Indian Journal of Science and Technology, *Vol 11(4), DOI: 10.17485/ijst/2018/v11i4/115405,* January *2018*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Using Public Open Data to Predict Dengue Epidemic: Assessment of Weather Variability, Population Density, and Land use as Predictor Variables for Dengue Outbreak Prediction using Support Vector Machine

Chiung Ching Ho[1], Choo-Yee Ting[1] and Dhesi Baha Raja[2]

[1]Data Science SIG, Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor DE, Malaysia; ccho@mmu.edu.my, cyting@mmu.edu.my
[2]Department of Community Medicine & Public Health, Faculty of Medicine and Health Sciences, Universiti Malaysia Sarawak; dr.dhesi@gmail.com

Abstract

Objectives: This study was performed to predict dengue outbreaks using predictor variables derived from weather variability, population density, land use and elevation in Malaysia. **Methods:** We used publicly available data associated with dengue from the Malaysian open data platform and historical dengue case data from the Ministry of Health Malaysia. We investigated the correlations between predictor variables related to Weather Variability, Population Density, Residential Building Types and Construction Sites; with that of outbreaks of dengue in the chosen site of study, and used the Support Vector Machine classifier to predict outbreak of dengue fever. **Results:** The model we proposed was evaluated through cross-validation, and returned an accuracy rate of 88.62%, while sensitivity was 93%, and specificity was recorded at 79.32%. Population density had the highest correlation with predicted dengue fever outbreak. **Conclusions:** In this study, we assessed weather variability, population density, and land use as predictor variables for predicting dengue fever outbreak using a Support Vector Machine classifier.

Keywords: Dengue, Infectious Disease, Meteorology, Projection and Predictions, Support Vector Machine

1. Introduction

1.1 Dengue Virus

The World Health Organization (WHO) recognises that dengue is a major international public health concern, with severe dengue affecting many Asian and Latin American countries. Dengue affects 2.5 billion people worldwide, with estimates of 50-100 million dengue infections occurring worldwide. These infections results in 50 thousands hospitalization, with the majority of those hospitalized being young children. Current dengue trends indicates that not only are the number of new cases increasing over the years, but are also occurring in new locales which were previously dengue-free, such as France, Croatia and Portugal in mainland Europe[1].

1.2 Current Surveillance

Current surveillance methods for dengue employed by agencies participating in dengue surveillance have been classified as either active, passive or sentinel site

*Author for correspondence

surveillance. The 2010 meeting between the Asia-Pacific Dengue Prevention Board (APDPB) and the Americas Dengue Prevention Board (ADPB) reported that passive surveillance of dengue is most common (76.9%), followed by sentinel site surveillance (19.2%) while active surveillance is done the least (3.9%)[2]. In the same meeting, weaknesses of current surveillance systems were expressed, including the following:

- The preference by decision makers for dengue control over that of dengue prediction,
- Lack of preventive dengue services,
- Difficulty in interpretation of diagnostic tests,
- Centralized reporting which prevents timely local actions,
- Lack of biological surveillance,

Lack of coordination and data sharing between surveillance performed by epidemiologist and mosquito control units.

The Support Vector Machine (SVM) model proposed in this paper can be used to address many of the weaknesses identified above, especially in the area of making accurate dengue outbreak prediction. Once accurate dengue outbreak prediction is at hand, effective dengue prevention actions can be conducted with greater confidence. The SVM model can be used as an engine for a centralized coordination web application, where epidemiologist can give expert advice on aedes prevention actions to mosquito control units, and this is demonstrated by a proof-of-concept application developed by the authors at http://bit.ly/1SyDFWc.

1.3 Open Data for Health

The problems of unverified dengue reports and misdiagnosis of dengue can be solved through the provision of data that has been verified and confirmed by a medical professional, who then sends it for compilation by the Ministry of Health and its related agencies. Information such as the number of cases, mortality rates, and accumulated length of dengue infection in a locale are valuable indicators on the spread of Dengue. These data are often aggregated, where the identity of the patients are removed, thus solving the problem of infringing on patient privacy. The data is then made publicly available, also known as Open Data.

