Chapter #

MINING WEATHER INFORMATION IN DENGUE OUTBREAK

Predicting future cases based on Wavelet, SVM and GA

Yan WU, Gary LEE, Xiuju FU, Harold Soh and Terence HUNG
yan_wu@scholars.a-star.edu.sg
Agency for Science, Technology and Research
*20 Biopolis Way, #07-01 Centros (A*GA), Singapore 138668*
{leekk, fuxj, sohsh, terence}@ihpc.a-star.edu.sg
Institute of High Performance Computing,
1 Fusionopolis Way, #16-16 Connexis, Singapore 13863

Dengue Fever has existed throughout the contemporary history of mankind and poses an endemic threat to most tropical regions. Dengue virus is transmitted to humans mainly by the *Aedes aegypti* mosquito. It has been observed that there are significantly more *Aedes aegypti* mosquitoes present in tropical areas than in other climate regions. As such, it is commonly believed that the tropical climate suits the life-cycle of the mosquito. Thus, studying the correlation between the climatic factors and trend of dengue cases is helpful in conceptualising a more effective pre-emptive control measure towards dengue outbreaks. In this chapter, a novel methodology for forecasting the number of dengue cases based on climactic factors is presented. We proposed to use Wavelet transformation for data pre-processing before employing a Support Vector Machines (SVM)-based Genetic Algorithm to select the most important features. After which, regression based on SVM was used to perform forecasting of the model. The results drawn from this model based on dengue data in Singapore showed improvement in prediction performance of dengue cases ahead. It has also been demonstrated that in this model, prior climatic knowledge of 5 years is sufficient to produce satisfactory prediction results for up to 2 years. This model can help the health control agency to improve its strategic planning for disease control to combat dengue outbreak. The experimental result arising from this model also suggests strong correlation between the monsoon seasonality and dengue virus transmission. It also confirms previous work that showed mean temperature and monthly seasonality contribute minimally to outbreaks.

Keywords— Dengue Fever Modelling, Wavelet, SVM, Infectious Disease, Data Mining.

1. INTRODUCTION

In 2006, the World Health Organization reported that "dengue is the most rapidly spreading vector-borne disease". Between 1985 and 2006, in Southeast Asia alone, an estimated 130 thousand people were infected annually with dengue virus, which is transmitted to humans mainly by the *Aedes aegypti* mosquito [1]. The Dengue virus is transmitted from a female *Aedes* mosquito to a human or vice versa during the feeding process, generally referred to as horizontal transmission. The Dengue virus may also be passed from parent mosquitoes to offspring via vertical transmission.

However, mosquitoes can only survive in certain climate conditions. For example, the *Aedes aegypti* mosquito cannot survive below the freezing temperature of 0°C [2]. However, climatic conditions worldwide have displayed enormous short-term abnormalities, which increases the likelihood of mosquitoes surviving longer and penetrating temperate regions that were previously free from mosquito-borne diseases [2].

In addition to climate conditions, other factors contribute to the time-series transmission profile of the dengue virus, e.g. geographical expansion of *Aedes* mosquito population and its density, the demographic distribution, mobility and susceptibility of the human population. This dynamic relationship is illustrated in the Fig. 1 below.

