J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

REVIEW ARTICLE

A Robust and Non-parametric Model for Prediction of Dengue Incidence

Atlanta Chakraborty[1] and Vijay Chandru[2]*

Abstract | Disease surveillance is essential not only for the prior detection of outbreaks, but also for monitoring trends of the disease in the long run. In this paper, we aim to build a tactical model for the surveillance of dengue, in particular. Most existing models for dengue prediction exploit its known relationships between climate and socio-demographic factors with the incidence counts; however, they are not flexible enough to capture the steep and sudden rise and fall of the incidence counts. This has been the motivation for the methodology used in our paper. We build a non-parametric, flexible, Gaussian process (GP) regression model that relies on past dengue incidence counts and climate covariates, and show that the GP model performs accurately, in comparison with the other existing methodologies, thus proving to be a good tactical and robust model for health authorities to plan their course of action.

Keywords: *Epidemic, Dengue, Non-parametric, Gaussian process, Covariance, Kernel, Robust, Tactical model*

1 Introduction

Dengue is a fast emerging pandemic-prone viral disease transmitted by *Aedes aegypti* and *Aedes albopictus* mosquitos. According to the World Health Organisation (WHO), each year, an estimated 390 million dengue infections occur all around the world. Cases across the Americas, South-East Asia and Western Pacific exceeded 1.2 million in 2008 and over 3.2 million in 2015[23]. Several precautionary measures include vector control tools, like controlling mosquito populations; however, implementation is a major challenge and effective dengue prevention is rarely achieved, specially in developing countries. Often, it is the emergency vector control operation that is usually applied when an outbreak occurs, such as insecticide fogging.

Accurate forecasts of incidence cases, or infected individuals are key to planning and resource allocation of dengue vaccines, medical centres, etc. Previous attempts to model dengue have made use of relatively simple models, such as generalised linear model and ARIMA, exploiting the relationship with other environmental variables[4, 7, 14]. However, most of the times, disease dynamics are not well understood and such models may fail to capture that [11]. Dengue is closely related to the seasonal changes, rainfall and humidity. Our model is trained on historical incidence data, mean surface temperature, humidity and rainfall, and makes use of a Bayesian non-parametric modelling framework, Gaussian processes (GP) that allows for flexibility in the model, thus being able to forecast the sudden peak increase of the incidence counts.

2 Related Work

A study and systematic review of existing dengue modelling methods, conducted by Louis et al.[11], has been instrumental in providing us with an overview of current modelling efforts and their limitations. The study enlists a wide range of predictors that were used to create dengue risk maps, such as socioeconomic and demographic data, climatic and environmental data, remote sensing and entomological data.

Several efforts[4, 6, 7, 14] have made use of parametric models such as logistic regression models, multinomial models and generalised linear models with climatic covariates as inputs. Climatic

Non-parametric: A parametric model simplifies the learning model to a known functional form, but a non-parametric model does not make any assumptions on the learning function and can hence learn any function from the training data.

Vector: In epidemiology, vectors are organisms that transmit infectious pathogens from animals to humans or between humans.

ARIMA: ARIMA models predict future values of a time series based on its past values, i.e., its lags.

[1] *Institute of Operations Research and Analytics, National University of Singapore, 3 Research Link, Innovation 4.0, #04-01, Singapore 117602, Singapore.*
[2] *Center for BioSystems Science and Engineering, Indian Institute of Science, Bangalore, India.*
**atlanta@u.nus.edu*

data have been found to be particularly useful for the generation of risk maps[2]. Additionally, other factors, such as human mobility or housing conditions, are also likely to be linked to the occurrence of dengue cases[8]. In contrast to the fields of malaria and other vector-borne diseases, the study shows that dengue is particularly challenging due to the high number of non-detectable breeding sites[5, 9, 12, 20]. There have been models making use of entomological data[1, 3, 16, 19–22] studying the link between several vector aspects (like larvae abundance, ovi-trapping) and dengue cases; however, the exact nature of the relationship remains unknown. Such surveys are not only labor-intensive and costly, but they also yield spurious results that are not useful for prediction[3].

