# Predicting Dengue Outbreaks in Sri Lanka using Human Mobility - A Genetic Algorithms integrated Support Vector Regression Approach

Lasantha Fernando[1], Amal Shehan Perera[2] and Sriganesh Lokanathan[1]

[1] LIRNEasia, 12, Balcombe Place, Colombo 08, Sri Lanka,
[2] University of Moratuwa, Moratuwa 10400, Sri Lanka

**Abstract.** We propose a machine learning model to predict dengue outbreaks 2 weeks ahead in Sri Lanka. In this work, we use support vector regression with human mobility as an input feature in addition to other input features such as past dengue cases, temperature, rainfall and vegetation index. Human mobility is derived using historical pseudonymised mobile phone call detail records (CDR) from multiple mobile phone operators in Sri Lanka. We also show that human mobility contributes to the spread of the disease even in dengue endemic regions. A genetic algorithms based approach was integrated to the training phase of our model to select the best feature combination out of more than 70 input features that were derived after performing feature engineering. As compared to a stand alone support vector regression, our approach is able to predict an epidemic curve that is more accurate while also maintaining a lower error.

**Keywords:** disease outbreak prediction, human mobility, CDR, support vector regression, genetic algorithms

## 1 Introduction

Dengue is a vector borne tropical infectious disease that affects 50-100 million people globally every year and is endemic in Sri Lanka. Predicting dengue outbreaks 2 weeks before hand in Sri Lanka would assist in executing preventive measures by the public health sector. In this work, we evaluate multiple machine learning techniques on a smaller dataset to determine which technique performs best and continue further optimisations and tuning on the selected technique to obtain dengue outbreak predictions for a larger dataset. In our work, we use human mobility as an input feature in addition to other input features. Human mobility has been established as an important factor in the spread of vector-borne pathogens such as dengue [14], but mathematical models such as gravity model used in earlier studies have not been widely successful in modeling human movement accurately.

The human mobility model was built using historical pseudonymised mobile phone call detail records for Sri Lanka by quantifying human movement as an

aggregate value for a given region at a given time. Multiple input features were derived from basic input data sources and provided as an input for multiple machine learning techniques. After selecting the best performing technique, genetic algorithms were used to further improve the fit of the prediction curve by selecting the best input features that contribute towards a better prediction curve. Our work provides a comparison of the performance of contemporary machine learning techniques mainly using two metrics, the root mean squared error (RMSE) and the coefficient of determination ($R^2$). We also show from our work that human mobility is an influential factor even in dengue endemic regions. We further go on to show that our genetic algorithms based approach can be used to improve the fit of the prediction curve while reducing the error of the model as well.

## 2 Related Work

Prediction of dengue outbreaks has been the focus of multiple studies globally [2, 12, 21] as well as in Sri Lanka [5, 19]. Even though the impact of human mobility on the propagation of dengue had been established in multiple studies previously, mathematical models such as gravity model that attempt to derive human mobility patterns using regional population have not yielded good accuracy [13]. However, with the advent of increased computational capabilities and big data processing techniques, multiple research studies have utilized Mobile Network Big Data (MNBD) and particularly mobile phone Call Detail Records (CDR) as a means of deriving large scale human mobility patterns [1,6,7]. While some of these studies have already explored the applications of MNBD in domains such as disease outbreak prediction and epdimedic modeling [15, 17, 18], we did not come across any study that focus on the effect of human mobility on dengue endemic regions.

A Malaysian study [21] compared two techniques, namely Least Squares - Support Vector Machines vs Neural Networks in determining the best techique to predict dengue incidence. However, there did not exist any literature that compared more than two machine learning techniques to show which technique was best in predicting dengue outbreaks.

Another Malaysian study made use of wavelet decomposition, support vector machines and genetic algorithms to detect climatic factors that contribute towards dengue incidence [20]. This study optimizes on the RMSE, as opposed to our approach of optimizing on the $R^2$. Our study also differs in how the best set of features is selected. The above study takes a feature if it appears in 70 % of its cross validation cycles, whereas the best set of features were taken by using the best instance of the final generation of our approach.