Health related Open Data has seen increased usage over the years. Epivue[3] an open source public health information system was used to manage and disseminate public health data from the Washington State Cancer Registry and Washington State Centre for Health Statistic. HealthData.gov has published 400 health dataset[4], which contains data ranging from epidemiology to Medicare. In the UK, open health data has been used to potentially lower the cost of statin prescribed under the United Kingdom's National Health Service[5] by examining a public data set containing prescription written by every family physician.

In Malaysia, open data for health is limited to press statement and press releases to the general public issued by the Ministry of Health and its officers, and also by statements made by the Minister of Health. A recent effort at using open public data for health can be found in iDengue[6], an informational website which lists out number and cases in Malaysia and their corresponding geo-location. I Dengue not do feature pattern recognition, which is addressed in our work. In our work, pattern recognition is performed by the SVM model using predictor variables from meteorology, land use and population density, which can be then used to predict future dengue outbreaks and is described fully in the Methods section of this paper.

1.4 Dengue Predictors

Predictors for the outbreak of dengue fever include weather variability, mosquito density and imported cases, overtop surveillance, social-economic factors, population density and land use. Weather variability has both positive and negative effects on the spread of dengue. Rainfall[7] has been proven to positively co-relate with the spread of dengue, while temperature variation and relative humidity has been shown to impede dengue[8]. A recent study[9] on the effects of weather on dengue fever spread showed that temperature, precipitation, vapour pressure and minimum relative humidity was positively associated, whilst being negatively associated with air pressure. The same study also showed that mosquito density was a positive indicator, which validates the usage of overtops as a dengue indicator; as shown by other related works[10-12]. In addition to all these factors, factors such as human population density[13] and proximity to construction sites[14] are beginning to be investigated with regards to dengue fever

spread. Socio-economic conditions[15] is another emerging area of research to be considered in the spread of dengue fever.

2. Methods

Overview: The Big App Challenge (BAC) version 1.0 is the inaugural national-level Big Data competition to be held in Malaysia. The aim of the BAC was to solve problems related to Malaysia's national and societal issues through the use of open data[16] the open data was provided by agencies belonging to the government of Malaysia[17], in particular the Ministry of Health and the Meteorology Department. The Ministry of Health provided data related to dengue fever cases in Malaysia, while the Meteorology Department provided weather related data in the same period of time. The provided open data was subsequently augmented with other data sources to create a classification model for dengue cases.

2.1 Locale Selection

The state of Selangor in the Federation of Malaysia to date[6] has the highest number of accumulated dengue cases in Malaysia, amounting to 59.7% of total accumulated dengue fever cases recorded in Malaysia. Selangor is the single largest cluster of dengue activity in Malaysia currently. In addition to this, Selangor is the state with the highest density of population and the greatest amount of physical infrastructure development. Previous investigations on dengue fever have been carried out in Selangor, focusing on issues such as the correlation of rainfall and dengue infection[7], overtop surveillance for mixed breeding of different mosquitoes[12], weather effects in addition to rainfall[18], and a recent studies on spatial-spatio-temporal patterns of dengue fever[19] as well land use factors and its correlation with dengue fever[20]. Figure 1 shows the heat-map of accumulated dengue hotspots from the years 2010-2014 in Malaysia.

3. Data Sources

Data was provided by the organizers of the BAC, which originated from the Ministry of Health Malaysia and the Meteorology Department. The data provided consisted of dengue cases per calendar week from 2010 to the 33rd week of year 2014 for all the states in Malaysia. Mortality rates of dengue fever for the same time period were provided as well. For dengue cases, related details such as the location, accumulated number of dengue fever cases