The weakness of the current dengue prediction efforts originates from the fact that dengue is highly dynamic and multifactorial. Factors such as host immunity and genetic diversity of circulating viruses also play an important role, but they are difficult and expensive to track. They continue to pose challenges and limit the ability to produce accurate and effective risk maps and models, thus failing to support the development of an early warning system. Many epidemiological models have been developed and have gained importance in the last decade; however, they cannot be used in the public health context due to their complexity and the extensive need for input data. Additionally, most models produce an average forecast of the numbers in the long run, instead of being able to predict the immediate rise and fall of the incidence counts. On account of this, we aim to build a robust but easily implementable model that would depend only on climatic variables and past historical data. The GP model has an added advantage of generalising the model to include several other factors that affect dengue, such as human mobility and vector data.

3 The Data

We have chosen Singapore as our case study due to the free availability of weekly dengue incidence counts. Data for the years 2005–2017 were downloaded from the Ministry of Health, Singapore bulletin[13]. Rainfall, relative humidity and surface air temperature data were also downloaded from the National Environment Agency website[15]. The years 2005–2016 were used as a training set, whereas the year 2017 was used as a test set. The incidence data are final counts, i.e. the total number of cases in each week. To avoid overfitting or underfitting, we have implemented the process of k-fold cross-validation ($k = 10$) to help us understand the performance of the model and select an appropriate one.

Figure 1 shows Singapore incidence counts across 2005–2017. It clearly depicts a yearly cycle. Not only do the data have a mean response which varies with time, but also the variability of the incidence counts is unequal across the months. The dengue count is a heteroskedastic variable when predicted by the month number.

Heteroskedastic: Heteroskedastic data is data with unequal variability across a set of predictor variables, here across time periods.

Prior: In Bayesian inference, the prior expresses one's belief before observing the data.

Posterior: The posterior distribution is the distribution of the parameters after taking into account the observed data.

Cross-validation: Cross validation tests the model's ability to predict unknown new data to avoid problems of over-fitting or selection bias.

4 Methodology

4.1 Gaussian Process Modelling

This paper proposes to model dengue incidence with Gaussian processes (GP), a non-parametric modelling framework[17], for the purpose of getting an added flexibility and making accurate predictions of the peak season, as it falls and rises. One can think of GP as defining a distribution over functions, and inference takes place directly in the space of functions.

A GP is completely specified by its mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}')$ of a process $f(\mathbf{x})$, and we write $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. We first assume our model to be of the form $y = f(\mathbf{x}) + \epsilon$, where ϵ is additive and independent identically distributed Gaussian noise with variance σ_n^2.

From our training data, we know $\{(\mathbf{x_i}, y_i)|i = 1, \ldots, n\}$, where n is the total number of observations. The joint distribution of the training outputs, y, and the test outputs f_* according to the prior is

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^* \mathbf{I} & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right).$$

If there are n_* test points. then $K(\mathbf{X}, \mathbf{X}_*)$ denotes the $n \times n_*$ matrix of the covariances evaluated at all pairs of training (X) and test points (X^*), and similarly for the other entries. To get the posterior distribution over functions, we need to restrict this joint prior distribution to contain only those functions which agree with the observed data points. The key predictive equations for Gaussian process regression are

$$f_*|X, y, X_* \sim \mathcal{N}(\bar{f_*}, \text{cov}(f_*)), \tag{1}$$

$$\bar{f_*} = K(X_*, X)[K(X, X) + \sigma_n^2 \mathbf{I}]^{-1} y, \tag{2}$$

$$\text{cov}(f_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 \mathbf{I}]^{-1} K(X, X_*). \tag{3}$$

