Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 7, No. 2, August 2019, pp.608-617
Available online at: http://pen.ius.edu.ba

Review on nowcasting using least absolute shrinkage selector operator (LASSO) to predict dengue occurrence in San Juan and Iquitos as part of disease surveillance system

Tang Sui Lan[1], Preethi Subramanian[2]
[1] Faculty of Computing, Engineering and Technology, Asia Pacific University of Technology & Innovation
[2] Faculty of Computing, Engineering and Technology, Asia Pacific University of Technology & Innovation

Article Info

Received Feb 8, 2019

Keyword:

Dengue
LASSO
Nowcasting
Iquitos
San Juan

ABSTRACT

Dengue which was first detected mainly in South East Asia during 1940s is now a serious public health concern across the subtropical and temperate regions of Americas, Europe and China due to the change in global climate and international travel. Hence, 3.9 billion people in 128 countries are exposed to the danger of potentially fatal dengue infection. This is a review paper of various dengue forecasting methodology to identify suitable models for predicting the disease occurrence in San Juan, Puerto Rico and Iquitos, Peru. Least Absolute Shrinkage Selector Operator (LASSO) model using climatic variables and Google Trends search terms as predictors was proposed to forecast dengue cases four weeks in advance. LASSO's flexibility in incorporating a variety of predictors and its ease of interpretation present LASSO as a compelling case against the general predictive models. Public health regulators could make use of such nowcasting model to facilitate the timing of vector control and public health campaigns along with the medical resource allocation to cope with potential dengue outbreaks.

Corresponding Author:

Dr. Preethi Subramanian,
Faculty of Computing, Engineering and Technology,
Asia Pacific University of Technology & Innovation,
Technology Park Malaysia, Bukit Jalil, Kuala Lumpur 57000, Malaysia.
Email: preethi@staffemail.apu.edu.my

1. Introduction

1.1. Dengue disease

Dengue is transmitted via Aedes mosquitoes affected by one of the four serotypes of dengue virus (Strauss, Castro, Reintjes and Torres [1], Bliman and Coelho [2]). The general symptoms of dengue include high fever, pain in the eyes, muscle and/or bone. Center for Disease Control and Prevention [3] cautioned that the potentially fatal Dengue Hemorrhagic fever (DHF) is suspected when the patient developed warning signs such as red spots or patches on the skin, bleeding from nose or gums and vomiting blood.

1.2. Predict the Next Pandemic Initiative

The "Predict the Next Pandemic Initiative" was launched by the Obama White House [4] on 5th June 2015 in response to this global health challenge. For the first time ever, several departments in the U.S. Federal

Government, Peru and Puerto Rico jointly released a series a dengue related data on disease incidence, weather, and the environment to promote data-intensive discoveries. The competition aims to predict the weekly dengue incidences at two cities in South America – San Juan of Puerto Rico and Iquitos of Peru. DrivenData [5] took over the hosting of the competition under "DengAI: Predicting Disease Spread" competition since 2017 for education purpose.

1.3. Objective

This paper aims to review the various dengue forecasting methodology in the search for suitable model in predicting the diseases occurrence in San Juan, Puerto Rico and Iquitos, Peru.

1.4. Scope

Table I lists the scope of the "Predict the Next Pandemic Initiative" / "DengAI: Predicting Disease Spread" competitions [5].

Table 1. Scope of the competition

Country	City	Temporal range for training data	Temporal range for testing data
Peru	Iquitos	2000-2009	2009 - 2013
Puerto Rico	San Juan	1990-2009	2009 – 2013

2. Dengue situation in the two cities

2.1. Iquitos, Peru

Torres, Orduna, Piña-Pozas, Vázquez-Vega and Sarti [6] noted that Iquitos has equatorial climate with seasonal dengue outbreak during the rainy season around March. The Andean area experienced heavy downpour induced by the La Niña phenomenon in early 2011 as reported by the International Federation of Red Cross and Red Crescent Societies (IFRC) [7]. Consequently, Loreto (the region where Iquitos was located) was the worst hit area in the country's 2011 most serious dengue crisis (IFRC [8]).

2.2. San Juan, Puerto Rico

Sougata, Acebedo and Chua [9] explained that San Juan reported higher incidences of dengue due to tropical monsoon climate and population density stood at a two and half times much higher than Iquitos. Puerto Rico went through an epidemic period in 2007 and 2010 but systematic review conducted by [6] did not support local weather as a crucial factor in explaining the changes in the annual case in Puerto Rico.