## 3 Deriving Human Mobility using CDR

We used CDR data of nearly 10 million mobile subscribers, which is approximately half the total population of Sri Lanka, spanning for more than 1 year

from multiple mobile operators to derive a human mobility value for our prediction models. The Medical Officer of Health (MOH) division, the smallest spatial administrative unit for the health sector in Sri Lanka, for each CDR was identified in order to derive this mobility value. The home MOH division of each subscriber was identified by deriving the most frequent night time location of a given subscriber during the complete study period [8]. The number of CDRs for a given subscriber for a given time period outside his or her home MOH area was obtained. The ratio of number of CDRs outside the home MOH division against the total number of CDRs per subscriber per given week was taken as the weekly mobility of that mobile subscriber.

An assumption was made in building our model that the proportion of CDRs within a given area is proportional to the proportion of time spent within that area by a subscriber. This assumption has been used in similar studies such as the one done in Senegal [4]. After deriving the ratio of CDRs outside the home MOH division for a subscriber, the mobility value for an MOH division was derived by aggregating the mobility value of all subscribers who had visited that particular MOH division during the given time period.

If we consider M as a set of all MOH divisions, and S as a set of all subscribers, our model can be defined as follows:

$CDR(m_i, s_j, w_k)$ = No. of CDRs in MOH division $m_i$, for subscriber $s_j$ during week $w_k$ where $\forall m_i \in M, \forall s_j \in S$

Mobility of subscriber $\mathbf{s_j}$ in MOH $\mathbf{m_i}$ can be defined as

$$mob(m_i, s_j) = \frac{CDR(m_i, s_j, w_k)}{\sum_i^M CDR(m_i, s_j, w_k)} \tag{1}$$

where $\forall m_i \in \{M - Home(s_j)\}, \forall s_j \in S$

Mobility for MOH $\mathbf{m_i}$ can be defined as

$$mob(m_i) = \frac{\sum_j^N mob(s_j)}{N} \tag{2}$$

where N equals the number of subscribers that travelled to $\mathbf{m_i}$ in that week

## 4 Feature Engineering

In addition to human mobility, multiple other data sources were also used as input sources. We used weekly reported dengue cases for each MOH division provided by the Epidemiology Unit of Ministry of Health, Sri Lanka. Rainfall and temperature data was also obtained for the study period from the Integrated Surface Data of National Oceanic and Atmospheric Administration, USA [11]. The mean Normalized Difference Vegetation Index (NDVI) was derived using the

MOD13Q1 dataset from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data [10]. All input data were projected to its corresponding MOH division with a temporal scale of 1 week.

The values of previous weeks for a particular data source was also derived and provided as an input for the model. The observations were lagged by 1 to 12 weeks to obtain input features for up to 12 weeks before. Missing values due to the lagging was imputed using predictive mean matching of the MICE package in R [16].The population of each MOH division was also considered as an input feature for the model. Population data was obtained from the estimates done by the Ministry of Health, Sri Lanka and was considered to be constant throughout the study period.

## 5 Methodology

Multiple machine learning techniques were evaluated to determine which technique provides the best prediction accuracy. Four techniques were selected for comparison out of which Support Vector Regression (SVR), Neural Networks (NN) and Random Forests (RF) were selected based on the success of its use in related literature while XGBoost [3] was selected due to its recent popularity and success in many prediction problems. For evaluation of the techniques, data from 6 MOH divisions from years 2012-2014 were used. MC-Nuwara Eliya, MC-Galle, MC-Kandy, Anuradhapura, Kurunegala and Dehiwala were chosen as the regions of study for the evaluation phase due to their high dengue incidence and mobility compared to other regions. Data from the first 117 weeks of the 2012-2014 period of each MOH division was used as the training set and the final 39 weeks were used as the test set. Lagged input values from 2 weeks to 12 weeks were used for mean temperature, minimum temperature, maximum temperature, rainfall, mean NDVI and mobility while one fixed population value per MOH division was used. During the evaluation phase, the input features were fixed for all machine learning methods and differed only based on whether mobility was used as an input or not.

### 5.1 Standalone Support Vector Regression model

From the initial evaluation, SVR was shown to have the best performance. Therefore SVR was selected as the technique to carry out further work on the prediction model. For the extended study 20 MOH divisions were selected. The test set was selected to be year 2014 for 5 MOH divisions that had different characteristics. The 5 MOH divisions were MC-Colombo, Trincomalee, MC-Galle, Haputale and Batticaloa. The model was tuned using the tooling provided by the R package itself [9] by exploring feasible ranges for different hyper parameters. Hyperparameter tuning information is provided in Table 1.