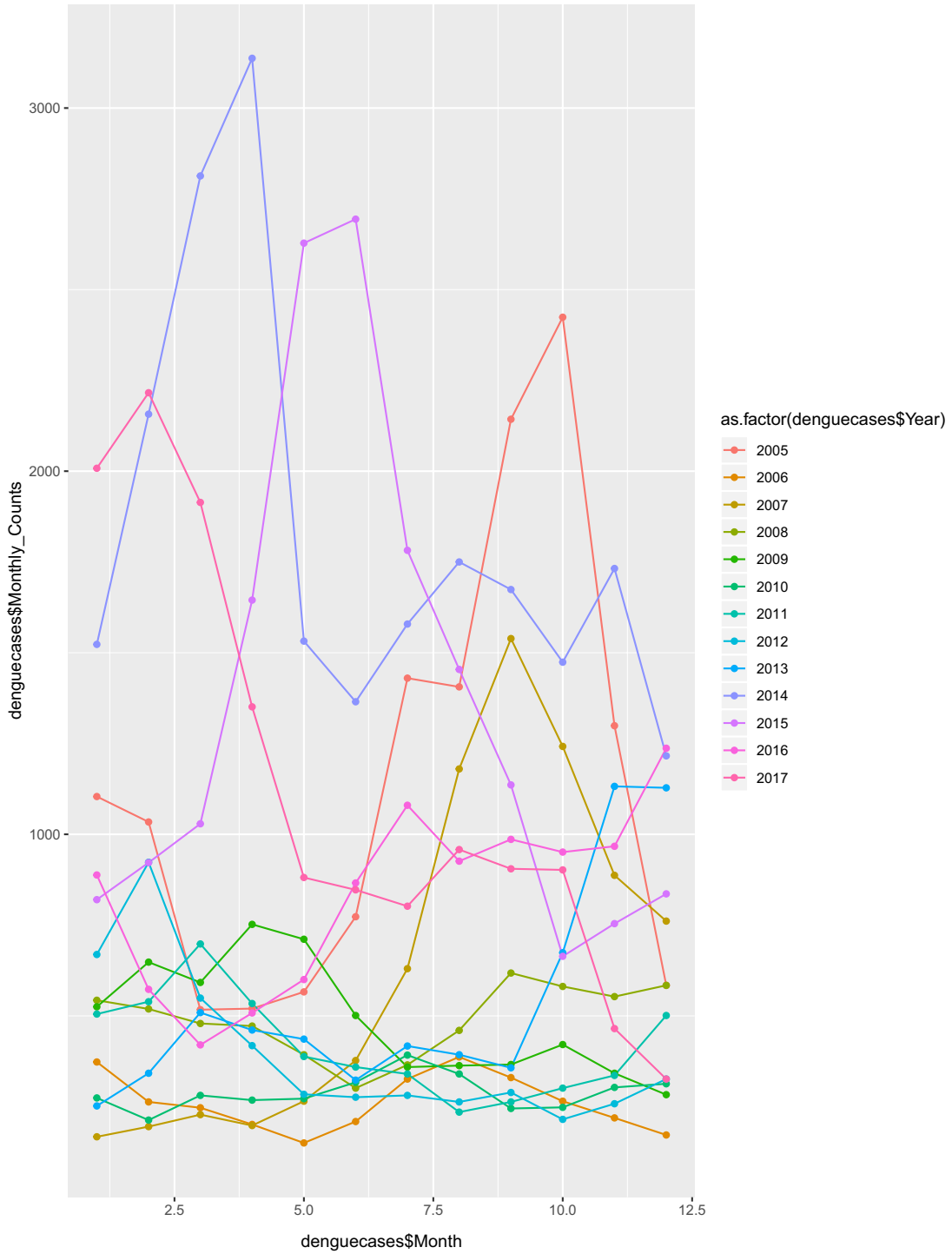

Figure 1: Monthly incidence of dengue vs month across years 2005–2017.

By this definition, GPs allow us to obtain the exact predictive distribution through a closed-form expression. They are also flexible, since one can use any positive semi-definite kernel as the covariance function K as a measure of similarity between points, providing rich insights about the dependencies between them.

Under the Gaussian process model, the prior is Gaussian, $f|X \sim \mathcal{N}(0, K)$, or

$$\log p(f|X) = \frac{-1}{2} f^T K^{-1} f - \frac{1}{2} \log |K| - \frac{n}{2} \log 2\pi \quad (4)$$

Positive semi-definite kernel: A positive semi-definite matrix has non-negative eigen values.

and the *likelihood* is a factorised Gaussian $y|f \sim \mathcal{N}(f, \sigma_n^2\mathbf{I})$. We thus arrive at the log marginal likelihood as

$$\log p(y|X) = \frac{-1}{2} y^T (K + \sigma_n^2\mathbf{I})^{-1} y - \frac{1}{2}\log \times |K + \sigma_n^2\mathbf{I}| - \frac{n}{2}\log 2\pi. \tag{5}$$

Marginal likelihood: Likelihood is the distribution of the observed data. The marginal likelihood is the distribution of the observed data marginalized (summed) over the parameters.