3. General dengue prediction models

There are two main approaches to dengue forecasting. The first mechanistic models are anchored on the vector's life cycle. Cheepsattayakorn and Cheepsattayakorn [10] highlighted the SEIR model (S: Susceptibles; E: Exposed; I: Infectious; R: Recovered/Immune) as a common framework used to illustrate the progression of dengue disease. Morin, Monaghan, Hayden, Barrera and Ernst [11] listed notable SEIR models, including Container Inhabiting Mosquito Simulation (CIMSIM), Fock/DENSiM (Dengue Simulation Model) and Dynamic Mosquito Simulation Model (DyMSiM).

The second empirical / statistical approach focuses more on improving forecast accuracy. Lee, Chung and Hwang [12] promoted artificial neural network (ANN). Specialized time series models such as Auto-Regressive Integrated Moving Average (ARIMA) have also been used by Shi et al. [13].

4. Least Absolute Selection and Shrinkage Operator (LASSO)

LASSO is a type of penalized regression which can perform feature selection to prevent over-fitting. It overcome the weaknesses of the above mentioned models such as mechanistic models' inflexibility to include a myriad of predictors and the challenging interpretability of some "black box" machine learning models.

4.1. Technical details

Kane [14] expounded that the total error of a model consists of a portion due to bias and another portion due to variance (Fig. 1). The complexity of a regression model increases as more features were introduced in the bid to reduce parameter bias. However, the trade-off between error due to bias and error due to variance as the constituents of total error resulted in prediction with low bias but high variation (Fig. 1). In other words, overfitting was resulted as the model does not generalize well beyond the training dataset.

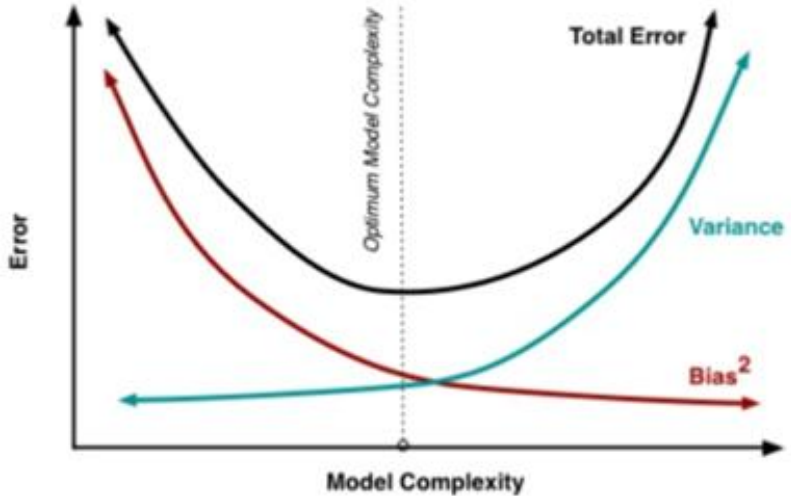

Figure 1. Bias and variance trade-off [14]

Thus a small value of Shrinkage Estimator, λ was introduced to "regularize" or shrink the magnitude of the coefficient as part of the remedy for overfitting. The unbiased Ordinary Least Squares (OLS) estimate was sacrificed for parameter estimates with less variance [14].

$$\beta'_k = \frac{1}{1+\lambda}\beta_k \quad (1)$$

where

β'_k = Parameter estimate

β_k = OLS estimate

λ = Shrinkage Estimator and λ >> 0

Ridge Regression is a variant of the new regression techniques which deploy λ to handle multi-collinearity among the predictors. However, Ridge does not yield parsimonious models as coefficients are reduced but not zerolized (Fig. 2). Lasso overcomes Ridge's weakness in feature selection by forcing the least important coefficients to zero. The most influential predictor will be identified by Lasso as its coefficient will need the largest λ to be zerolized (Fig. 3) [14].

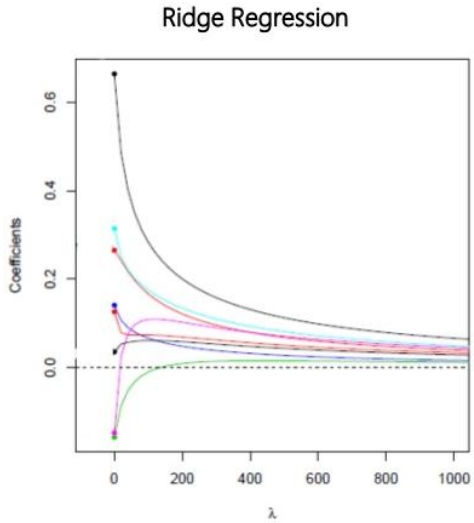

Figure 2. Ridge Trace for Ridge regression [14]

