other approaches (Table 1). The values of accuracy are larger than weighted accuracy and precision values (Table 1, 3). For instance, in Table 1 the accuracy of XGBoost under DEN is 0.87 which is higher than corresponding weighted accuracy 0.78 and precision 0.85. This may happen due to class imbalance issue. More specifically, with respect to normal cases (normal class), we have fewer infected cases.

Table 2 demonstrates the standard error of classification performance of various ML algorithms on CHIK, DENV, and ZIK virus prediction for all clusters. Table 4 indicates the standard error of classification performance of various ML algorithms on CHIK, DEN, and ZIK predictions for non-clusters. We presented the standard error of all clusters and non-clusters predictive results in terms of accuracy, weighted accuracy, precision, recall and F1 measure (Table 2, 4). XGBoost is performing the best for CHIKV, DENV, and ZIKV. The standard error of XGBoost is more stable than other baseline approaches. For example, in Table 2, the standard error of XGBoost under DEN for accuracy, weighted accuracy precision, recall and F1 score are respectively 0.07, 0.08, 0.13, 0.07 and 0.09, which are lower than most of the standard errors of the other baseline approaches.

For CHIKV, DENV, and ZIKV, the influence of socio-economics attributes is larger than the influence of climate attributes for some cases. For example, from Figure 2A, 2B, 3A, 4A we find socio-economics attributes have more impact than climate attributes. The weighted socio-economic attributes SHAP value is 0.61, and the weighted climate attributes SHAP value is 0.39 (Figure 3A).

From Figure 2A, 2B, 3A, 3B, 4A, 4B, we find the attributes altitude and minimum rainfall volume have marginal influence on the model output. For the rainfall volume, the average rainfall volume and the maximum rainfall volume are more important than the minimum rainfall volume with respect of the model output.

**Model validation**:

The model was evaluated based on classification accuracy, weighted accuracy, precision, recall and F1 score. We provide an extended discussion on each of these in the supplementary text. Based on the results presented in Table 1, 2, 3, 4, the accuracy of different approaches is higher than the corresponding weighted accuracy, precision, recall and F1 score. For example, in Table 1, with respect to the method decision tree under DENV scenario, the accuracy is 0.84, the corresponding weighted accuracy, precision, recall, and F1 score are 0.76, 0.73, 0.76 and 0.74 separately.

**Figure legends**

Figure 1. Spatial clusters of CHKV, DENV, and ZIKV in Mexico

Figure 2A. Average impact on model output magnitude for CHKV all clusters, 2B. Average impact on model output magnitude for CHKK non clusters

Figure 3A. Average impact on model output magnitude for Dengue all clusters, 3B. Average impact on model output magnitude for Dengue non clusters

Figure 4A. Average impact on model output magnitude for Zika all clusters, 4B. Average impact on model output magnitude for Zika non clusters

**Supplement figures**

Figure S1: CHIKV cluster 1

Figure S2: CHIKV cluster 2

Figure S3: CHIKV cluster 3

Figure S4: DENV cluster 1

Figure S5: DENV cluster 2

Figure S6: DENV cluster 3

Figure S7: DENV cluster 4

Figure S8: DENV cluster 5

Figure S9: DENV cluster 6

Figure S10: DENV cluster 7

Figure S11: DENV cluster 8

Figure S12: DENV cluster 9

Figure S13: DENV cluster 10

Figure S14: DENV cluster 11

Figure S15: DENV cluster 12

Figure S16: ZIKV cluster 1

Figure S17: ZIKV cluster 2

Figure S18: ZIKV cluster 3

Figure S19: ZIKV cluster 4

Figure S20: ZIKV cluster 5

Figure S21: ZIKV cluster 6

**Table**

Table 1: Classification Performance of Various ML algorithms on Dengue, Zika and CHKV Virus prediction for all clusters

Table 2: Standard error of classification performance of various ML algorithms on Dengue, Zika and CHKK virus prediction for all clusters

Table 3: Classification Performance of Various ML algorithms on Dengue, Zika and CHKV Virus prediction for non-clusters

Table 4: Standard error of classification performance of various ML algorithms on Dengue, Zika and CHKV virus prediction for non-clusters

**Supplement Tables**

**Tables Supplement**

Table S1: Socio-economic and climate features impacts on DENV, ZIKV and CHIKV virus prediction at different clusters

Table S2: Prediction results of different machine learning approaches for cluster 1 in Dengue clusters

Table S3: Prediction results of different machine learning approaches for cluster 2 in Dengue clusters

Table S4: Prediction results of different machine learning approaches for cluster 3 in Dengue clusters

Table S5: Prediction results of different machine learning approaches for cluster 4 in Dengue clusters