We estimate the hyper-parameters of K by maximising the marginal likelihood (or minimising the negative log likelihood). We can use several gradient-based optimisers, since it is necessary to compute the partial derivatives of the marginal likelihood w.r.t. the hyper-parameters. For our purpose, we use the "BFGS" method.

Hyper-parameters: Model hyper-parameters differ from model parameters. Whilst the model parameter is a part of the learning model and is estimated from the data automatically, the model hyper-parameter is external to the model and is set manually, can be tuned for a given predictive model.

We apply a logarithmic (one plus) transformation on the response variable and model this transformation as a GP. This is done to ensure that the largest variances are stabilised. The main task in modelling via GPs is to define an appropriate covariance structure. We assume a zero-mean GP by centering the response variable about its mean.

4.2 Defining the Covariance Function

Covariance functions encode our assumptions about the function which we wish to learn. It is a basic assumption that input points which are "close" to each other are likely to have similar target values y. Based on this, training points that are close to a test point should provide information about the prediction at that point. It is the covariance function that defines this nearness or similarity.

A complex covariance function is derived by combining several different kinds of simple covariance functions. The covariance structure imposed by the GP prior should reflect what we expect from the data. We make use of standard kernels defined in the GP literature[17]. Our goal is to model the transformed incidence counts as a function of $x_i = (x_1, x_2, x_3, x_4)_i$, i.e. the ith observation and its corresponding month number, total monthly rainfall, mean relative humidity and mean surface air temperature, respectively.

Kernels: A kernel (also called a covariance function) describes the joint variability between two variables. For rationale on the different kernels used, the reader is referred to[15].

To enforce the assumption that the test input is highly correlated with its pre-ceding inputs, we use a 5/2 Matern kernel which is defined as

$$k_1(x_i, x_j) = \sigma_1^2\left(1 + \frac{\sqrt{5}\Delta x}{l_1} + \frac{5\Delta x^2}{3l_1^2}\right) \times \exp\left(\frac{-\sqrt{5}\Delta x}{l_1}\right), \tag{6}$$

where $\Delta x = |x_i - x_j|$ is the absolute distance between the inputs. Its hyper-parameters, σ_1 and l_1 are used to control the strength of correlation signal and the span of time that should correlate, respectively.

We use a second component to exploit the periodicity observed in dengue incidence, while still giving more importance to closer periods of time.

$$k_2(x_i, x_j) = k_{21}(x_i, x_j) \times k_{22}(x_i, x_j), \tag{7}$$

$$k_{21}(x_i, x_j) = \sigma_2^2 \exp\left(\frac{-\Delta x^2}{2l_2^2}\right), \tag{8}$$

$$k_{22}(x_i, x_j) = \exp\left(-2\sin^2\left(\frac{\pi\Delta x}{p}\right)/l_{\text{per}}^2\right). \tag{9}$$

k_{21} is a squared-exponential kernel (also called radial basis function kernel) and k_{22} is a periodic kernel. The hyper-parameters of k_{21}–l_2 and σ_2 are used to control the number of months that should impact the incidence and strength of the correlation signal, respectively. p and l_{per}, the hyper-parameters of k_{22} are used to control the periodicity and length scale of the signal, respectively.

Next, we model the small irregularities with a rational quadratic term. The rational quadratic kernel allows us to model the data varying at multiple scales.

$$k_3(x_i, x_j) = \sigma_3\left(1 + \frac{\Delta x^2}{2\alpha l_3^2}\right)^{-\alpha}. \tag{10}$$

σ_3 is the magnitude, $\alpha > 0$ is the scale parameter and l_3 is the characteristic length scale.