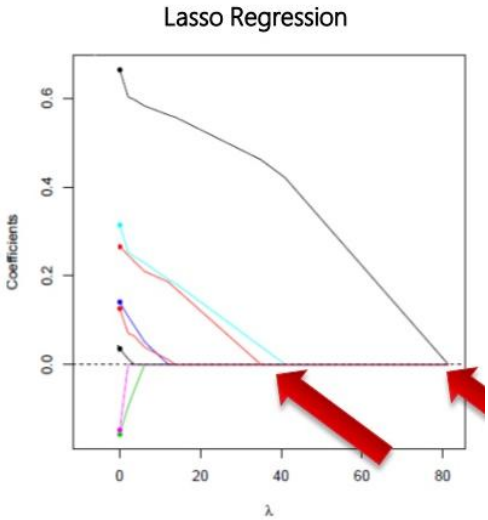

Figure 3. Ridge Trace for LASSO regression [14]

Lasso ℓ1 uses the $\|\beta\|_1$ penalty while Ridge ℓ2 uses the $\|\beta\|_2^2$ penalty [14]. Lasso characterizes a convex problem (Tibshirani [15]) and has higher efficiency when it comes to storage and computation (Jain [16]).

$$\hat{\beta}^{\text{lasso}} = \text{argmin } \|y - X\beta|_2^2 + \lambda \|\beta\|_1 \quad (2)$$

$$\hat{\beta}^{\text{ridge}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \quad (3)$$

[15] stated that while he introduced Lasso back in 1996, it was popularized only in 2002 when LARS algorithm modified forward stage wise regression to derive an effective Lasso solution.
Lasso models could have lower Mean Square Error (MSE) in comparison to Ridge as influential predictors which are correlated with the chosen predictors are zerolized by Lasso. Elastic Net, a new Ridge-Lasso hybrid tries to balance between the 2 models by allowing grouped selection of correlated features [14], [16].

$$\hat{\beta}^{\text{ridge}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \quad (4)$$

$$\hat{\beta}^{\text{lasso}} = \text{argmin } \|y - X\beta|_2^2 + \lambda \|\beta\|_1 \quad (5)$$

$$\hat{\beta}^{\text{elastic}} = \text{argmin } \|y - X\beta\|_2^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \quad (6)$$

4.2. Nowcasting with LASSO

Nowcasting is the methodology of using currently available predictors' data to extrapolate occurrence of an interested outcome in near future. Nowcasting originated from meteorology (World Meteorological Organization [17]). It has since been adopted in other fields such as finance to understand the current state of economy before the release of much delayed, official macro economy indicators from central banks (European Central Bank [18], CRAN [19]).

According to [13], correlative statistical approaches are superior to the conventional population dynamic models when it comes to dengue nowcasting. The former allows for integration with real time input data and tend to be accurate if forthcoming circumstances do not deviate from the historical pattern. Such attempts since 2011 include Autoregressive Integrated Moving Average (ARIMA), Knorr-Held two-component (K-H), Poisson multivariate regression, step-down linear regression, generalized boosted regression, and negative binomial regression [13]. Recent researches such as Chen, Chu, Chen and Cook [20] looked at LASSO as nowcasting model for its feature selection benefit along with the capability of incorporating autoregressive terms as predictors just as in the classical time series models. The final LASSO model is geared for real life implementation as it was calibrated with multiple training and validation stages [13].

The LASSO-based dengue nowcasting framework operated by Singapore's National Environment Agency (NEA) was documented by [13]. Multiple data streams were integrated on weekly basis – dengue cases, Breeding Percentage (BP, the dominance of Ae. Aegypti among the breeding sites), weather (temperature and humidity), population from midyear census, seasonality and trend decomposition by Breaks for Additive Seasonal and Trend (BFAST) algorithm in R. A total of 226 covariates (with lags up to 20 weeks) were used to construct twelve LASSO sub-models corresponding to forecast window of 1-12 weeks ahead in training-validation-testing cross validation. A final ensemble model was built to derive an overall prediction for next 12 weeks. Higher average weekly temperatures and BP were significant predictors over the short term while high levels of absolute humidity over the last month influenced medium term prediction. While the outbreak size prediction of the LASSO ensemble underperformed its epidemic timing prediction, it withstood the test of reliability whereby advance alerts were raised during the 2013 and 2014 dengue epidemic. The Singapore Ministry of Health and the Environmental Public Health Operations Department of the NEA were able to redirect hospital sources and dengue campaign accordingly.

