RAIS RESEARCH ASSOCIATION for INTERDISCIPLINARY STUDIES
OCTOBER 2017

Dengue Spread Modeling in the Absence of Sufficient Epidemiological Parameters: Comparison of SARIMA and SVM Time Series Models

Jerelyn Co[1],
Research and Graduate Assistant
MS Computer Science, Ateneo de Manila University
Jason Allan Tan[2],
Ma. Regina Justina Estuar[3], Kennedy Espina[4]
Ateneo de Manila University
jerelynco@gmail.com[1], jasonallantan@gmail.com[2]
restuar@ateneo.edu[3], kespina@ateneo.edu[4]

ABSTRACT: Dengue remains to be a major public health concern in the Philippines, claiming hundreds of lives every year (Jaymalin 2017). Given limited data for deriving necessary epidemiological parameters in developing deterministic disease models, forecasting as a means in controlling and anticipating outbreaks remains a challenge. In this study, two time series models, namely Seasonal Autoregressive Integrated Moving Average (SARIMA) and Support Vector Machine (SVM), were developed without the requirement for prior epidemiological parameters. Performances of the models in predicting dengue incidences in the Western Visayas Region of the Philippines were compared by measuring the Root Mean Square Error (RMSE) and Mean Average Error (MAE). Results showed that the models were both effective in forecasting Dengue incidences for epidemiological surveillance as validated by historical data. SARIMA model yielded average RMSE and MAE scores of 16.8187 and 11.4640, respectively. Meanwhile, SVM model achieved scores of 11.8723 and 7.7369, respectively. With the data and setup used, this study showed that SVM outperformed SARIMA in forecasting Dengue incidences. Furthermore, preliminary investigation of one-month lagged climate variables using Random Forest Regressor's feature ranking yielded rain intensity and value as top possible dengue incidence climate predictors.
KEY WORDS: SARIMA, SVM, Dengue Fever, Time Series Modeling, Feature importance.

1. Introduction

Dengue is a mosquito-borne viral disease transmitted to humans by female mosquitoes primarily of the species *Aedes aegypti* and, to a lesser extent, *Ae. albopictus*. It has been a major public health concern across the tropical and subtropical regions of the world, imposing great burden on populations, health systems, and economies in the affected countries (World Health Organization 2012). The virus is now endemic in more than 100 countries, significantly affecting Southeast Asia and Western Pacific. The reports published in 2016 showed that there were 375,000 suspected cases of Dengue in the Western Pacific Region, where 176,411 cases were reported from the Philippines (World Health Organization 2017).

The first recorded epidemic in the Southeast Asia came from Manila, Philippines in 1954. Since 1958, Dengue has been a notifiable disease in the country (Dominguez 1997, 41-47). From 2008 to 2012, the Philippines' Department of Health (DOH) reported 585,324 cases with a Case Fatality Rate (CFR) of 0.55 or 3,195 deaths (Edillo et al. 2015, 360-66). In 2011, the country was ranked 4th in the highest incidences of Dengue in the list of Association in the Southeast Asian Nations (ASEAN) member-countries (Mateo 2017). An endemic disease in the Philippines, Dengue is still considered to be one of the top priority communicable diseases in the Philippine Health Agenda 2016-2022 (Cabral 2016, 1-11).