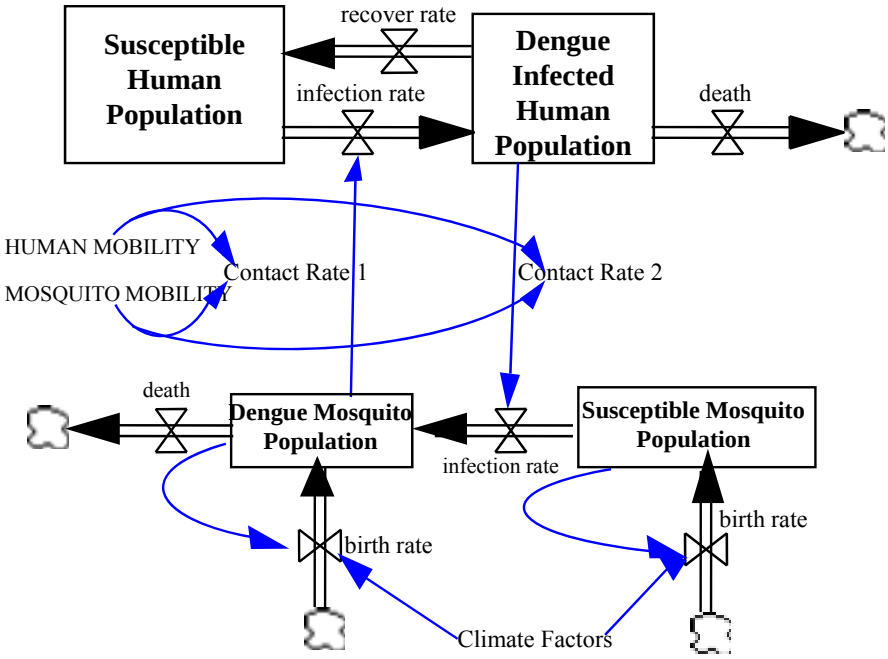


Figure #-1. The dynamics of dengue virus spread

The ideal method of modelling the dynamics of the dengue outbreak is to precisely capture each and every contributing element in the chain. However, associating factors such as mobility of human population to an exact function in the model remains a challenging problem. In this chapter, we proposed a method to simplify the problem in a small relatively well-defined area.

Previous work by Fu et al [3] incorporated time-lags of weather features into their model and selected contributing features with a genetic algorithm. However, the work assumed the usefulness of all levels of detail, in terms of time resolutions/time lags, including noise that was likely present in every feature. Although the method was able to detect seasonal factors that contributed to the outbreak of dengue fever, the resulting model possessed a large number of features, which raised the curse of dimensionality [5]. The work also ignored the use of regression learning for testing the generality of its feature selection model for case forecasting.

The following sections of this chapter describe how a simple yet effective model can be constructed to accommodate the above limitations. Briefly, the proposed method pre-processes the data using Wavelet Transformation (WT) to separate information at different time resolutions superposed in the time-series. Genetic Algorithms (GA) with Support Vector Machines (SVM) are then applied to select features. Following this, regression based on SVM is used to perform forecasting using the model. Conclusions drawn from the empirical modelling using the dengue cases in Singapore are studied, discussed and compared against [3].

2. MODEL CONSTRUCTION

2.1 Data Representation

Data selected for the model should represent the many variables affecting the spread of the dengue virus. Prior studies have shown that the population size of the mosquito depends largely on the presence of suitable breeding sites and climate conditions [4]. As such, we can make the following assumptions of the input data:

- Variations in the weather data represent the breeding patterns of the *Aedes* mosquitoes and
- The time-series of dengue cases in a relatively confined area represents the ensemble of dengue spread information such as susceptibility of human population to the virus.

The coding pattern of the above two pieces of information is unknown but likely non-linear. Moreover, not all information at different time resolutions superposed in the time-series of an input feature, such as temperature, contributes to the fluctuation of dengue cases. Perhaps only one or a few seasonal levels of detail in a particular input feature contribute to the outbreak. Hence, the use of an input feature without eliminating irrelevant levels of details induces unnecessary noise into the model.

In our model, the input features consisted of the past weekly dengue cases and 8 daily weather features, namely rainfall intensity, cloudiness, {maximum, mean and mininum} temperature and humidity. As the dengue cases were only recorded when a patient was diagnosed and the incubation period of dengue virus is known to differ across individuals, daily dengue cases do not accurately represent the outbreak. Thus, summing the daily cases into weekly buckets reduces such inaccuracy. As such, all daily weather factors were averaged into weekly means after performing wavelet decomposition as described in Section 2.2 below.