[20] evaluated LASSO algorithm in nowcasting infectious diseases in Japan (chickenpox and Hand, foot, and mouth disease (HFMD)), Singapore (dengue and HFMD), Taiwan (dengue) and Thailand (chickenpox, dengue and malaria). Wavelet analyses was first performed to study the periodicity (amplitude, time and frequency) of each disease in each locality, out of which lags of disease incidence and climatic variables up to 26 weeks were determined as autogressive predictors. The dengue incidences underwent logarithmic transformation while the predictors where standardized prior to modelling. Multiple LASSO models were created for each disease in different countries for 1-26 weeks ahead. The optimal constraint parameters were determined using ten-fold cross validation in R. The ability of models to act as early warning system was tested by identifying the threshold corresponding to previous year's 75th percentile high incidence, alongside Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and R-squared. While high accuracy was reported for HFMD in all countries, the performance of the other models waned beyond 4 weeks' horizon. The climatory variables were found to add more value in temperate region where cyclic patterns were observed for the diseases and climatic data, in comparison to the more stable climate in equatorial region. The calendar month indicator was found to be influential in certain models, such as Thailand's chickenpox. The models are better than predicting the timing, rather than the magnitude of disease outbreak.

The predictive prowess of LASSO can oftentimes be boosted using crowd generated signals from the internet and social media such as the Google Trends.

5. Google Trends

The main conclusions of the study may be presented in a short Conclusions section, which may stand alone or form a subsection of a Discussion or Results and Discussion section.

Seung-Pyo, Hyoung and San [21] stated that Google Trends (https://trends.google.com/trends/) was launched in 2006 while the underlying data dates back to 2004. Google News Initiative Google [22] described the compilation of the internet searches from its three platforms, viz. Search, Google News, or YouTube using its

in house Knowledge Graph technology. Google Trends (GTs) analyses topical searches conduction over myriad time range, localities and languages. The public accessible version of GTs is normalized on a 0–100 scale with 100 being the maximum value of the time frame selected. Google News Initiative [23] remarked that the value of zero was used to indicate searches that below a certain volume threshold instead of absence of search. A downward trend is to be interpreted as the reducing popularity of search term relative to other topics.

Yang et al. [24] highlighted that GTs sparked great interest as a real time explanatory variable in epidemiology research on the premise that the more a person is affected by a disease, the more likely that he or she will query Google for related information. Google [25] hosted the Google Flu Trends (GFT) and Google Dengue Trends (GDT) for limited number of countries (excluding Peru and Puerto Rico) from 2008 until 2014. [24] cited a research which reported inconsistent performance of GDT across Mexico states. However, earliest publications in 2011 demonstrated that GTs data can be used to provide timely prediction of dengue cases in Bolivia, Brazil, India, Indonesia and Singapore [21].

An influenza focused ARGO (Autoregressive model with Google search queries as exogenous variables) model was extended to predict national level dengue cases in Brazil, Mexico, Singapore, Taiwan and Thailand [24]. Autoregressive models were built for each country using lags of 1-12 and 24 months under a rolling two-year window. Next, top ten dengue-related search terms for respectively country were identified and incorporated into the earlier autoregressive models to form ARGO. Lasso was tweaked to prevent the coefficient of important time lags from being zerolized. Four alternative models were constructed – a seasonal autoregressive (SAR) model; two models using official Google Dengue Trends (GDT) and GTs as explanatory variables respectively; a combination of SAR and GDT; naïve model with 1 month lag autoregressive term. ARGO consistently outperformed the rivals by RMSE, MAE, RMSPE, MAPE and Pearson correlation standards, except in Taiwan. The authors attributed the outlying Taiwan model performance to the country having minimal dengue occurrence until the 2014-2015 outbreak plus Google is not the leading search engine in Taiwan. Further tests revealed that ARGO is robust to both the sampling variation of GTs (by collecting the search terms in ten different sessions) and the scenario that reporting delay caused the latest month dengue cases to be unavailable as autoregressive term. Consequently, the authors concluded that ARGO framework is flexible to forecast dengue in multiple countries with high Google market share and seasonal dengue outbreak patterns.