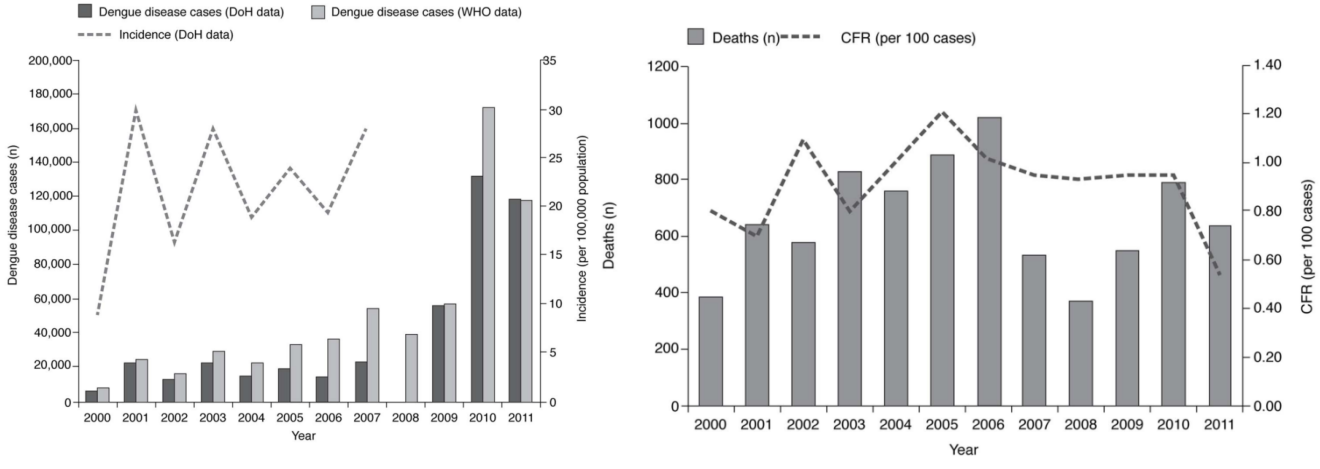


Figure 1. Reported Dengue cases (left) and deaths (right) (Bravo et al. 2011, 1-11)

Mathematical models and tools have been developed to model and forecast the disease dynamics in population (Andraud et al. 2012, 1-14; Shen 2014, 1-68; Lima et al. 2016, 1-21). However, constructing such models require prior information on the epidemiological parameters which are not easy, if not impossible, to obtain due to limitation in data and high requirement for disease understanding. Several mathematical models have been proposed in describing the dynamic behavior of dengue transmission (Johansson, Hombach, and Cummings 2011, 5860-68). Dengue, having a complex nature, is expected to have varying and several parameters to approximately capture its dynamics.

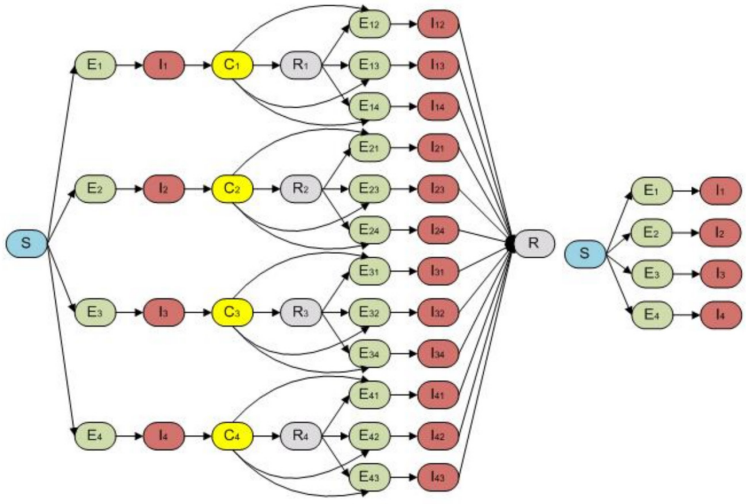


Figure 2. SpatioTemporal Epidemiological Modeler (STEM) Dengue Full Model (Kaufman 2014)

Figure 2 shows an example compartment model made for the Spatio-Temporal Epidemiological Model software. This model has 51 compartments with around 12 epidemiological parameters to search for (Kaufman 2014), which considers the immunity rate brought by Dengue's serotypes. Other examples are from Yingyun (2014) and Derouich et al. (2003), which employ less number of compartments but with around the same number of parameters (Shen 2014; Derouich, et al. 2003, 1-10). Complexity of compartment models are often attributed to their specificity. For example, a study by Shim (2016) went further to the age-dependency nature of the disease transmission (Shim 2016, 1137-47). Using these compartment models, apart from creating the compartments and formulating the Ordinary Differential

Equations (ODEs), each of the parameters must be estimated and adjusted from literature and validation.

To develop models with less parameter requirements, this paper will apply two known time series methods in forecasting real Dengue incidence data in the Western Visayas Region of the Philippines. The first method is based on statistical modeling **Seasonal Autoregressive Integrated Moving Average (SARIMA)**, and the second method is based on machine learning-based modeling **Support Vector Machine (SVM)**. Using several validation measures, this paper aims to discuss and compare the performance of the two methods. For multivariate study, feature importances of climate variables to dengue incidences will also be investigated.

2. Literature Review

Time series modeling methods have been attracting researchers over the last few decades. Particular to health, some of the applications are on physiologic studies, critical care medicine, epidemiology, environment, demographics, and health services (Zeger, Irizarry and Peng 2006, 56-79).