2.2 Wavelet Decomposition

In order to separate the different levels of details encoded in the input elements of the system, wavelet decomposition of the time-series inputs was chosen. WT are similar to Fourier transforms. The fundamental difference is that instead of deriving the spectral density of a give signal, WT projects the signal onto functions called *wavelets* with dilation parameter ***a*** and translation parameter ***b*** of their mother wavelet:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{1}$$

subject to the condition of:

$$\Psi_o = \int_0^{+\infty} \frac{\left|\Psi(\omega)\right|^2}{\left|\omega\right|} d\omega \quad \text{is bounded} \tag{2}$$

where $\Psi(\omega)$ is the Fourier transform of $\psi(t)$

The continuous WT is defined as follows:

$$T_{a,b}(f) = \int_{-\infty}^{+\infty} f(t)\psi^*(a,b)dt \qquad \text{or}$$

$$T_{a,b}(f) = \left\langle f(t), \psi_{a,b}(t) \right\rangle \tag{3}$$

where $\langle *,* \rangle$ denotes the scalar product in the space of f(t)

Hence the reconstructed signal from the WT is:

$$f(t) = \frac{1}{\Psi o} \int_{-\infty}^{+\infty} \int_{0}^{+\infty} T_{a,b}(f)\psi(a,b)\, da/a^2 dc \tag{4}$$

If a finite number of terms are taken to represent the original signal, (4) can be rewritten as a multi-resolution approximation [5]:

$$\hat{f}(t) = \sum_{i=1}^{N} \Psi_i \psi(a_i, b_i) \tag{5}$$

This can be interpreted as an ensemble of N levels of details $\{D_1 ... D_N\}$ and a level of approximation $\{A_N\}$ which is illustrated in (6).

$$\hat{f}(t) = A_0 = A_N + \sum_{n=1}^{N} D_n \text{ , where } A_n = A_{n+1} + D_{n+1} \tag{6}$$

While the decomposed signals resemble the fundamental ideas of the use of time-lags, it separates signals at different levels which in turn separate potential noise from signals.

2.3 Genetic Algorithm

After Wavelet decomposition, each signal input into the model generates a series of daughter Wavelets inputs. As mentioned earlier, not all levels of detail of every input signal may be relevant. Thus, it is important to filter sub-signals that are irrelevant to the output.

For this purpose, we implemented a biologically-inspired, robust, general-purpose optimisation algorithm: the genetic algorithm (GA). In the GA, each input feature is represented as an allele in a chromosome [6]. The evolution of chromosomes begins from the original (randomly initialized) population and propagates down generationally. During each generation, the fitness of all chromosomes in the population is evaluated. Then, competitive selection and modification are performed to generate a new population. This process is repeated until a predefined limit or convergence is achieved. In order to confirm and validate the feature selection model and to generate a better representative set of contributing factors, a 10 cross-validation technique was introduced into the methodology. The converged solution of weights in 0s and 1s represented the non-selection/selection of a particular sub-signal by the GA.

2.4 Support Vector Machines

The core component of GA is to construct the learning classifier. Among the supervised learning techniques used to correlate input and output data, the Support Vector Machine (SVM) stands firmly as one of the most successful methods. The SVM maps input vectors to a higher dimensional space where a maximum hyperplane is drawn to separate classes [7]. Given a set of input$\{(\mathbf{x}_i, y_i)\}$, where y_i is the corresponding output $\{-1,1\}$ to the i^{th} input feature vector $\mathbf{x}_i$ $\{\mathbf{x_i} \in R^n\}$. $\Phi(\mathbf{x}_i)$ maps $\mathbf{x}_i$ into a higher dimensional space where a possible linear estimate function can be defined:

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + b = \mathrm{sgn}(\sum_{i \varepsilon SVs} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b) \qquad (6)$$

where $K(\mathbf{x}_i, \mathbf{x}) = \Phi(\mathbf{x})^T \Phi(\mathbf{x}_i)$ is the kernel function, *SVs* are the Support Vectors, α is the Lagrange multipliers and *b* is the bias. The SVM problem can be formulated as a quadratic programming problem by optimising α as:

$$\min_{\alpha_i} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad (7)$$

subject to the condition of:

$$0 \le \alpha_i \le C \quad \forall_i \text{ and } \sum_i \alpha_i y_i = 0$$

where C is the capacity control penalty term.