Teng et al. [26] appraised Google Trends (GTs) as a source of real time, external regressor to complement time series model in predicting Zika virus disease (ZVD). The global ZVD case data from February until November 2016 were downloaded from the websites of Pan American Health Organization (PAHO) and World Health Organization (WHO). The "Interest over time" measure of Zika related search terms during the same window was downloaded from GTs website and was normalized to be 100 on the 1st day of the training dataset. The data compiled was segregated into two datasets, i.e. training (the initial 37 weeks) and testing (final 3 weeks which corresponds to Zika epidemic). Pearson Product-Moment Correlation during the training period established that strong linear correlation exists between Zika related GTs and reported ZVD cases. A simple linear regression was constructed as the first baseline model. A second baseline model of ARIMA (0, 1, 3) was fitted using Box-Jenkins method. An augmented ARIMA (0, 1, 3) was constructed using GTs as additional explanatory variable. The augmented ARIMA model registered lower Akaike Information Criterion (AIC) than the linear baseline model and the basic ARIMA model.

6. Related work on San Juan and Iquitos from

6.1. Non-ensemble models

Laureano-Rosario [27] made use of the climatic and population variables from National Oceanic and Atmospheric Administration (NOAA) to forecast the weekly dengue incident rate (per 100,000) for two population subgroups. Confirmed daily cases specific to the population at risk (those younger than 24 years old) and vulnerable population (small children and old folks) were obtained from two medical sources. Sea Surface Temperature (SST) data were extracted from NOAA Advanced Very High Resolution Radiometer (AVHRR) as an additional predictor. The authors ran two multiple linear regressions as baseline models. The ANN from Radar Pluvial Flooding Identification for Drainage System (RAPIDS) was modified with factor, a = 3 and genetic algorithm II (NSGA-II) to create two ANN models for the two population subgroups. The ANNs were run as binary models on whether there would be a potential dengue outbreak against

predetermined thresholds for each population subgroups. The authors found the non-linear ANN outperformed the Multiple Linear Regression (MLR). However, ANN models have lower accuracy rate for the "potential outbreak" period than the "no outbreak" period. Population size and date were found to have an excitatory influence (positive ANN weights) upon dengue incidents while previous cases were found to be an inhibitory influence (negative ANN weights). Maximum air temperature seemed to have inconsistent impact on the two population subgroups.

Freeze, Erraguntla and Verma [28] expanded their Data Integration and Predictive Analysis System (IPAS) for Influenza like Illness (ILI) to predict dengue cases in San Juan and Iquitos. Feature engineering was mostly centered on the weekly dengue incidences with normalization of dengue incidence to per hundred thousand of annual population; square and cube of the normalized dengue incidences as nonlinear terms; slope or the change in normalized incidence over 1-4 week horizons for trend analysis. R's caret package was used to create combined models to predict dengue incidences for 1-week and 4-weeks ahead. Features were filtered based on the significance of their coefficients in a MLR model. Shortlisted features were used in the final MLR, Support Vector Machine, Random Forests and Boosting models. Significant features found comprised of seasonal effect (represented by week number), nonlinear terms, trends and regional effect (San Juan vs. Iquitos). Random Forest and Boosting had the smallest and largest testing MSE respectively.

6.2. Ensemble models

Yamana, Kandula and Shaman [29] highlighted the need to reconcile the competing disease forecasting models. The authors created three sub-models (F1, F2 and F3) without using the climatic and population data provided in the competition. F1 involved making simplistic assumptions about the distributions of S, I, D, R0 components within the Susceptible–Infectious–Recovered (SIR) model. Next, multiple SIR models was aggregated using Ensemble Adjustment Kalman Filter (EAKF) data-assimilation method. It under-predicted pre-dengue outbreak incidences. F2 enlisted Bayesian Model Averaging (BMA) in forecasting current week's dengue incidence as the weighted sum of the previous weeks' dengue incidences from the previous seasons. It was the best peak incidences timing predictor among the three models. F3 focused on the historical likelihood whereby peak timing was proxy by Gaussian distribution while Gamma distribution was used to generate peak incidence and total incidence in a season. Its forecast deteriorated for outbreaks which did not converge to long term average. Superensemble (SE) models consisting of different combinations of F1, F2 and F3 were constructed by weighting their respective performance in the preceding weeks using BMA. SEs reported overall lower MSE than the individual models, implying the strength of SE framework in offsetting the bias within sub-models. The reliability of F2 among the individual models was carried forward into the outperformance of SE(F1, F2) among the SEs.