Finally, we specify the noise model as the sum of a squared exponential contribution and an independent component. Noise in the series could be due to measurement inaccuracies. It could also be due to the changes in weather phenomena every year, hence we assume that there is a little amount of correlation in time.

$$k_4(x_i, x_j) = \sigma_f^2 \exp\left(\frac{-\Delta x^2}{2l_4^2}\right) + \sigma_n^2\delta_{x_i x_j}, \tag{11}$$

where σ_f is the signal variance, l_4 is its length scale and σ_n is the magnitude of the independent noise component.

The final covariance function is

$$k(x_i, x_j) = k_1(x_i, x_j) + k_2(x_i, x_j) + k_3(x_i, x_j) + k_4(x_i, x_j) \quad (12)$$

with 12 hyper-parameters.

Note that most of the above defined covariance functions are stationary, i.e. invariant to translations in the input space. Sampson and Guttorp[18] introduced the method of warping in 1992, which allows us to introduce an arbitrary non-linear map $u(x)$ of the input space x, and then use stationary covariance functions in the $u(x)$ space. This is yet another reason for the logarithmic transformation of the incidence counts.

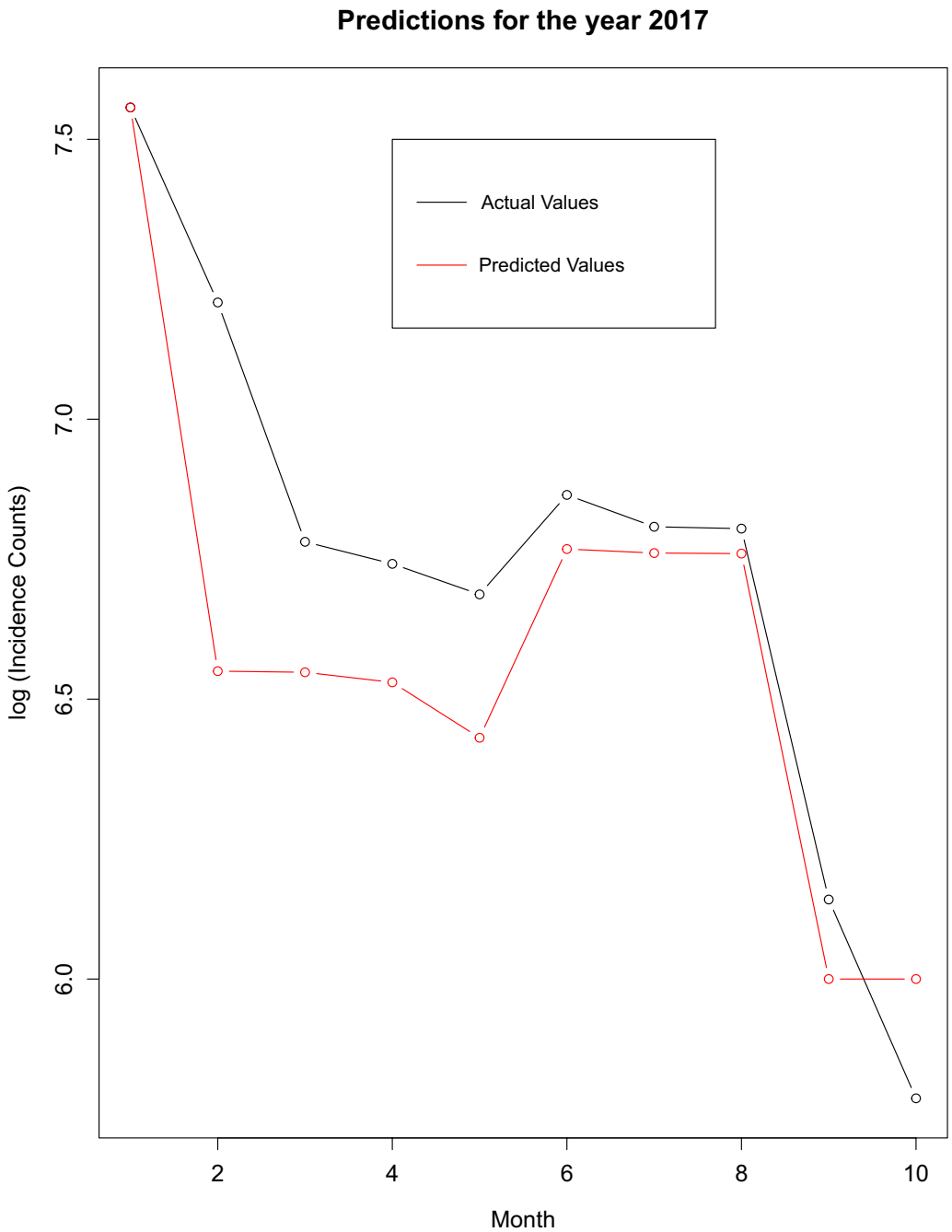

Figure 2: Gaussian model on the test set.

The code has been run on R, mainly using the `GauPro` package. All the above-mentioned kernels are imported from the `kergp` package. The training data are then fit and the marginal likelihood is optimised using the "BFGS" algorithm.

5 Results and Discussion

We compare our GP model with three different existing methodologies— time series forecasting (ARIMA), generalised additive models (GAM) and predictions from random forests (RF), on the basis of two different metrics-root mean squared error (RMSE) and mean absolute deviation (MAD). The performance across various methods is reported as follows, for both in-sample and out-of-sample forecasts.

Model	Training RSME	Test RMSE	Training MAD	Test MAD
GAM	0.608	0.682	0.623	0.883
Time series	0.521	0.562	0.493	0.512
Random forest	0.676	0.719	0.713	1.101
GP	8.3e−07	0.260	9.63e−07	0.262

Overfitting is taken care of by cross-validation. As can be seen from the above performance metrics, GP is very accurate for forecasting and easily implementable. It also has room for adding more covariates to the model. For the out-of-sample forecasts, Fig. 2 shows the predictions for the year 2017.

6 Conclusion and Future Work

As we have seen in the last section, the model fit by Gaussian process serves as a good tactical model. This is due to its non-parametric nature and its flexibility, thus being able to automatically adapt to different scenarios. The other advantage of this model is the nature of the input, incidence numbers correlated with climate variables. In the future, we would like to investigate the role of human and vector factors in helping us forecast dengue incidence in a public health context. The GP model can accommodate such factors by introducing kernel functions based on human and vector interactions and add it to the already defined kernel function in this paper.