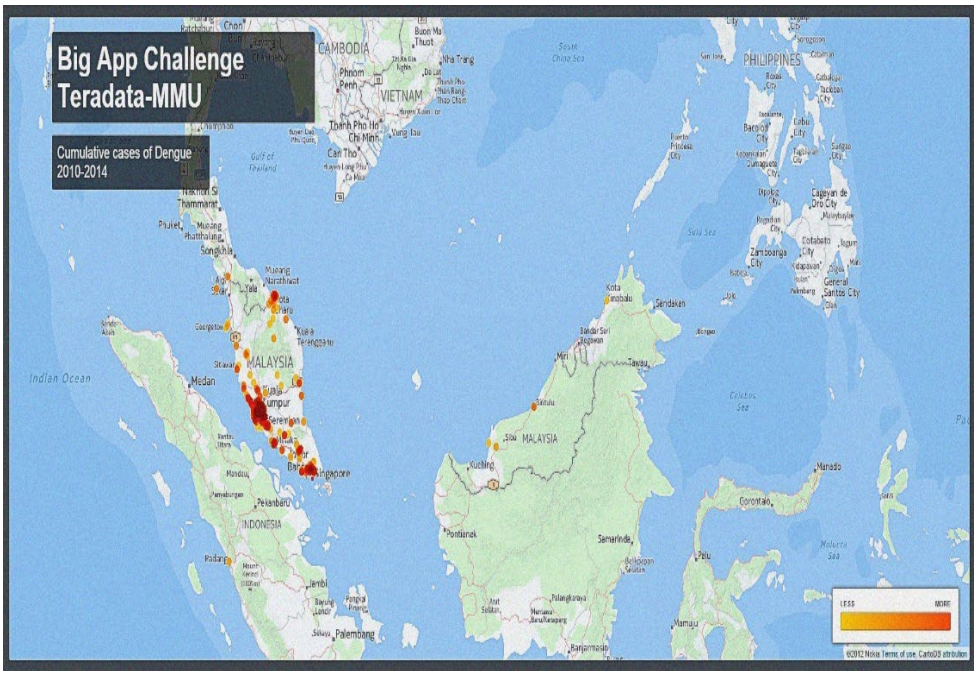

Figure 1. Heat-map of accumulated dengue hotspots in Malaysia from 2010-2014.

and the duration of the outbreak was provided for each epic week. The accumulated number of dengue cases is counted upon confirmation of a dengue case by public health officials. Location was given at the state level, district, zone, local authority, and locality. In Malaysia, diagnosis of dengue cases is performed via clinical diagnosis.

The Meteorology Department provided information from 44 weather stations. The longitude and latitude of each station was provided, as well as the height of the station from the mean sea level. Each station provided the daily rainfall (measured in millimetres), 24 hour mean wind (measured in metre-per-second), and daily solar radiation (measured in Mega Joules per square metre). While daily rainfall and 24 hour mean wind was available from all weather stations, the daily solar radiation was only available from weather stations. Additional information in the form of number of occurrences of thunderstorms in the hotspots investigated was sourced from the Meteorology Department, as moisture is a key ingredient of a thunderstorm. Moisture is also a positive indicator for the spread of dengue[21].

The 2010 population distribution report[22] published by the Department of Statistic Malaysia was used an additional data source to supplement the provided data, showing population density and the majority race in a given area. The population density of major dengue hotspots were scaled accordingly.

1 = Less than 100 person per square km
2 = 101 - 500 person per square km
3 = 501 - 1000 person per square km
4 = 1001 - 1500 person per square km
5 = At least 1501 person per square km

Population density is an important indicator on the ease of dengue spread, as it is directly related to human-Aedes mosquito contact. In addition to that, in densely populated areas, people are on the move more, further increasing those chances for human-to-mosquito transmission of dengue fever[23]. The same report also showed the majority race in a given area.

The geo-coding of dengue fever hotspots on Google maps gave further insights on the elevation of the hotspots, as well as the proximity of construction sites to a dengue fever hotspot. Investigation of construction sites is important as stagnant water is frequently found in such sites, and has been proven to promote spread of aides mosquitos[24]. In addition to this, geo-coding of dengue fever hotspots also showed the type of accommodation in use.

In summary, the data used in our work originated from public data that was made public, augmented with publicly available data from map service providers.

4. Data Perprocessing

Data pre-processing is an important step in ensuring the formation of a valid model to investigate the co-relation of dengue fever with the various indicators investigated in this work. The data provided by the Meteorology Department from the weather stations had missing data on some days, and these missing data are assumed to be missing completely at random. These data was normalized to zero. Other pre-processing work done was the correction of addresses of dengue hotspots in terms of spelling and address conventions. Non-numeric data including types of accommodation, race, development etc. were transformed into categorical data, which is acceptable for use using the Support Vector Machine (SVM) classifier.