**Table 1.** Hyper parameter tuning for SVR

| **Parameter** | **Tuning Range** | **Optimum Value** |
|---|---|---|
| Type | $\epsilon$-regression, $\nu$-regression | $\nu$-regression |
| Kernel | Radial, Polynomial, Linear | Radial |
| $\gamma$ | 0.001 - 0.1 | 0.004 |
| $\nu$ | 0.1 - 0.8 | 0.35 |
| Cost | 1 - 10 | 3 |

### 5.2 Integrating Genetic Algorithms

The support vector regression models developed were able to provide comparative RMSE and $R^2$ measures after tuning. However, some MOH divisions such as MC-Colombo provided much lower accuarcy when compared to the overall average. A Genetic Algorithms (GA) based approach was integrated to the training phase of the SVR model to improve the $R^2$ measure. The models developed using only support vector regression used all of the available engineered input features where some of features might contribute towards the adding noise to the model. The initution behind incorporating GA was to select the best features that contribute towards increasing the $R^2$ of the model. Input features were represented as a binary chromosome and the fitness function was designed to minimize the $R^2$ of the model after a support vector regression model was trained optimizing for RMSE using the selected input features. The population size was selected to be 100 after experimenting with different population sizes and the the GA model was trained for 50 generations. Crossover probability was set at 0.8 while the mutation probability was set at 0.1.

## 6 Results

After deriving the input features and performing feature engineering, distance correlation was used as a measure to determine the dependence between dengue incidence and other input features. The correlation graph for mobility and other input features is depicted in Fig. 1. This graph shows significant correlation between the mobility value derived using our model and dengue incidence. Also high correlation is shown between mean NDVI and dengue incidence.

The results from the evaluation phase detailed in Table 2 show that the model performance increased for each technique for the evaluated dataset when mobility was introduced to the model. While the improvement is minimal in some techniques, overall consistency in improvement of the model performance when mobility is introduced suggests that mobility is an influential factor in predicting dengue outbreaks.

The results additionally show us that support vector regression has the best performance when considering both the metrics, RMSE and $R^2$.

After making use of the data from 20 MOH divisions and performing hyper paramter tuning in order to build a more generic model, we were able to obtain

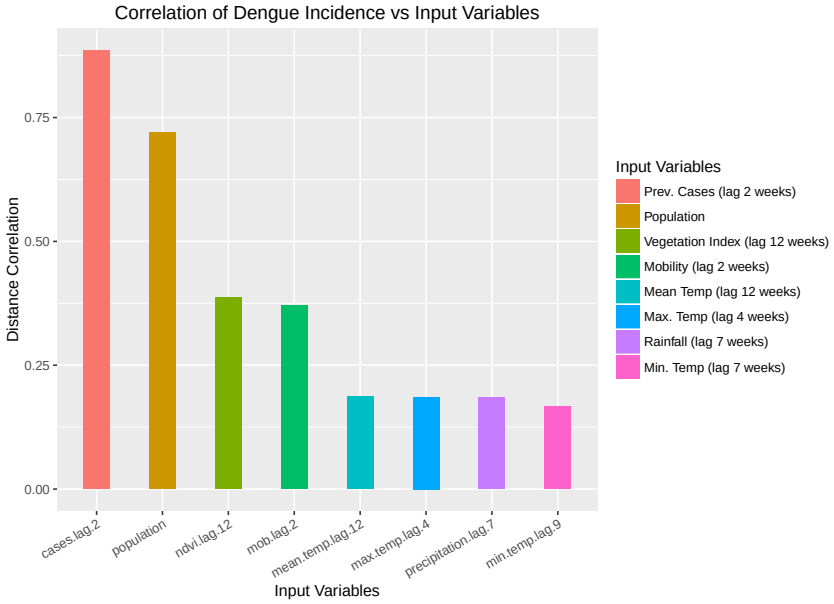


**Fig. 1.** Correlation of Input Variables with Dengue Incidence

**Table 2.** RMSE and $R^2$ values of different machine learning methods

| **Model** | **Tuning Parameters** | **RMSE** | | **$R^2$** | |
|---|---|---|---|---|---|
| | | **- Mobility** | **+ Mobility** | **- Mobility** | **+ Mobility** |
| Random Forests | Max. Nodes = 5, n-trees = 120 | 6.907 | 6.812 | 0.628 | 0.639 |
| Neural Networks | Hidden = 3, Err. = SSE, Act. Func = logistic | 10.966 | 9.239 | 0.063 | 0.335 |
| XGBoost | Max. Depth = 4, $\eta$ = 0.05, n-folds = 4 | 6.892 | 6.794 | 0.63 | 0.64 |
| SVR | Kernel = Radial, $\nu$ = 0.3, Cost = 5 | 6.408 | 6.17 | 0.68 | 0.704 |