The GP model provides a sufficient window for health authorities to be aware of the incoming dengue counts and hence carefully plan and take necessary actions. Its easy implementation can act as a very accurate early warning system when implemented on a weekly basis. To make the model functional on a weekly basis, one may consider data on a weekly scale spatially and consider resource allocation and facility problems to effectively implement an operational model.

Publisher's Note

Received: 17 August 2020 Accepted: 14 September 2020
Published online: 17 October 2020

References

1. Adams B, Kapan DD (2003) Man bites mosquito: understanding the contribution of human movement to vector-borne disease dynamics. PLoS ONE 4:e6763
2. Albinati J, Wagner M, Pappa GL (2016) An accurate Gaussian process-based early warning system for dengue fever [stat.AP]. arXiv: 1608.03343v1
3. Banu S, Hu W, Hurst C, Tong S (2011) Dengue transmission in the Asia–Pacific region: impact of climate change and socio-environmental factors. Trop Med Int Health 16:598–607
4. Chen SC, Hsieh MH (2012) Modeling the transmission dynamics of dengue fever: implications of temperature effects. Sci Total Environ 431:385–391
5. Dambach P, Machault V, Lacaux JP, Vignolles C, Sie A, Sauerborn R (2012) Utilization of combined remote sensing techniques to detect environmental variables influencing malaria vector densities in rural West Africa. Int J Health Geogr 11(1476–072X (Electronic)):8–20
6. Dayama P, Kameshwaran S (2013) Predicting the dengue incidence in singapore using univariate time series models. In: Annual symposium proceedings. AMIA, pp 285–292
7. Gharbi M, Quenel P, Gustave J, Cassadou S, Ruche GL, Girdary L, Marrama L (2011) Time series analysis of dengue incidence in Guade-loupe, French West Indies: forecasting models using climate variables as predictors. BMC Infect Dis 11(1):166
8. Hii YL, Zhu H, Ng N, Ng LC, Rocklov J (2012) Forecast of dengue incidence using temperature and rainfall. Plos Neglect Trop Dis 6:e1908
9. Jancloes M, Thomson M, Costa MM, Hewitt C, Corvalan C, Dinku T, Lowe R, Hayden M (2014) Climate services to improve public health. Int J Environ Res Public Health 11:4555–4559
10. Johnson LR, Gramacy RB, Cohen J, Mordecai E, Murdock C, Rohr J, Ryan SJ, Stewart-Ibarra AM, Weikel D (2017) Phenomenological forecasting of dengue incidence using heteroskedastic Gaussian processes: a dengue study.arXiv 2017