The corrected addresses of hotspots were then subsequently geo-coded to points on Google Map. From the geocoded points, other related information such as the altitude and proximity to construction site were acquired. Visual inspection of the geocoded points also identified whether the hotspot was a high-rise building or a terrace house.

4.1 Model Selection and Validation

Heuristicin Malaysia, a dengue outbreak is deemed to have occurred when there are 2 or more cases of dengue fever happening within one incubation period in a radius of 200-400 metres, according to clinical practice guidelines published by the Academy of Medicine of Malaysia[25]. Based on these guidelines, dengue outbreak is deemed to have taken place based on the following heuristics.

For any given epid week, we take the number of accumulated cases before that epic week as T1. Subsequently, the accumulated number of cases four weeks after the reference epic week is denoted as T4. This heuristic also considers population density. Should the difference of accumulated number of cases between T1 and T4 exceed 4 cases happens in an area which has population den-

sity of 3 (indicating a density of 501-1000 person per square kilometre), a dengue outbreak is deemed to have occurred. Similarly a dengue outbreak is deemed to have happened if the difference between accumulated number of cases between T1 and T4 exceeds 8 cases, and it happens in areas which has population density of 5 (indicating at least 1501 person per square kilometre).

Using logical inference, the heuristic is described as such: Let p beT4-T1>4 and Population Density =3, while q be T4-T1>8 and Population Density = 5, and r be Dengue Outbreak. The logical disjunction expression representing the heuristic used in determining whether a dengue outbreak has occurred is shown below:

$$(p \ \ vq) \rightarrow r$$

5. Model Selection

In our work, we applied a Support Vector Machine (SVM) learning algorithm to identify the association between a dengue outbreak and the various indicators investigated. SVM is a non-probabilistic binary linear classifier, when used to classify training examples into one of two categories, as is the case for the work we performed. SVM has been used successfully for the purposes of forecasting dengue using climate-based models[26], and prediction of dengue outbreaks through search query surveillance[26]. The former and latter results both showed that SVM performed well in discriminating between two classes, as compared to results achieved using techniques such as Rough Set Access, Naive Bays, Association Classification and combination techniques[27].

The SVM classifier was applied to predictor variables comprising meteorological, epidemiological, geo-social, and demographical elements. The predictor variables used in the outbreak model included total number of accumulated dengue cases, number of days since the onset of dengue fever, total number of dengue cases 1 week prior to an epid week, wind direction (azimuth degrees), wind speed (M/s), rain volume (mm), occurrences of thunderstorm (categorical yes or no), altitude (metres above sea-level), race, population density (scalar range of 1-5 as described in the previous section), type of accommodation (high-rise or low-rise), proximity to a construction site(metres), and whether an outbreak is predicted to have happened or not. The determination of whether an outbreak has occurred is based on the heuristic shown previously.

The predictor variables subsequently were normalized to between a range of 0.0 to 1.00, as SVM has been shown to perform better when the data has been scaled. Scaling reduces the effect of attributes in higher numeric range from dominating attributes in lesser numeric range. A radial basis function is chosen as the kernel of the SVM as it is a versatile technique which works in any dimension of data, and it is highly accurate in situations where data is dense. A grid search was performed to choose the best value of cost and gamma, two parameters in a SVM. The SVM classification was performed using the WEKA machine learning tool[28].

6. Results

The assessment of the fitness of use for the chosen predictor variables for the task of dengue outbreak prediction was firstly done based on their relative Pearson coefficient, which indicates the worth of the variable in determining the class as shown in Figure 2. The predictor variables evaluated are the same set of predictor variables as the ones described in the Model Selection sub-section.