These methods have long contributed to epidemiologic studies of both infectious and chronic diseases for the purposes of predicting future values for surveillance (Gran et al. 2009, 221-39; Ture and Kurt 2016, 41-46). Some of the most known and used method in this area are the Autoregressive Integrated Moving Average (ARIMA) model and machine learning-based models, such as the Support Vector Machine (SVM).

2.1 Autoregressive Integrated Moving Average (ARIMA)

Statistical methods have been extensively adopted for time series forecasting. These methods assume that future values of the time series result based from past values as well as to past errors.

ARIMA, being a popular time series modeling technique, has been frequently used in disease spread forecasting. ARIMA can model historical information by using the autoregressive (AR) stage to consider past values and moving average (MA) stage to consider the current and previous residual series (Zhang et al. 2014, 1-16). It has been used to predict incidences of various diseases including hepatitis, hemorrhagic fever, and malaria (Li et al. 2012, 364-70; Ture and Kurt 2006, 41-46; Abeku et al. 2002, 851-57) .

2.2 Support Vector Machine (SVM)

With the rise of Machine Learning methods, Artificial Neural Networks have been successfully applied for modeling infectious diseases due to its nonlinear mapping ability, an advantage to the linear statistical method ARIMA. Following Neural Networks is SVM with the same property and a few more advantages in endemic forecasting (Zhang et al. 2014, 1-16).

SVM was originally developed for pattern recognition to improve the generalization property of Neural Networks. There are two main categories for SVM: Support Vector Classification (SVC) and Support Vector Regression (SVR). With its ability to solve nonlinear regression estimation problems, the SVR, shows great potential in time series forecasting (Okasha 2014; Msiza, Nelwamondo, and Marwala 2007, 638-43).

SVM works by mapping the dataset into a high-dimensional feature space and including a penalty term in the error function (Chen, and Lee 2015. 99-116; Basak, Pal and Patranabis 2017, 203-24). The idea of SVM is to separate the dataset and find the hyperplane that maximizes the margin. In regression problems, a linear learning machine learns nonlinear function in a kernel-induced feature space (Msiza, Nelwamondo, and Marwala 2007, 638-43).

3. Methodology

Several datasets were obtained from the respected departments of the country for this study. First is the Philippines Integrated Disease Surveillance and Response (PIDSR) data for years 2012 to 2016 of the Department of Health which holds the incidence count for numerous diseases including dengue. For the Western Visayas Region, PIDSR has incidence records from 122 municipalities/cities which were the subject of this study.

Another dataset is the climate data from the Advanced Science and Technology Institute (ASTI) of the Department of Science and Technology (DOST) consisting measures for wind speed and direction, rain amount and duration, temperature, pressure, and humidity for six municipalities/cities in the same region. For preliminary analysis of climate variables' contribution to dengue incidence, Random Forest Regressor's feature importance measures of the climate variables were derived by applying the regressor on one-month lagged climate variables in predicting dengue incidence data.

3.1 Models Development

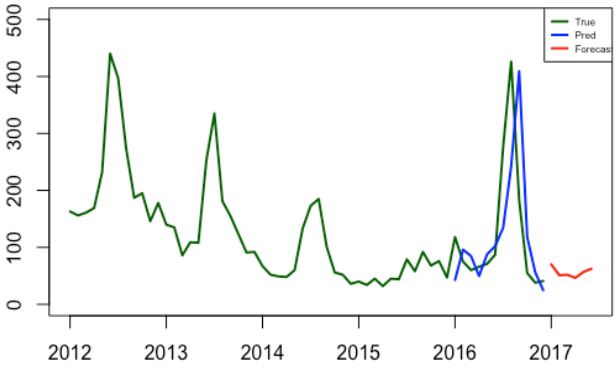
A requirement in time forecasting is to transform data into its autocorrelated form. This means that if the input time series data is $\{x_1, x_2,.., x_n\}$ and $\{x_t\}$ is the goal value for forecasting, then $\{x_{t-1}, x_{t-2},..., x_{t-p}\}$ should be the correlated value where p is the embedding dimension or the number of lags. To accommodate wide range of seasonal data, seasonal ARIMA (SARIMA) was used. Additions to ARIMA to form this seasonal ARIMA are the seasonal terms. Identifying the model orders for both nonseasonal (p, d, q) and seasonal level (P, D, Q) was based from the investigation of Akaike's Information Criterion (AIC). For SVM, since there is no standard optimal embedding value p, experiments were performed using values p={1,2,3,4,5}. Additionally, for every experiment, 5-fold cross validation was performed to determine the best combination of parameters, and C.