an overall RMSE of 10.005 with an $R^2$ value of 0.891 for the final model. The performance of the model for each MOH division in the test set is given in Table 3. The application of the genetic algorithm based optimization resulted in overall improvement to the final model. As an example, the RMSE and $R^2$ value for the year 2014 for MC-Colombo was 21.99 and 0.502 respectively before applying the optimisation. After GA optimization was applied, the RMSE value was reduced to 18.385 while $R^2$ was increased to 0.652.

**Table 3.** RMSE and $R^2$ values for different MOH divisions *(Negative $R^2$ values were not included)*

| MOH | RMSE | | $R^2$ | |
|---|---|---|---|---|
| | Without GA | GA Optimized | Without GA | GA Optimized |
| MC-Colombo | 21.99 | 18.385 | 0.502 | 0.652 |
| Haputale | 1.924 | 1.61 | - | - |
| MC-Galle | 3.752 | 3.579 | - | - |
| Batticaloa | 3.842 | 3.797 | 0.212 | 0.23 |
| Trincomalee | 1.784 | 1.864 | - | - |

## 7 Discussion

This work introduces a human mobility model derived from CDR data that can be applied to multiple machine learning techniques directly. A measure is derived for a given spatial region for a given temporal scale and we have shown that this derived measure shows significant correlation with dengue incidence in the context of Sri Lanka. The improvement did vary when this mobility measure was applied to different machine learning techniques, sometimes showing more than 4 times improvement in $R^2$ for Neural Networks, while showing only a 1.59 % increase in $R^2$ for XGBoost. However, there is a consistent improvement in all the considered techniques when mobility is introduced, which corroborates the high correlation shown with dengue incidence.

We have been able to select the best technique by doing a quantitative comparison for the two metrics RMSE and $R^2$ used in the evaluation phase. However, we cannot observe a significant qualitative difference in performance between the different techniques. The number of input features were kept fixed in the evaluation phase to maintain consistent input features across the evaluated machine learning methods. It should be noted that the high dimensionality of the input would affect some machine learning techniques such as Neural Networks significantly while not becoming a critical factor for other techniques such as Random Forests and XGBoost. In the case of Neural Networks, the number of hidden layers and the number of nodes within a layer were determined in order to reduce the introduction of noise due to high dimensionality as much as possible, we would still have to consider the issue of dimensionality when factoring these results objectively. High dimensionality can also explain the relatively poor performance of Neural Networks when compared to the other 3 techniques that were considered. A genetic algorithms based optimization for Neural Networks similar to what was applied for SVR can be considered in future work as a technique to reduce the errors due to dimensionality.

The genetic algorithms based optimization was developed due to the fact that some MOH divisions like MC-Colombo, which has the highest dengue incidence in the country, did not have a good prediction performance, which is reflected in the prediction graph for MC-Colombo in Fig. 2. However, for the purpose of applying this work to execute dengue preventive measures, it is important to

obtain a prediction curve that matches the actual epidemic curve closely. The reason is that epidemiological trend or the possibility of a potential outbreak would be of more interest to public health sector officials rather than the exact number of predicted dengue cases for a given region. Therefore, we explored possible optimizations to improve the $R^2$ of predictions for regions where the model is performing poorly. The GA optimization technique used in our work optimizes $R^2$, while the SVR training optimizes RMSE, thereby ensuring that the model evolution optimizes for both $R^2$ and RMSE, which is observable in the improved prediction graph (Fig. 3) for MC-Colombo.