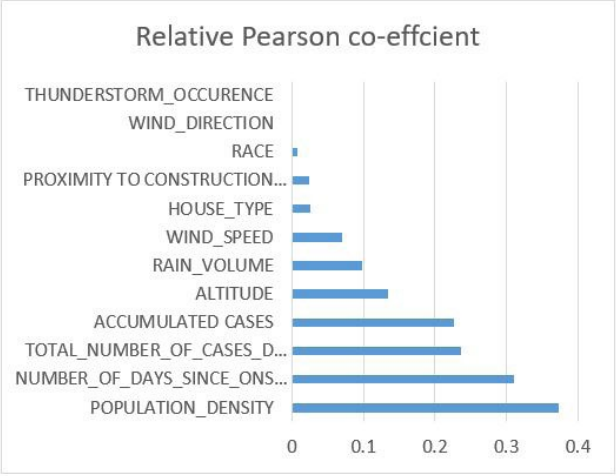

Figure 2. Value of Pearson correlation for all predictor variables in relation to each other.

Since there are five predictor variables that showed a Pearson correlation, r, value of between 0.1 and 0.3 (categorized as a "weak" correlation[29] all of the predictor values were used by the SVM classifier.

Evaluation of the model was done through 10-fold cross validation of the data set comprising normalized predictor variables. The original dataset is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

The model we proposed performed well when it was evaluated with the SVM classifier. The measures used to evaluate the model are accuracy, sensitivity (also known as True Positive Rate (TPR)), specificity, Area under ROC Curve (AUC), and False Positive Rate (FPR).

Accuracy of the model is 88.62%, while sensitivity was 93%, and specificity was recorded at 79.32%. The Area under ROC (AUC) was 0.8618 as shown in Figure 3. TPR for the model was 93%, while FPR was 20.7%.

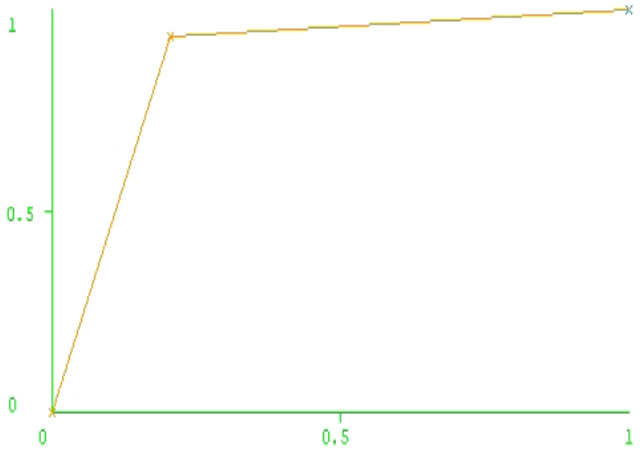

Figure 3. Plot of ROC curve for SVM classifier with AUC of 0.8618 using Weka.

7. Discussion

The results of our work have showed that the usage of open data can be useful in modelling the meteorological, epidemiological, geo-social, and demographical predictors for a dengue outbreak. The model is accurate, and is both sensitive and specific towards outbreak of dengue. Our model has shown that the key variables for predicting a dengue outbreak are related to population density as well as the number of dengue cases, as well as the duration of time since the first instance of a dengue case in a geographical area. Areas with higher population density are prone to future outbreaks of dengue, especially if the dengue fever outbreak is not contained, as shown by high accumulated number of cases and longer duration of time since the first recorded case.

The second grouping of variables which are important is that of meteorology-linked variables, including wind speed and volume of rain. This is consistent with other studies which have concluded that wetter weather is conducive for mosquito propagation, as is stronger wind speeds.

The final grouping of variable which we studied was related to demographics and geographical features. It would seem that proximity to a construction site, as well the type of residence do not impact as much on dengue fever outbreak. Likewise, dengue outbreaks are not prevalent to any category of demographics. Our results compared favourably with similar work. Rainfall and temperature was used as an indicator for a dengue outbreak in India, using a Neural Network model. The accuracy for this work[30] ranged from 81% to 92%, depending on the number of hidden nodes. In Peru, indicators used for the prediction of dengue outbreaks included rainfall, temperature, vegetation cover, atmospheric pressure, sea surface temperature, elevation, as well as socio-economic and demographic data. The fuzzy-association rules based model used for this work resulted in near 100% specificity, although sensitivity was reported as about 60%. Our model outperformed the multiple rule-based model[30] which was used to predict dengue outbreak in Malaysia, with an AUC of 73.78 produced by the Naive Bays classifier.