Table S6: Prediction results of different machine learning approaches for cluster 5 in Dengue clusters

Table S7: Prediction results of different machine learning approaches for cluster 6 in Dengue clusters

Table S8: Prediction results of different machine learning approaches for cluster 7 in Dengue clusters

Table S9: Prediction results of different machine learning approaches for cluster 8 in Dengue clusters

Table S10: Prediction results of different machine learning approaches for cluster 9 in Dengue clusters

Table S11: Prediction results of different machine learning approaches for cluster 10 in Dengue clusters

Table S12: Prediction results of different machine learning approaches for cluster 11 in Dengue clusters

Table S13: Prediction results of different machine learning approaches for cluster 12 in Dengue clusters

Table S14: weights of socio-economic features and climate features by using SHAP, RDC and Pearson approaches.

Table S15: Prediction results of different machine learning approaches for cluster 1 in Zika clusters

Table S16: Prediction results of different machine learning approaches for cluster 2 in Zika clusters

Table S17: Prediction results of different machine learning approaches for cluster 3 in Zika clusters

Table S18: Prediction results of different machine learning approaches for cluster 4 in Zika clusters

Table S19: Prediction results of different machine learning approaches for cluster 5 in Zika clusters

Table S20: Prediction results of different machine learning approaches for cluster 6 in Zika clusters

Table S21: weights of socio-economic features and climate features by using SHAP, RDC and Pearson approaches.

Table S22: Prediction results of different machine learning approaches for cluster 1 in CHIKV clusters

Table S23: Prediction results of different machine learning approaches for cluster 2 in CHIKV clusters

Table S24: Prediction results of different machine learning approaches for cluster 3 in CHIKV clusters

Table S25: weights of socio-economic features and climate features by using SHAP, RDC and Pearson approaches.

**References**

1. Lubinda J, Trevino CJ, Walsh MR, et al. Environmental suitability for Aedes aegypti and Aedes albopictus and the spatial distribution of major arboviral infections in Mexico. *Parasite Epidemiol Control* 2019; **6**: e00116.

2. World Health Organization (WHO). (2020). Available at [<https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>], last accessed 12.26.2020.

3. Cattarino L, Rodriguez-Barraquer I, Imai N, Cummings DAT, Ferguson NM. Mapping global variation in dengue transmission intensity. *Sci Transl Med* 2020; **12**(528).

4. Ananth S, Shrestha N, Trevino CJ, et al. Clinical Symptoms of Arboviruses in Mexico. *Pathogens* 2020; **9**(11).

5. de Castro DB, Sampaio VS, de Albuquerque BC, et al. Dengue epidemic typology and risk factors for extensive epidemic in Amazonas state, Brazil, 2010-2011. *BMC Public Health* 2018; **18**(1): 356.

6. Mammen MP, Pimgate C, Koenraadt CJ, et al. Spatial and temporal clustering of dengue virus transmission in Thai villages. *PLoS Med* 2008; **5**(11): e205.

7. Lai WT, Chen CH, Hung H, Chen RB, Shete S, Wu CC. Recognizing spatial and temporal clustering patterns of dengue outbreaks in Taiwan. *BMC Infect Dis* 2018; **18**(1): 256.

8. Joglas Souza, Carson K. Leung, Alfredo Cuzzocrea. An Innovative Big Data Predictive Analytics Framework over Hybrid Big Data Sources with an Application for Disease Analytics. Advanced Information Networking and Applications. 2020; 1151: 669–680.

9. Fuller TL, Calvet G, Genaro Estevam C, et al. Behavioral, climatic, and environmental risk factors for Zika and Chikungunya virus infections in Rio de Janeiro, Brazil, 2015-16. *PLoS One* 2017; **12**(11): e0188002.

10. Chien LC, Sy F, Perez A. Identifying high risk areas of Zika virus infection by meteorological factors in Colombia. *BMC Infect Dis* 2019; **19**(1): 888.

11. Parham PE, Waldock J, Christophides GK, et al. Climate, environmental and socio-economic change: weighing up the balance in vector-borne disease transmission. *Philos Trans R Soc Lond B Biol Sci* 2015; **370**(1665).

12. Spiegel JM, Bonet M, Ibarra AM, Pagliccia N, Ouellette V, Yassi A. Social and environmental determinants of Aedes aegypti infestation in Central Havana: results of a case-control study nested in an integrated dengue surveillance programme in Cuba. *Trop Med Int Health* 2007; **12**(4): 503-10.

13. Colon-Gonzalez FJ, Lake IR, Bentham G. Climate variability and dengue fever in warm and humid Mexico. *Am J Trop Med Hyg* 2011; **84**(5): 757-63.