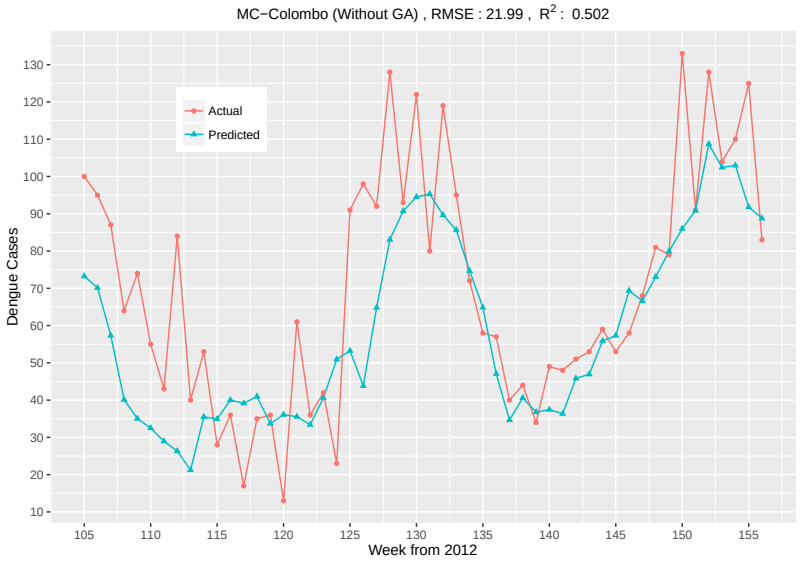


**Fig. 2.** Predicted vs Actual for MC-Colombo (without GA optimization)

## 8 Conclusion

We have shown that human mobility contributes to dengue incidence even in regions where the disease is endemic. We also hope that the methodology introduced in building our mobility model can be extended to build more complex human mobility models that reflects the contribution of mobility to dengue incidence more accurately. The evaluation of different machine learning technique provides direction on which machine learning technique is most suited in the context of dengue incidence prediction in Sri Lanka. The genetic algorithms based optimization technique was shown to provide significant improvements to prediction performance. Our work can also be extended to apply the GA optimization technique to other machine learning methods and obtain a similar comparison to determine whether the GA technique will affect the final ranking between different machine learning methods evaluated in this study.

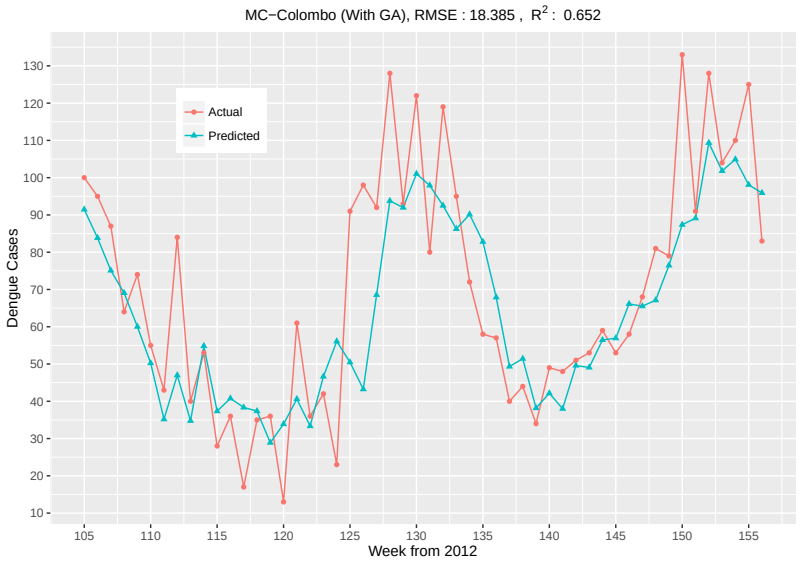


**Fig. 3.** Predicted vs Actual for MC-Colombo (with GA optimization)

## Acknowledgments

The authors would like to thank the Dr. Hasitha Tissera and Dr. Azhar Ghouse of the Epidemiology Unit, Ministry of Health, Sri Lanka for providing dengue incidence data as well as expert knowledge on the disease dynamics and the epidemiology of dengue. We would also like to thank Amila De Silva from the University of Moratuwa for deriving the mean vegetation index values using MODIS satellite data. This research was funded through a grant from the International Development Research Centre (IDRC) of Canada and a grant from the Senate Research Committee (SRC) of University of Moratuwa.