While the results validated the usage of open public data, there are significant issues that need to be addressed in order to improve the usability of open and public data. First and foremost, ease of access to public data needs to be improved. Most dengue related work performed in Malaysia requires the collaboration of the owner of the data, particularly dengue cases and meteorological data to share their data. The situation is slowly improving, as dengue related data are now being made available on social media. Currently, dengue case data is being made available on the national Malaysian Open Data website.

The second challenge is that of timeliness of the data provided. While census data is periodic by nature, dengue and meteorological data requires a significant turn-around time before it is provided on demand. This is a significant challenge for any effort to implement near-real-time monitoring of dengue.

A third challenge is the uniformity of data format and data sources. Whilst public data is beginning to be shared via the official open data portal, it is shared either as a spreadsheet, portable document format, or as a graphic file. Adoption of standards for data definition, documents and interoperability is a step needed to encourage the use of open data.

8. Conclusion

We have contributed a model for predicting breakout of dengue using publicly assessable open data. This model performs well and compares favourably with other similarly effort. We hope that the success of creating this model will encourage further sharing of data by all parties, especially when the data is to be used to combat a public health threat.

9. Acknowledgement

We would like to thank Multimedia Development Corporation (MDeC) and TentSpark for organizing the Big App Challenge. We are also very thankful to the Ministry of Health Malaysia and the Department of Meteorology for providing open data that made the creation of this model possible. We also acknowledge the input given by TeraData Malaysia in the process of creating this model.

10. References

1. WHO Media Centre. WHO Dengue and severe dengue. Dengue and severe dengue. Available from: crossref. Date accessed: 04/2017.
2. Beatty ME, Stone A, Fitzsimons DW, Hanna JN, Lam SK, Vong S. Best Practices in Dengue Surveillance: A Report from the Asia-Pacific and Americas Dengue Prevention Boards. Gubler DJ, editor. PLOS Public Library of Science Neglected Tropical Diseases. 2010 Nov; 4(11):1-890. crossref.
3. Yi Q, Hoskins RE, Hillringhouse EA, Sorensen SS, Oberle MW, Fuller SS. Integrating open-source technologies to build low-cost information systems for improved access to public health data. International Journal of Health Geographics. 2008; 7(1):1-29. crossref. PMid:18541035 PMCid:PMC2432052.
4. Open Health Data. Available from: crossref. Date accessed: 24/01/2013.
5. Open data and health care: Beggar thy neighbour. Available from: crossref. Date accessed: 08/12/2012.
6. Agensi Remote Sensing Malaysia (ARSM). Available from: crossref. Date accessed: 08/04/2017.
7. Li CF, Lim TW, Han LL, Fang R. Rainfall, abundance of Aedes aegypti and dengue infection in Selangor, Malaysia. Southeast Asian Journal of Tropical Medicine and Public Health. 1985 Dec; 16(4):560-8. PMid:3835698.
8. Wu P-C, Guo H-R, Lung S-C, Lin C-Y, Su H-J. Weather as an effective predictor for occurrence of dengue fever in Taiwan. Acta Tropica. 2007 Jul; 103(1):50-7. crossref. PMid:17612499.
9. Sang S, Yin W, Bi P, Zhang H, Wang C, Liu X. Predicting Local Dengue Transmission in Guangzhou, China, through the Influence of Imported Cases, Mosquito Density and Climate Variability. PLOS Public Library of Science ONE. 2014 Jul; 9(7):102-755. crossref.
10. Chen CD, Benjamin S, Saranum MM, Chiang YF, Lee HL, Nazni WA. Dengue vector surveillance in urban residential and settlement areas in Selangor, Malaysia. Tropical Biomedicine. 2005 Jun; 22(1):39-43. PMid:16880752.
11. Pessanha JEM, Brandao ST, Almeida MCM, Cunha M da CM, Sonoda IV, Bessa AM. Ovitrap surveillance as dengue epidemic predictor. Journal Health Biology Science. 2014 Jun; 2(2):51-6. crossref.
12. Chen CD, Nazni WA, Lee HL, Seleena B, Mohd Masri S, Chiang YF. Mixed breeding of Aedes aegypti (L.) and Aedes albopictus Skuse in four dengue endemic areas in Kuala Lumpur and Selangor, Malaysia. Tropical Biomedicine. 2006 Dec; 23(2):224-7. PMid:17322826.
13. Padmanabha H, Durham D, Correa F, Diuk-Wasser M, Galvani A. The Interactive Roles of Aedes aegypti Super-Production and Human Density in Dengue Transmission. PLOS Public Library of Science Neglected Tropical Diseases. 2012 Aug; 6(8):17-99. crossref.
14. Brown CA, Anang Y, Okorie PN. Role of The Construction Industry In Promoting Mosquito Breeding In And Around The Accra Metropolis, Ghana. International Journal Science Technology Resource. 2014 Jul; 3(7):94-100.
15. Parham PE, Waldock J, Christophides GK, Hemming D, Agusto F, Evans KJ. Climate, environmental and socio-economic change: weighing up the balance in vector-borne disease transmission. Philosophical Transactions of the Royal Society Biological Sciences. 2015 Apr; 370(1665):2013-0551.