2.5 Regression Learning

After the SVM-based GA selects the most relevant features, the input is fed into a regression learning algorithm which generates a prediction model. An SVM regression (SVR) method based on the material discussed in Section 2.4 above can be used for non-linear learning by introducing a threshold ε analogous to the function of C in SVM Classification [8].

2.6 Model Assessment Criteria

To assess the quality of the final model, we compared predicted results to the actual data using two simple metrics: the Mean Squared Error (MSE) and the Coefficient of Determination (R^2). A small MSE and large R^2 ($0 <= R^2 <= 1$) suggest a better modelling of the data.

3. RESULTS AND DISCUSSIONS

Because Dengue fever occurs throughout the year in Singapore, we gathered the dengue case and weather data in Singapore from 2001 to 2007. As shown in Fig. 2 below, this set of dengue data includes 3 significant peaks at Weeks 190, 245 and 339.

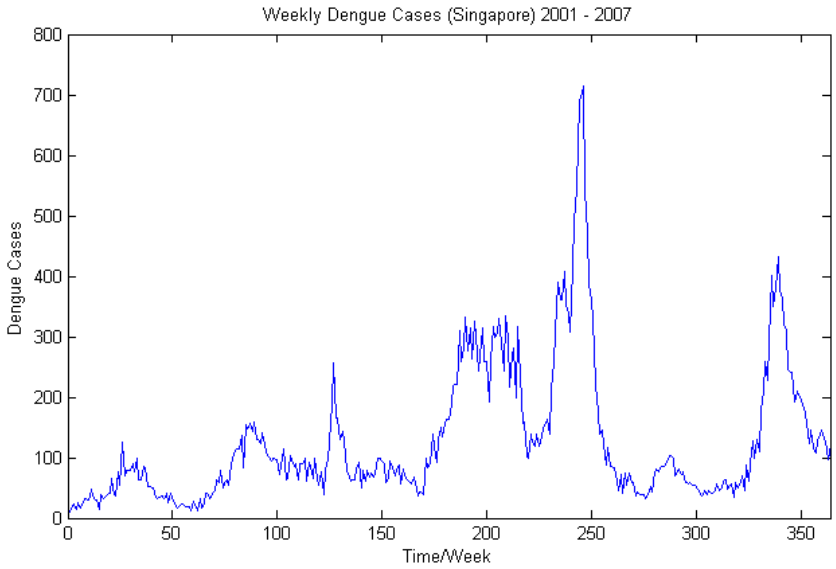


Figure #-2. Weekly Dengue Cases in Singapore from 2001 to 2007

Dengue cases over these years in Singapore are reported weekly while climate data from weather stations were collected daily. We corrected the resolution difference by first performing Wavelet transform of the weather data before sorting the samples into weekly bins. Additionally, we discarded the finest three levels of detail generated by Wavelet transforms because the down-sampling by seven makes $\log_2(7) \approx 3$ levels of detail redundant. The mother Wavelet employed here is the Daubechies[9] 6 Wavelet which was selected because reconstruction of the signal after suppression of the finest 3 levels of detail showed the best approximation of the original signal compared to Daubechies 2 and 4.

Before the GA and SVM were applied, all the input and output vectors were normalised. The normalisation process ensured that none of the variables overwhelmed the others in terms of the distance measure, which was critical for proper feature selection. In our experiment, a zero-mean unit-variance normalisation was employed to give equal distance measures to each feature and levels of detail.

The current dengue cases and decomposed current weather series were used as input features with future dengue cases as output into GA with 10 cross-validations to generate a general model of the data represented. A feature was finally selected if and only if it was selected in at least 7 of the 10 validation cycles. The method employed in GA was a tournament with the aim of minimising the root mean squared error. A recursive method was used to tune the parameter C for best performance (C=1 in this case). There are only 13 terms selected by GA out of the 63 input signals, tabulated in Table 1 below. As compared to the 50 features selected by the model in [3], the dimensionality required to describe the number of dengue cases was significantly smaller in this model. In the Table, the 1^{st} term of a feature denotes its most general trend (A_N) while the 2^{nd} to 7^{th} term, D_N to D_1 respectively, denotes different levels of detail from the coarsest to the finest.