## References

1. Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., Rebaudet, S., Piarroux, R.: Using mobile phone data to predict the spatial spread of cholera. Scientific reports 5, 8923 (2015), http://www.ncbi.nlm.nih.gov/pubmed/25747871
2. Chen, C.C., Chang, H.C.: Predicting dengue outbreaks using approximate entropy algorithm and pattern recognition. Journal of Infection 67(1), 65–71 (2013), http://dx.doi.org/10.1016/j.jinf.2013.03.012
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. KDD '16, ACM, New York, NY, USA (2016)
4. Finger, F., Genolet, T., Mari, L., de Magny, G.C., Manga, N.M., Rinaldo, A., Bertuzzo, E.: Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proceedings of the National Academy of Sciences 113(23), 201522305 (2016)

5. Herath, N., Perera, A.A.I., Wijekoon, H.P.: Prediction of dengue outbreaks in Sri Lanka using artificial neural networks. International Journal of Computer Applications 101(15), 1–5 (2014), http://research.ijcaonline.org/volume101/number15/pxc3898862.pdf
6. Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., Willinger, W.: Human mobility modeling at metropolitan scales. Proceedings of the 10th international conference on Mobile systems, applications, and services - MobiSys '12 p. 239 (2012), http://dl.acm.org/citation.cfm?id=2307659
7. Jiang, S., Ferreira, J., González, M.C.: Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data : A Case Study of Singapore. ACM KDD UrbComp'15 pp. 1–13 (2015)
8. Maldeniya, D., Lokanathan, S., Kumarage, A.: Origin-Destination Matrix Estimation for Sri Lanka Using 2 . the Four Step Model. Proceedings of the 13th International Conference on Social Implications of Computers in Developing Countries (May), 785–794 (2015)
9. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2015), https://CRAN.R-project.org/package=e1071, r package version 1.6-7
10. NASA LP DAAC: MOD13Q1 — Vegetation Indices 16-Day L3 Global 250m (2016), https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod13q1
11. National Centers for Environmental Information (NCEI), NOAA: Integrated Surface Database (ISD), https://www.ncdc.noaa.gov/isd
12. Rachata, N., Charoenkwan, P., Yooyativong, T., Chamnongthai, K., Lursinsap, C., Higuchi, K.: Automatic prediction system of dengue haemorrhagic-fever outbreak risk by using entropy and artificial neural network. 2008 International Symposium on Communications and Information Technologies, ISCIT 2008 (Iscit), 210–214 (2008)
13. Sarzynska, M., Udiani, O., Zhang, N.: A study of gravity-linked metapopulation models for the spatial spread of dengue fever. arXiv preprint arXiv:1308.4589 2008, 1–32 (2013), http://arxiv.org/abs/1308.4589
14. Stoddard, S.T., Morrison, A.C., Vazquez-Prokopec, G.M., Soldan, V.P., Kochel, T.J., Kitron, U., Elder, J.P., Scott, T.W.: The Role of Human Movement in the Transmission of Vector-Borne Pathogens. PLoS Negl Trop Dis 3(7) (2009)
15. Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C.M., Blondel, V., Smoreda, Z., Gonzlez, M.C., Colizza, V.: On the use of human mobility proxies for modeling epidemics. PLOS Computational Biology 10(7), 1–15 (07 2014)
16. van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 45(3), 1–67 (2011), http://www.jstatsoft.org/v45/i03/
17. Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., Buckee, C.O.: Quantifying the impact of human mobility on malaria. Science (New York, N.Y.) 338(6104), 267–70 (oct 2012)
18. Wesolowski, A., Qureshi, T., Boni, M.F., Sundsøy, P.R., Johansson, M.A., Rasheed, S.B., Engø-Monsen, K., Buckee, C.O.: Impact of human mobility on the emergence of dengue epidemics in pakistan. Proceedings of the National Academy of Sciences 112(38), 11887–11892 (2015)
19. Wickramaarachchi, W.P.T.M., Perera, S.S.N., Jayasinghe, S.: Modelling and analysis of dengue disease transmission in urban Colombo: A wavelets and cross wavelets

approach. Journal of the National Science Foundation of Sri Lanka 43(4), 337–345 (2016)
20. Wu, Y., Lee, G., Fu, X.J., Hung, T.: Detect climatic factors contributing to dengue outbreak based on wavelet, support vector machines and genetic algorithm. World Congress on Engineering 2008 Vols IIi 1, 303–307 (2008)
21. Yusof, Y., Mustaffa, Z.: Dengue Outbreak Prediction : A Least Squares Support Vector Machines Approach. International Journal of Computer Theory and Engineering 3(4), 489–493 (2011)