16. MDeC, Tentspark kick off big data analytics/open data app challenge | Digital News Asia. Available from: crossref. Date accessed: 25/02/2014.
17. Malaysian developers challenged to produce 'Big Data' apps. Available from: crossref. Date accessed: 24/09/2014.
18. Cheong YL, Burkart K, Leitao PJ, Lakes T. Assessing weather effects on dengue disease in Malaysia. International Journal of Environmental Research Public Health. 2013 Dec; 10(12):6319-34. crossref. PMid:24287855 PMCid:PMC3881116.
19. Ling CY, Gruebner O, Kramer A, Lakes T. Spatio-temporal patterns of dengue in Malaysia: combining address and sub-district level. Geospatial Health. 2014; 9(1):131-40. crossref. PMid:25545931.
20. Cheong YL, Leitao PJ, Lakes T. Assessment of land use factors associated with dengue cases in Malaysia using Boosted Regression Trees. Spat Spatio-Temporal Epidemiology. 2014; 10:75-84. crossref. PMid:25113593.
21. Schreiber KV. An investigation of relationships between climate and dengue using a water budgeting technique. International Journal of Biometeorology. 2001 Jul; 45(2): 81-9. crossref. PMid:11513051.
22. Department of Statistics Malaysia. Available from: crossref. Date accessed: 25/10/2017.
23. Management of dengue infection in adults. Clinical practice guidelines. 2008; p. 1-68.
24. Descloux E, Mangeas M, Menkes CE, Lengaigne M, Leroy A, Tehei T. Climate-Based Models for Understanding and Forecasting Dengue Epidemics. Anyamba A, editor. PLOSPublic Library of Science Neglected Tropical Diseases. 2012 Feb; 6(2). crossref.
25. Althouse BM, Ng YY, Cummings DAT. Prediction of Dengue Incidence Using Search Query Surveillance. Crockett RJK, editor. PLOSPublic Library of Science Neglected Tropical Diseases. 2011 Aug; 5(8). crossref.
26. Bakar AA, Kefli Z, Abdullah S, Sahani M. Predictive models for dengue outbreak using multiple rule base classifiers. 2011; p. 1-6.
27. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. ACM SIGKDD Explorations Newsletter. 2009 Nov; 11(1):1-10. crossref.
28. Dancey CP, Reidy J. Statistics without Maths for Psychology: Using SPSS for Windows. Upper Saddle River, NJ, USA: Prentice-Hall. 2004. PMCid:PMC2361735.
29. Munasinghe A, Premaratne HL, Fernando MGNAS. Towards an Early Warning System to Combat Dengue. International Journal of Computer Science Electronic Engineering. 2013; 1(2):252-6.
30. Buczak AL, Koshute PT, Babin SM, Feighner BH, Lewis SH. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Medical Informatics and Decision Making. 2012; 12(1): 1-124. crossref. PMid:23126401 PMCid:PMC3534444.