Table #-1. GA Selected Features

Features	Terms selected
Current Dengue Cases	Entire series
Cloudiness	3^{rd} , 4^{th}, 7^{th}
Maximum Temperature	4^{th}
Mean Temperature	None
Minimum Temperature	5^{th}
Maximum Humidity	5^{th}, 7^{th}
Mean Humidity	5^{th}
Minimum Humidity	2^{nd}
Rainfall Intensity	2^{nd}, 5^{th}

From Table 1, we observed that either the 4^{th} or 5^{th} term was selected in almost all the input features. These levels of detail corresponded to the

averaged details between 2 to 4 months. In Singapore, the climate can be divided into two main seasons, the Northeast Monsoon and the Southwest Monsoon season, separated by two relatively short inter-monsoon periods [10]. Each of these periods has a span of 2 to 4 months, which correlates well with our finding and increases the likelihood of a close relationship between dengue outbreak and seasonality in Singapore, i.e. the seasonal change of temperature, humidity and rainfall. From the literature reviewed in [2], it had been concluded that the rise in temperature does accelerate the life-cycle of the mosquito. This agrees with the findings from our model, i.e. the 1st terms of all temperature features were not selected.

In [3], Fu suggested that time-lags of {0, 2, 6, 8, 12} weeks are important. In contrast, our model selected a time-lag of 4 weeks which corresponded to an averaged detail of 4 weeks, i.e. 6th term in our feature vectors. From Table 1, none of the 6th terms were selected in this model. Taken together, both models suggest that the contribution of monthly fluctuations to dengue virus transmission may not be significant. Moreover, both models also suggest that mean temperature does not contribute to the prediction models at any level. The general trend (7th term) present in Cloudiness and Maximum Humidity suggests that some positive correlation exists between the growth of dengue cases and these two features.

In our previous work [11], we demonstrated that for future dengue outbreak prediction, SVR was likely a better model than linear regression due to its robustness even in the event of over-fitting. In [11], to investigate our claim that the coding pattern of weather data into dengue cases was non-linear, we performed regression with a Radial Basis Function (RBF) SVR with 5 years of training data and 1 year of testing input. The width γ used in RBF was set to be the inverse of the sample size (312). The range of data used for training and testing were taken from 2001 – 2006.

To confirm the validity of the model, we carried out a similar experiment with a different range of training and testing data (2002 – 2007) while keeping the model variables unchanged. Instances of the results from these two sets of data are tabulated in Table 2 and plotted in Fig. 3 (red and blue lines denote actual and predicted No. of cases respectively). The performance statistics show that mapping the weather data into the RBF space (which is non-linear) produced a high correlation between the input data and the actual dengue cases, reinforcing our claim of non-linearity.

Statistics in Table 2 show that on average, SVR with GA yields a 70% reduction in MSE and a 25% increase in correlation with the actual data as compared to that of SVR without GA. From this observation, we can infer that a significant number of irrelevant input features were removed by the GA, supporting our claim that not all levels of detail influenced the spread of dengue.