14. Johansson MA, Cummings DA, Glass GE. Multiyear climate variability and dengue--El Nino southern oscillation, weather, and dengue incidence in Puerto Rico, Mexico, and Thailand: a longitudinal data analysis. *PLoS Med* 2009; **6**(11): e1000168.

15. Moreno-Banda GL, Riojas-Rodriguez H, Hurtado-Diaz M, Danis-Lozano R, Rothenberg SJ. Effects of climatic and social factors on dengue incidence in Mexican municipalities in the state of Veracruz. *Salud Publica Mex* 2017; **59**(1): 41-52.

16. Undurraga EA, Betancourt-Cravioto M, Ramos-Castaneda J, et al. Economic and disease burden of dengue in Mexico. *PLoS Negl Trop Dis* 2015; **9**(3): e0003547.

17. Brunkard JM, Robles Lopez JL, Ramirez J, et al. Dengue fever seroprevalence and risk factors, Texas-Mexico border, 2004. *Emerg Infect Dis* 2007; **13**(10): 1477-83.

18. Watts MJ, Kotsila P, Mortyn PG, Sarto IMV, Urzi Brancati C. Influence of socio-economic, demographic and climate factors on the regional distribution of dengue in the United States and Mexico. *Int J Health Geogr* 2020; **19**(1): 44.

19. Cortes-Escamilla A, Lopez-Gatell H, Sanchez-Aleman MA, Hegewisch-Taylor J, Hernandez-Avila M, Alpuche-Aranda CM. The hidden burden of Chikungunya in central Mexico: results of a small-scale serosurvey. *Salud Publica Mex* 2018; **60**(1): 63-70.

20. Rodriguez-Morales AJ, Villamil-Gomez WE, Franco-Paredes C. The arboviral burden of disease caused by co-circulation and co-infection of dengue, chikungunya and Zika in the Americas. *Travel Med Infect Dis* 2016; **14**(3): 177-9.

21. Willey GR, Parkes HB, Palerm A, Bernstein MD, Howard F. Cline, Griffin EC. Mexico. 2019. [www.britannica.com/place/Mexico](file:///C:\Users\mh0797\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\TWQ7GAEF\www.britannica.com\place\Mexico).

22. Laureano-Rosario AE, Garcia-Rejon JE, Gomez-Carro S, Farfan-Ale JA, Muller-Karger FE. Modelling dengue fever risk in the State of Yucatan, Mexico using regional-scale satellite-derived Sea Surface Temperature. *Acta Tropica* 2017.

23. Serrano-Pinto V, Moreno-Legorreta M. Dengue hemorrhagic fever in the Northwest of Mexico: a two decade analysis. *Revista de Investigacion Clinica* 2017; (69): 152-8.

24. Jimenez Corona ME, De la Garza Barroso AL, Rodriguez Martinez JC, et al. Clinical and Epidemiological Characterization of Laboratory-Confirmed Autochthonous Cases of Zika Virus Disease in Mexico. *PLoS Curr* 2016; **8**.

25. Instituto Nacional de Estadística y Geografía [INEGI] (2020). México en Cifras. Available online: <https://www.inegi.org.mx/app/areasgeograficas/?ag=00> (accessed on 27 December 2020).

26. Jaenisch T, Tam DT, Kieu NT, et al. Clinical evaluation of dengue and identification of risk factors for severe disease: protocol for a multicentre study in 8 countries. *BMC Infect Dis* 2016; **16**: 120.

27. Saha S, Moorthi S, Pan HL, Wu X, Wang J, Nadiga S, et al. NCEP climate forecast system reanalysis (CFSR) selected hourly time-series products, January 1979 to December 2010. Research data archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. 2010. <http://rda.ucar.edu/datas> ets/ds093 .0/.

28. Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci Data. 2015;2:150066).

29. Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). (2010). "Población total, indicadores, índice y grado de rezago social, según entidad federativa, 2000, 2005, 2010 y 2015". [Base de datos en línea]. Recuperado el 1 de agosto de 2018 de (<https://www.coneval.org.mx/Medicion/Documents/Indice_Rezago_Social_2015/IRS_2000_2015_vf.zip>).

30. Kulldorff M (1997) A spatial scan statistic. Communications in Statistics - Theory and Methods 26: 1481–1496.

31. Benesty, Jacob, Jingdong Chen, Yiteng Huang, Israel Cohen. "Pearson correlation coefficient." Noise reduction in speech processing. Springer, Berlin, Heidelberg, 2009. 1-4.

32. Lopez-Paz, David, Philipp Hennig, and Bernhard Schölkopf. "The randomized dependence coefficient." Advances in neural information processing systems. 2013.

33. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems. 2017.