Johnson et al. [30] was one of the six winners of the 2015 Dengue Forecasting Challenge project. The weekly incidences, first of all, underwent a novel square-logarithmic transformation before followed by the creation of hetGP and GLM. hetGP (heteroskedistic Gaussian Process) disregard climatic variables in favor of a phenomenological approach. It overcomes heteroscedasticity by deploying a nonparametric Gaussian Process (GP) regression fitting. The model relies on four predictors (season-time, sine wave, starting level and severity indicator) to match current season to the most similar historical trajectory. GLM (Generalized Linear Model) involved negative binomial with a log link while augmenting given climatic variables (excluding humidity) with Southern Oscillation Index (SOI) and value of El Nino 1/2. The covariates underwent extensive feature engineering which entailed the combination of autoregressive (climatic lags), trend (cumulative, average, smoothing, simple linear regression), trigonometric (sine and cosine functions to capture seasonality) and transformation (logarithmic and squared) elements. Monte Carlo simulations were run to generate the forecasts followed by evaluation against the other top entries from the competitions plus [27]. The hybrid hetGP was among the top three models for San Juan while its performance was less consistent for Iquitos. The authors noticed that the parsimonious F2 model from [27] shared the advantageous historical seasonal pattern memorization and matching capability as their hetGP.

[9] performed massive feature engineering with the creation of 103 new variables from Mosquito lifecycle related lagged effects, up to 32 weeks, interaction between climatic variables and forecasted dengue cases from decomposition-based time series models. An ensemble model took the median prediction from the following sub-models, namely MLR, WMLR and CPM. Multiple Linear Regression (MLR) used overall data which enabled San Juan and Iquitos to learn from each other's dengue experience. Weighted Multiple Linear

Regression (WMLR) emphasized more on the epidemic periods. Covariate Pattern-Matching (CPM) relied on the closest matching historical window of each 19 raw climatic variables to predict the daily case changes. CPM is likely to over fit the training data as its testing dataset's MAD of 26.05% was much larger than training set's (at 7.23%). The ensemble model outperformed all sub-models, with the exception for CPM in the training set. The authors were able to deduce from the significant coefficients from MLR and WMLR models that vegetation and precipitation had negative impact while humidity and maximum temperature had positive influence on dengue occurrence.

Buczak et al. [31] was another winning team from the 2015 challenge. Their ensemble algorithm weighted the top 300 performing models from among the following by comparing their forecast errors in the past four years: (1) 1248 distinct additive seasonal Holt-Winters (HW) models which were created by tuning various parameters (with / without wavelet smoothing, periods of seasonality, ending weeks, optimization using RMSE or mean absolute percentage error (MAPE)). HW pretty much dominated the ensemble. (2) Two-dimensional Method of Analogues models which relied on historical sequence of precipitation with 2-weeks lag to forecast the weekly dengue incidences for 4-52 weeks in advance. These models performed better for San Juan's peak height and total cases. (3) Simple historical models of peak height, peak weak and total cases. They only outperformed during the start of individual forecast years, thus are the weakest of the three sub-models. The authors' team was the strongest contender for Iquitos' seasonal peak height and total cases but their San Juan predictions were mixed.

7. Conclusions

According to NOAA [32], with the exception of dengue in the study locations or nearby locations, participants are allowed to augment the model with exogenous data sources, such as social media and demography. However, most participants mostly stick to the climatic data, especially to take advantage of the satellite reanalysis data which is not found in most dengue forecasting research.

Furthermore, most of the participants approached the competition purely from historical perspective. This is evidenced by only one of the six papers published [28] which truly adopted forward looking nowcasting (forecasting the present or into the near future), which Marques-Toledo [33] emphasized is the most practical for situational awareness. [31] adopted nowcasting in one of the sub-model but not at the ensemble model level.

LASSO has been used in dengue nowcasting but none of them utilized internet searches as exogenous regressor [13], [20]. The limited dengue forecasting model which incorporated GTs on the other hand, was not built for nowcasting and did not take climatory conditions into consideration [24]. Hence, GTs can be utilized as external regressor to supplement the official DengAI climatic dataset to nowcast dengue cases in Iquitos, Peru and San Juan, Puerto Rico under a LASSO framework.

Acknowledgements

Tang Sui Lan would like to thank Dr. Imran Medi as the second marker of this research paper. Not forgetting Prof. Dr. Mandava Rajeswari, for being her teacher in LASSO hyperparameter tuning.

References

[1] Strauss,R.A., Castro,J.S., Reintjes,R. and Torres,J.R. Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case. *Int. J. Med. Inform.,* Vol. 104, pp. 26–30, 2017.