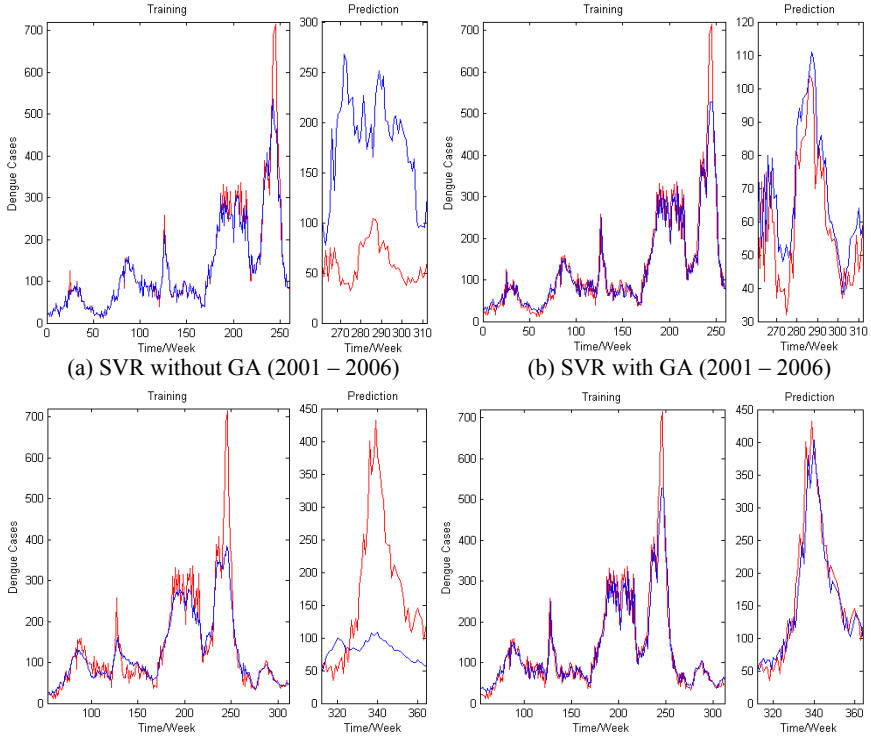
In order to test the stability of our proposed model, a further experiment was carried out. We extended the testing samples to two consecutive years (2006 – 2007) using the same training data (2001 – 2005). The result is tabulated in Table 3 in comparison with the previous two experiments.

Table #-2. Performance of an instance of RBF SVR approach

RBF SVR	With GA	MSE	R^2	Figure
Training data: 01 -05	No	0.27	0.75	3 (a)
Testing data: 06	Yes	0.091	0.92	3 (b)
Training data: 02 -06	No	0.37	0.71	3 (c)
Testing data: 07	Yes	0.085	0.91	3 (d)

Table #-3. Performance of different sample combinations

RBF SVR with GA	MSE	R^2
Training: 01 -05 Testing: 06	0.091	0.92
Training: 01 -05 Testing: 06 & 07	0.094	0.90
Training: 02 -06 Testing: 07	0.085	0.91



(a) SVR without GA (2001 – 2006) (b) SVR with GA (2001 – 2006)

(c) SVR without GA (2002 – 2007) (d) SVR with GA (2002 – 2007)

Figure #-3. Prediction results of SVR with and without GA

The performance statistics in Table 3 clearly show that the difference in MSE is minimal (< 0.01) and the correlation coefficient is within 2% of each other. These empirical results suggest that the proposed model is capable of the followings:

1. Producing relatively reliable predictions for up to 2 years ahead
2. Constructing a stable model instance using only 5 years of data

4. CONCLUSION

In this chapter, a novel model for predicting future dengue outbreak was presented. We proposed the use of Wavelet decomposition as a mean of data pre-processing. This technique, together with SVM-based GA feature selection and SVR learning, proved to be effective for analysing a special case of infectious diseases – dengue virus infection in Singapore. Our empirical results strongly suggest that weather factors influence the spread of dengue in Singapore. However, the results also showed that mean temperature and monthly seasonality contribute minimally to outbreaks. Analysis of the results in this chapter gives rise to further epidemiological research questions on dengue transmissions:

- As low frequency (higher order) terms denote general trends of the signal, the results tabulated in Table 1 indicate that dengue incidence seems to be closely related to the rainfall and humidity trends. Can this be proven epidemiologically?
- Although the rise in average temperature is thought to accelerate mosquito's breeding cycle, why was there no low frequency terms of temperature represented in the model? Moreover, why are some moderate frequencies (between 2 – 4 months) of both maximum and minimum temperatures selected in the model?
- As Singapore is an area surrounded by sea, does this research finding fit well into other regions with dengue prevalence?

To provide answers to these questions, further work has to be performed, particularly in the collection of more precise data over a longer period of time. Predictions of longer durations ahead, such as 2 weeks onwards, can also be constructed to measure the performance of the model.