3.2 Forecasting

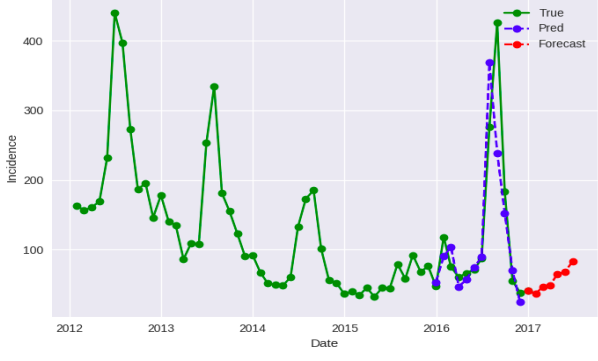
A dataset split of 80% and 20% of the total number of rows per municipality or city were used as the training and testing sets, respectively. Forecasting future incidences was based on recursive strategy, where a single model is trained to perform one-step ahead forecast which subsequently becomes the input for the next forecast. Six-month ahead forecasts were obtained from the models. To compare the two models' performances, Root Mean Square Error (RMSE) and Mean Average Error (MAE) were computed using the predicted and true values on the test set.

4. Results and Discussion

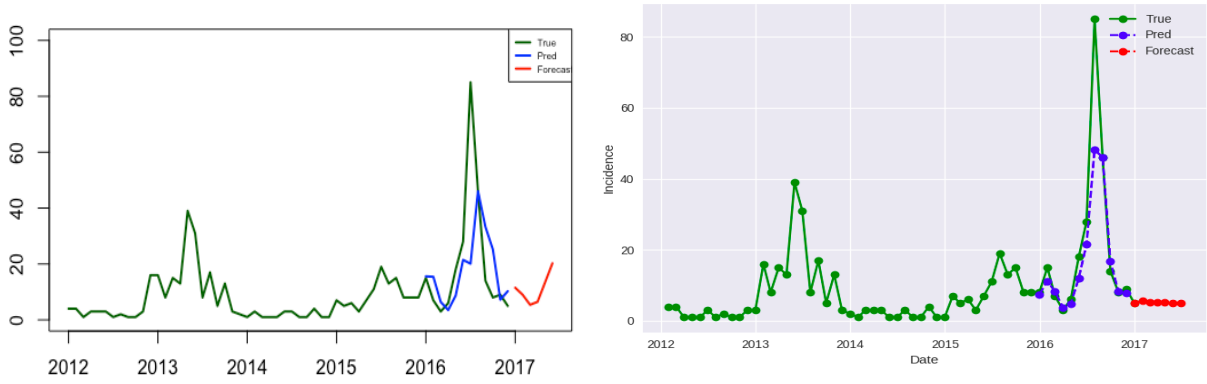
Figure 3 shows time series plots of Iloilo City and Isabela for both SARIMA and SVM models. Included in the plots are the municipality/city's true, predicted, and 6-month ahead forecast incidence value, respectively.



(a) SARIMA for Iloilo (RMSE: 111.78, MAE: 65.5)



(b) SVM for Iloilo (RMSE: 70.21, MAE: 59.24)



(c) SARIMA for Isabela (RMSE: 19.22, MAE:11.01) (d) SVM for Isabela (RMSE: 21.74, MAE:13.20)

Figure 3. Dengue Incidence plots using SARIMA and SVM

From figure 3, SVM tightly fit with the actual values for the time period with minimal variation, while SARIMA performed with slightly higher errors. For the spike nearing the end of 2016, however, SVM consistently underestimated the prediction, while SARIMA was closer to the actual values for Figure 3a. This inconsistency in performances then results to differing model of preference per municipality/city. Evident from the same figure, SVM performed better for Iloilo City with RMSE and MAE difference of 41.5 and 6.26, while SARIMA performing marginally better for Isabela with the difference of 2.52 and 2.19, respectively.