[2] Bliman,P.A., Codeco,C. and Coelho,F. From Detection to Forecasting: Big Data and Models in Epidemiology, the Example of Dengue in Rio de Janeiro. June 2016. [Online]. Available at: https://project.inria.fr/digitalforhealth/files/2016/06/PABlimanvf.pdf. [Accessed Jun. 13, 2018].

[3] Symptoms and What To Do If You Think You Have Dengue. *Center for Disease Control and Prevention*, 2012. [Online]. Available at: https://www.cdc.gov/dengue/symptoms/index.html. [Accessed Jul. 8, 2018].

[4] Back to the Future: Using Historical Dengue Data to Predict the Next Epidemic. *Obama White House* Article, 2015. [Online]. Available at: https://obamawhitehouse.archives.gov/blog/2015/06/05/back-future-using-historical-dengue-data-predict-next-epidemic. [Accessed Jul. 8, 2018].

[5] DrivenData. DengAI: Predicting Disease Spread - Challenge Summary. 2018. [Online]. Available at: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/. [Accessed Dec. 12, 2018].

[6] Torres,J.R., Orduna,T.A., Piña-Pozas,M., Vázquez-Vega,D. and Sarti,E. Epidemiological Characteristics of Dengue Disease in Latin America and in the Caribbean: A Systematic Review of the Literature. *Journal of Tropical Medicine*, Vol. 2017, March, 2017. [Online serial]. Available at: https://doi.org/10.1155/2017/8045435. [Accessed Dec. 22, 2018].

[7] Peru and Bolivia: Dengue outbreak - DREF operation n° MDR46001 21 September 2011. *The International Federation of Red Cross and Red Crescent Societies*, 2011. [Online]. Available at: https://reliefweb.int/sites/reliefweb.int/files/resources/Full_Report_2391.pdf. [Accessed Dec. 5, 2018].

[8] Peru and Bolivia: Dengue outbreak - DREF operation n° MDR46001 18 February 2011. *The International Federation of Red Cross and Red Crescent Societies*, 2011. [Online]. Available at: https://reliefweb.int/sites/reliefweb.int/files/resources/1507AAFA08C7F39E8525783B00764030-Full_Report.pdf. [Accessed Dec. 5, 2018].

[9] Sougata,D., Acebedo,C.M.L. and Chua,M.C.H. An ensemble prediction approach to weekly Dengue cases forecasting based on climatic and terrain conditions. *J. Health. Soc. Sci.*, Vol. 2, no. 3, pp. 257-272, 2017.

[10] Cheepsattayakorn,A. and Cheepsattayakorn,R. Climate Changes and Human Infectious Diseases. *ECMI.*, Vol. 14, no. 6, pp. 299-311, 2018.

[11] Morin,C.W., Monaghan,A.J., Hayden,M.H., Barrera,R. and Ernst,K. Meteorologically Driven Simulations of Dengue Epidemics in San Juan, PR. *PLOS Negl. Trop. Dis.*, August, 2015. [Online serial]. Available at: http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0004002. [Accessed Jun. 13, 2018].

[12] Lee,K.Y., Chung,N. and Hwang,S.T. Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas. *Ecol. Inform.*, vol. 36, pp. 172–180, 2016.

[13] Shi,Y., Liu,X., Kok,S.Y., Rajarethinam,J., Liang,S., Yap,G., Chong,C.S., Lee,K.S., Tan,S.S., Chin,C.K., Lo,A., Kong,W., Ng, L.C. and Cook,A.R. Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts and Policy Decision Support in Singapore. *Environ. Health. Perspect*., vol. 124, no. 9. pp. 1369-1375, 2016.

[14] Kane,D. Data Science - Part XII - Ridge Regression, LASSO, and Elastic Nets. 2015. [Online] Available at: https://www.slideshare.net/DerekKane/data-science-part-xii-ridge-regression-lasso-and-elastic-nets. [Accessed Jul. 21, 2018].

[15] Tibshirani,R. Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc. Series. B Stat. Methodol.*, vol. 73, no. 3, pp. 273–282, 2011.

[16] Jain,S. A comprehensive beginners guide for Linear, Ridge and Lasso Regression. *Analyticsvidhya.com*, Jun. 22, 2017. [Online] Available at: https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/. [Accessed Jul. 21, 2015].

[17] Nowcasting. *World Meteorological Organization*, 2017. [Online]. Available at: http://www.wmo.int/pages/prog/amp/pwsp/Nowcasting.htm. [Accessed Nov. 9, 2018].