Because outbreaks are stochastic events, it is not possible to precisely predict the specific realization of one particular outbreak. However, models, such as the one proposed in our work, are useful for providing quantitative risk assessments of disease spread in a well-defined area, which is essential information for constructing effective public-health strategies.

5. REFERENCES

[1] World Health Organisation: Dengue Reported Cases. 28 July 2008. W.H.O. <http://www.searo.who.int/en/Section10/Section332_1101.htm>

[2] Andrick, B., Clark, B., Nygaard, K., Logar, A., Penaloza, M., and Welch, R., "Infectious disease and climate change: detecting contributing factors and predicting future outbreaks", *IGARSS '97: 1997 IEEE International Geoscience and Remote Sensing Symposium*, Vol. 4, pp.1947 - 1949, Aug 1997.

[3] Fu, X., Liew, C., Soh, H., Lee, G., Hung, T., and Ng, L.C. "Time-series infectious disease data analysis using SVM and genetic algorithm", *IEEE Congress on Evolutionary Computation (CEC) 2007*, pp. 1276-1280, Sep 2007.

[4] Favier, C., Degallier, N., Vilarinhos, P.T.R., Carvalho, M.S.L., Yoshizawa, M.A.C., and Knox, M.B., "Effects of climate and different management strategies on Aedes aegypti breeding sites: a longitudinal survey in Brasília (DF, Brazil)", *Tropical Medicine and International Health 2006*, Vol. 11, No. 7, pp. 1104-1118, Jul 2006.

[5] Mallat, S.G., "Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$", *Transactions of the American Mathematical Society*, Vol. 315, No. 1, pp. 69-87, Sep 1989.

[6] Grefenstette, J.J., *Genetic algorithms for machine learning*, Kluwer Ac Pub, 1993.

[7] Burges, C.J.C., "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 955-974, 1998.

[8] Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., and Vapnik, V., "Support Vector Regression Machines", *Advances in Neural Info Processing Systems 9*, pp. 155-161, 1996.

[9] Daubechies, I. "Orthonormal Bases of Compactly Supported Wavelets." *Comm. Pure Appl. Math*. Vol. 41, pp. 909-996, 1988.

[10] National Environment Agency, Singapore: Climatology of Singapore. 20 Aug 2007. NEA, Singapore. <http://app.nea.gov.sg/cms/htdocs/article.asp?pid=1088>

[11] Wu, Y., Lee, G., Fu, X., and Hung, T., "Detect Climatic Factors Contributing to Dengue Outbreak based on Wavelet, Support Vector Machines and Genetic Algorithm", *World Congress on Engineering 2008*, Vol. 1, pp. 303 – 307, July 2008.

[12] Bartley, L.M., Donnelly, C.A., Garnett, G.P., "Seasonal pattern of dengue in endemic areas: math models of mechanisms", *Trans R Soc Trop Med Hyg*, pp. 387-397, Jul 2002.

[13] Shon, T., Kim, Y., Lee, C., and Moon, J., "A machine learning framework for network anomaly detection using SVM and GA", *Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop 2005*, pp. 176- 183, Jun 2005.

[14] Nakhapakorn, K., and Tripathi, N. K., "An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence", *International Journal of Health Geographics*, Vol. 4, No. 13, 2005.

[15] Ooi, E., Hart, T., Tan, H., and Chan, S., "Dengue seroepidemiology in Singapore", *The Lancet*, Vol. 357, I. 9257, pp. 685-686, Mar 2001.

[16] Ministry of Health, Singapore: Weekly Infectious Diseases Bulletin. 28 July 2008. M.O.H. Singapore. <http://www.moh.gov.sg/mohcorp/statisticsweeklybulletins.aspx>

[17] Gubler, D.J., "Dengue and dengue hemorrhagic fever", *Clinical Microbiology Reviews*, Vol. 11, No. 3, pp. 480–496, July 1998.

ACKNOWLEDGMENT

Our heartfelt gratitude is expressed to William C. Tjhi from Institute of High Performance Computing, Singapore for reviewing the Chapter. We also wish to acknowledge NOAA of the U.S.A. for releasing daily weather data free-of-charge for research and education.