11. Louis VR, Phalkey R, Horstick O, Ratanawong P, Wilder-Smith A, Tozan Y, Dambach P (2014) Modeling tools for dengue risk mapping—a systematic review. Int J Health Geogr 13(1):1–15
12. Machault V, Vignolles C, Pagès F, Gadiaga L, Tourre YM, Gaye A, Sokhna C, Trape J-F, Lacaux J-P, Rogier C (2012) Risk mapping of Anopheles gambiae s.l. densities using remotely-sensed environmental and meteorological data in an urban area: Dakar, Senegal. PLoS One 7:e50674
13. (2018) Ministry of Health Singapore. https://www.moh.gov.sg/resources-statistics/infectious-disease-statistics/2018/weekly-infectious-diseases-bulletin
14. Naish S, Dale P, Mackenzie JS, Mengersen K, Tong S (2014) Climate change and dengue: a critical and systematic review of quantitative modelling approaches. BMC Infect Dis 14(1):167
15. National Environment Agency, Singapore. http://www.weather.gov.sg/climate-historical-daily
16. Nevai AL, Soewono E (2013) A model for the spatial transmission of dengue with daily movement between villages and a city. Math Med Biol 30:dqt002
17. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, London
18. Sampson PD, Guttorp P (1992) Nonparametric estimation of nonstationary spatial covariance structure. J Am Stat Assoc 87(417):108–119
19. Syed M, Saleem T, Syeda U-R, Habib M, Zahid R, Bashir A, Rabbani M, Khalid M, Iqbal A, Rao EZ, Shujja-ur-Rehman Saleem S (2010) Knowledge, attitudes and practices regarding dengue fever among adults of high and low socioeconomic groups. J Pak Med Assoc 3:243–7
20. Thomas CJ, Lindsay SW (2000) Local-scale variation in malaria infection amongst rural Gambian children estimated by satellite remote sensing. Trans R Soc Trop Med Hyg 94:159–163
21. Troyo A, Fuller DO, Calderón-Arguedas O, Solano ME, Beier JC (2009) Urban structure and dengue fever in Puntarenas, Costa Rica. Singap J Trop Geogr 30:265–282
22. Wilder-Smith A, Gubler DJ (2008) Geographic expansion of dengue: the impact of international travel. Med Clin N Am 92:1377–1390
23. (2018) World Health Organization. https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue

Atlanta Chakraborty is a third year PhD. Student at the Institute of Operations Research and Analytics, National University of Singapore. She completed her undergraduate study in 2018 from the Indian Institute of Science, Bangalore with a major in Mathematics. Her main research area is in computational statistics, developing new computational and statistical methods for complex processes with intractable likelihoods. She is also interested in the application of statistical methods to disciplines like healthcare, ecology and epidemiology. She is passionate about working with a diverse range of people in different disciplines and help them make informed decisions by making best use of their data.

Vijay Chandru is computational mathematician who is an INAE Distinguished Technologist and an Adjunct Professor at the Indian Institute of Science (IISc) in Bangalore. He co-founded Strand Life Sciences, a spinoff of IISc in precision medicine which he led as executive chairman from inception in 2000 till his retirement in 2018. He also co-founded several non-profit organizations - CHET, Metastring and OPFORD Foundations which are dedicated to health policy, the democratization of data and healthcare access for the underserved orphan diseases. A former president of ABLE, the biotechnology industry apex body, Professor Chandru is a technology pioneer of the World Economic Forum where he has served on the industry advisory council on the future of health.