[18] Now-casting and the real-time data flow, Working Paper Series No. 1564. *European Central Bank*, Jul. 2013. [Online]. Available at: https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1564.pdf. [Accessed Nov. 11, 2018].

[19] Comprehensive R Archive Network, nowcasting: Nowcast Analysis and Create Real-Time Data Basis. cran.r-project.org, Nov. 27, 2018 [Online]. Available at: https://cran.r-project.org/web/packages/nowcasting/index.html. [Accessed Nov. 9, 2018].

[20] Chen,Y., Chu,C.W., Chen,M.I.C. and A.R.. Cook The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. *J. Biomed. Inform.*, vol. 81, pp. 16–30, 2018.

[21] Seung-Pyo,J., Hyoung,S.Y. and San,C. Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *TF&SC.*, vol. 130. pp. 69–87, 2018.

[22] Google Trends: See what's trending across Google Search, Google News and YouTube. *Google News Initiative.* storage.googleapis.com, 2018. [Online]. Available at: https://storage.googleapis.com/gweb-news-initiative-training.appspot.com/upload/GO802_NewsInitiativeLessons_Fundamentals-L03-GoogleTrends.pdf. [Accessed Nov. 10, 2018].

[23] Google Trends: Understanding the data. *Google News Initiative*. storage.googleapis.com, 2018. [Online]. Available at: https:// storage.googleapis.com/ gweb-news-initiative-training.appspot.com/ upload/GO802_NewsInitiativeLessons_Fundamentals-L04-GoogleTrends.pdf. [Accessed Nov. 10, 2018].

[24] Yang,S., Kou,S.C., Lu,F., Brownstein,J.S., Brooke,N. and Santillana,M. Advances in using Internet searches to track dengue. *PLOS Comput. Biol.*, July, 2017. [Online serial]. Available at: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005607. [Accessed Jul. 9, 2018].

[25] Thank you for stopping by. *Google Trends* google.org, 2018. [Online]. Available at: http://www.google.org/flutrends/about/. [Accessed Jul. 9, 2018].

[26] Teng,Y., Bi,D., Xie,G., Jin,Y., Huang,Y., Lin,B., An,X., Feng,D. and Tong,Y. Dynamic Forecasting of Zika Epidemics Using Google Trends. *PLOS One.*, January, 2017. [Online serial]. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0165085. [Accessed Jul. 11, 2018].

[27] Laureano-Rosario,A.E., Duncan,A.P., Mendez-Lazaro,P.A., Garcia-Rejon,J.E., Gomez-Carro,S., Farfan-Ale,J., Savic,D.A. and Muller-Karger,F.E. Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico. *Trop. Med. Infect. Dis.*, vol. 3, no. 1, January, 2018. [Online serial]. Available at: https://doi.org/10.3390/tropicalmed3010005. [Accessed Dec. 22, 2018].

[28] Freeze,F., Erraguntla,M. and Verma,A. Data Integration and Predictive Analysis System for Disease Prophylaxis: Incorporating Dengue Fever Forecasts. *Proceedings of the 51st Hawaii International Conference on System Sciences*, pp. 913-922, Jan 2-6, 2018.

[29] Yamana,T.K., Kandula,S. and Shaman,J. Superensemble forecasts of dengue outbreaks. *J. Royal Soc. Interface*., vol. 13, no. 123, October, 2016. [Online serial]. Available at: https://doi.org/10.1098/rsif.2016.0410. [Accessed Dec. 22, 2018].

[30] Johnson,L.R., Gramacy,R.B., Cohen,J., Mordecai,E., Murdock,C., Rohr,J., Ryan,S.J., Stewart-Ibarra,A.M. and Weikel,D. Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: a dengue case study. *Ann. Appl. Stat.*, vol. 12, no. 1, pp. 27-66, 2018.

[31] Buczak,A.L., Baugher,B., Moniz,L.J., Bagley,T., Babin,S.M. and Guven,E. Ensemble method for dengue prediction. *PLOS One.*, vol. 13, no. 1, January, 2018. [Online serial]. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0189988. [Accessed Jul. 8, 2018].

[32] Dengue Forecasting. *National Oceanic and Atmospheric Administration,* 2018. [Online]. Available at: http://dengueforecasting.noaa.gov/. [Accessed Jul. 8, 2018].

[33] Marques-Toledo,C.A., Degener,C.M., Vinhal,L., Coelho,G., Meira,W., Codeço,C.T. and Teixeira,M.M. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. *PLOS Negl. Trop. Dis.*, July, 2017 [Online serial]. Available at: http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005729. [Accessed Jul. 10, 2018].