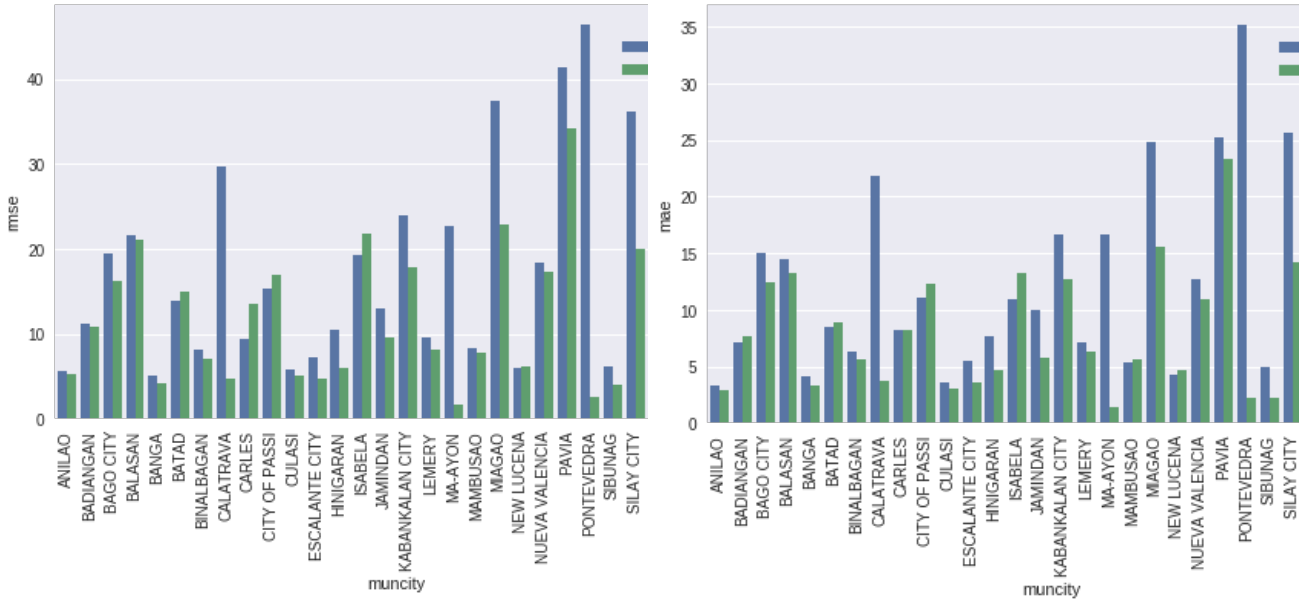


Figure 4. (left) RMSE and (right) MAE bar charts for SARIMA and SVM

Supporting the finding from figure 3, figure 4 shows variations of RMSE and MAE for the two models using 30 municipalities / cities. There are municipalities/cities where SVM model performed better, while there are some that SARIMA performed better.

Overall, computing the average scores for the two models resulted to SVM models performing better than SARIMA models for this specific application. SVM model achieved average RMSE and MAE scores of 11.8723 and 7.7369, respectively. Meanwhile, SARIMA model yielded average scores of 16.8187 and 11.4640, respectively.

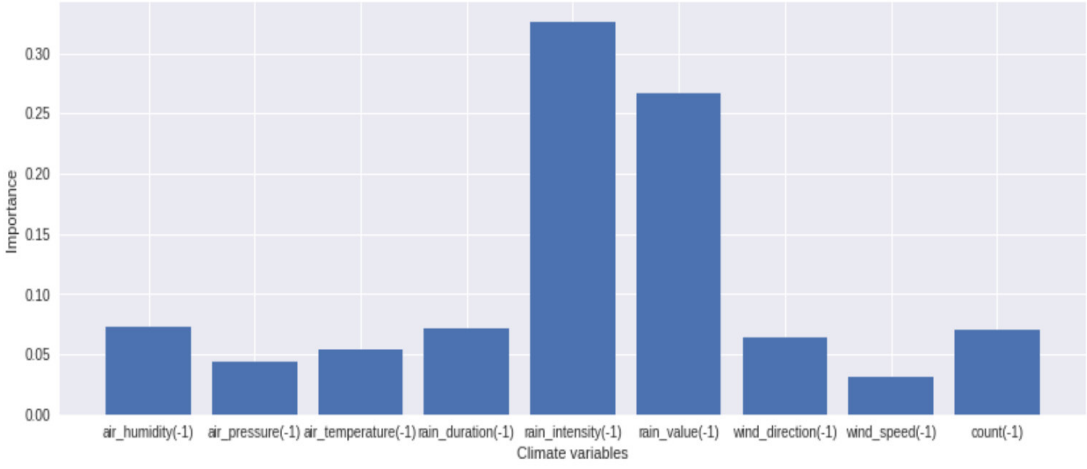


Figure 5. Feature Importances Scores of Climate Variables

Figure 5 shows the feature importances chart of the climate variables. Prominent on the chart is the high importance scores of rain intensity and rain value, following with a large gap are air humidity, rain duration, and wind direction.

Conclusion

Disease spread mathematical modeling, although a well established technique, has high development requirements such as epidemiological parameter values. The problem, however, is that these parameter values are not always easily accessible. To present alternatives, this study proved and compared the feasibility of two models, namely SVM and SARIMA, in dengue incidence forecasting. For this specific purpose,

averaging the municipality/city error scores led to SVM generally performing better than SARIMA with RMSE and MAE differences of 4.9464 and 3.7271, respectively. Results also showed that model choice may differ depending on the time series, municipality/city incidence in this case, the model was fitted to. This suggests that instead of sticking to just one of the techniques, a model with the better performance may be chosen for each of the municipality/city. For future studies, preliminary results of feature importances scores using Random Forest Regressor for climate variables showed that rain intensity and rain value are the top predictors of dengue incidence. Furthermore, a hybrid of direct-recursive forecasting strategy may be performed to include the newly predicted values in the model.

References

Abeku, Tarekegn A., Sake J. De Vlas, Gerard Borsboom, Awash Teklehaimanot, Asnakew Kebede, Dereje Olana, Gerrit J. Van Oortmarssen, and J. D F Habbema. 2002. "Forecasting Malaria Incidence from Historical Morbidity Patterns in Epidemic-Prone Areas of Ethiopia: A Simple Seasonal Adjustment Method Performs Best." *Tropical Medicine and International Health* 7 (10): 851–57. doi:10.1046/j.1365-3156.2002.00924.x.

Andraud, Mathieu, Niel Hens, Christiaan Marais, and Philippe Beutels. 2012. "Dynamic Epidemiological Models for Dengue Transmission: A Systematic Review of Structural Approaches." *PloS One* 7 (11): e49085. doi:10.1371/journal.pone.0049085.

Basak, Debasish, Srimanta Pal, and Dipak Chandra Patranabis. 2007. "Support Vector Regression." *Neuronal Information Processing - Letters and Reviews* 11 (10): 203–24. doi:10.4258/hir.2010.16.4.224.

Bravo, Lulu, Vito G. Roque, Jeremy Brett, Ruby Dizon, and Maïna L'Azou. 2014. "Epidemiology of Dengue Disease in the Philippines (2000–2011): A Systematic Literature Review." *PLoS Neglected Tropical Diseases* 8 (11). doi:10.1371/journal.pntd.0003027.

Cabral, Esperanza I. 2016. "The Philippine Health Agenda for 2016 to 2022." *Phillippine Journal of Internal Medicine* 54 (2): 1–11.

Chen, Thao-tsen, and Shie-jue Lee. 2015. "A Weighted LS-SVM Based Learning System for Time Series Forecasting." Information Sciences 299. Elsevier Inc.: 99–116. doi:10.1016/j.ins.2014.12.031.

Derouich, M, A Boutayeb, EH Twizell, HW Hethcote, EA Newton, P Reiter, L Esteva, et al. 2003. "A Model of Dengue Fever." *BioMedical Engineering OnLine* 2 (1): 4. doi:10.1186/1475-925X-2-4

Dominguez, Ma. Nerissa. 1997. "Current DF / DHF Prevention and Control Programme in the Philippines." *Dengue Bulletin* – 21: 41–47.

Edillo, Frances E., Yara A. Halasa, Francisco M. Largo, Jonathan Neil V Erasmo, Naomi B. Amoin, Maria Theresa P Alera, In Kyu Yoon, Arturo C. Alcantara, and Donald S. Shepard. 2015. "Economic Cost and Burden of Dengue in the Philippines." *American Journal of Tropical Medicine and Hygiene* 92 (2): 360–66. doi:10.4269/ajtmh.14-0139.

Gran, J.M., L. Wasmuth, E.J. Amundsen, B.H. Lindqvist, and O.O. Aalen. 2009. "Growth Rates in Epidemic Models: Application to a Model for HIV/AIDS Progression." *Statistics in Medicine* 28 (July 2006): 221–39. doi:10.1002/sim.

Jaymalin, Mayen. "More dengue deaths recorded in 2016." *Philstar.com*, 10 Feb. 2017, www.philstar.com/headlines/2017/02/10/1670758/more-dengue-deaths-recorded-2016. Accessed 13 Sept. 2017.

Johansson, Michael A., Joachim Hombach, and Derek A T Cummings. 2011. "Models of the Impact of Dengue Vaccines: A Review of Current Research and Potential Approaches." Vaccine 29 (35). Elsevier Ltd: 5860–68. doi:10.1016/j.vaccine.2011.06.042.

Kaufman, James. "Dengue Disease Transmission Model." Dengue Disease Transmission Model - Eclipsepedia. April 2014. Accessed September 10, 2017. https://wiki.eclipse.org/Dengue_Disease_Transmission_Model

Li, Qi, Na Na Guo, Zhan Ying Han, Yan Bo Zhang, Shun Xiang Qi, Yong Gang Xu, Ya Mei Wei, Xu Han, and Ying Ying Liu. 2012. "Application of an Autoregressive Integrated Moving Average Model for Predicting the Incidence of Hemorrhagic Fever with Renal Syndrome." *American Journal of Tropical Medicine and Hygiene* 87 (2): 364–70. doi:10.4269/ajtmh.2012.11-0472.

Lima, Tiago Frana Melo de, Raquel Martins Lana, Tiago Garcia de Senna Carneiro, Claudia Torres Codeco, Gabriel Souza Machado, Lucas Saraiva Ferreira, Lilliam Caesar de Castro Medeiros, and Clodoveu Augusto Davis Junior. 2016. "Dengueme: A Tool for the Modeling and Simulation of Dengue Spatiotemporal Dynamics." *International Journal of Environmental Research and Public Health* 13 (9): 1–21. doi:10.3390/ijerph13090920.

Mateo, Ibarra. 2017. "Philippines Ranks 4th in ASEAN-Wide Dengue Incidence | News | GMA News Online." Accessed August 20. http://www.gmanetwork.com/news/news/nation/223701/philippines-ranks-4th-in-asean-wide-dengue-incidence/story/.

Msiza, Ishmael S., Fulufhelo V. Nelwamondo, and Tshilidzi Marwala. 2007. "Artificial Neural Networks and Support Vector Machines for Water Demand Time Series Forecasting." *2007 IEEE International Conference on Systems, Man and Cybernetics*, 638–43. doi:10.1109/ICSMC.2007.4413591.

Okasha, M. K. (2014). Using Support Vector Machines in Financial Time Series Forecasting Using Support Vector Machines in Financial Time Series Forecasting, (January). https://doi.org/10.5923/j.statistics.20140401.03

Shen, Yingyun. 2014. "Mathematical Models of Dengue Fever and Measures to Control It." Electronic Theses, Treatises and Dissertations. http://diginole.lib.fsu.edu/etd/9093.1-

Shim, Eunha. 2016. "Dengue Dynamics and Vaccine Cost-Effectiveness Analysis in the Philippines." *American Journal of Tropical Medicine and Hygiene* 95 (5): 1137–47. doi:10.4269/ajtmh.16-0194.

Ture, Mevlut, and Imran Kurt. 2006. "Comparison of Four Different Time Series Methods to Forecast Hepatitis A Virus Infection." *Expert Systems with Applications* 31 (1): 41–46. doi:10.1016/j.eswa.2005.09.002.

World Health Organization. 2012. "Treatment, Prevention and Control Global Strategy for Dengue Prevention and Control."

World Health Organization. 2017. "Dengue and Severe Dengue." *World Health Organization*. http://www.who.int/mediacentre/factsheets/fs117/en/.

Zeger, Scott L., Rafael Irizarry, and Roger D. Peng. 2006. "On Time Series Analysis of Public Health and Biomedical Data." *Annual Review of Public Health* 27 (1): 57–79. doi:10.1146/annurev.publhealth.26.021304.144517.

Zhang, Xingyu, Tao Zhang, Alistair A Young, and Xiaosong Li. 2014. "Applications and Comparisons of Four Time Series Models in Epidemiological Surveillance Data" 9 (2): 1–16. doi:10.1371/journal.pone.0088075.