IMPROVING DISEASE OUTBREAK FORECASTING MODELS FOR EFFICIENT TARGETING OF PUBLIC HEALTH RESOURCES

Fernando, Lasantha;Perera, Amal S;Lokanathan, S;

;

IDRC Grant/ Subvention du CRDI: 108008-001-Leveraging Mobile Network Big Data for Developmental Policy

Annex 12: Improving Disease Outbreak Forecasting Models for Efficient Targeting of Public Health Resources

Lasantha Fernando, Sriganesh Lokanathan
LIRNEasia

Amal Shehan Perera
University of Moratuwa

Azhar Ghouse, Hasitha Tissera
Epidemiology Unit of Sri Lanka

Introduction

Dengue is estimated to have approximately 390 million infections annually out of which an estimated 96 million manifest (Bhatt et al., 2013). WHO estimates almost half the global population to be at risk from this neglected tropical infectious disease. The enormous economic burden of the disease is evident when considering that an estimated 264 disability-adjusted life years (DALYs) per million population is lost due to dengue each year (World Health Organization, 2012). In Sri Lanka, it was estimated that within the Colombo district, where the nation's capital is situated, a financial burden of US$ 971,360 was imposed upon the national health system in 2012 alone just for the execution of preventive measures (Thalagala et al., 2016). Considering all of these factors, optimizing resource allocation and reducing the economic burden of dengue should be a key element of any long term strategy for dealing with the disease. In this context, the ability to predict dengue outbreaks for a particular region 2 weeks in advance would lead to better resource mobilization and would be invaluable asset for the public health sector.

Additionally, a disease outbreak forecasting model developed for dengue need not be limited to tackling resource allocation for that disease only, but can also be used to forecast outbreaks of other arboviral diseases such as Ebola, Zika or Chikungunya. In Sri Lanka at least, Chikungunya is already prevalent and it is just a matter of time before Zika comes to Sri Lanka given that cases have been detected in the Asian region including Singapore. In an epidemic or outbreak of an infectious disease, we would need to develop a deep understanding of human mobility patterns within the infected regions to identify the potential hotspots and also to forecast to which regions are most likely to be infected next. For the purpose of understanding human mobility in disease propagation, Mobile Network Big Data (MNBD) has become a low cost data exhaust that provides rich insight into human mobility patterns with better spatial and temporal granularity when compared to statistical methods which rely mostly on macro level population parameters.

In this work, we evaluate multiple machine learning techniques such as Neural Networks (NN), Support Vector Machines (SVM), Random Forests and XGBoost to determine which technique performs best. A comparison of the model performance between different techniques is provided. We go on to use a genetic algorithm based optimization to further improve the accuracy of these models. Our work shows that Call Detail Records (CDR) can be used to derive proxy indicators for human movement patterns which are applicable across multiple machine learning models. Our results show that human mobility has an impact on dengue incidence, even in dengue endemic regions. The forecasting models developed in this work can be utilized to effect a significant impact on the issue of allocating resources effectively to combat dengue which in turn would lead to reduced economic burden as well as reduced mortality and morbidity.

Literature Review

Prediction of dengue outbreaks has been the focus of multiple studies globally (Hales, de Wet, Maindonald, & Woodward, 2002; Rachata et al., 2008; Chen & Chang, 2013) as well as in Sri Lanka (Wickramaarachchi, Perera, & Jayasinghe, 2016; Herath, Perera, & Wijekoon, 2014). The impact of human mobility on the propagation of dengue had also been established in multiple studies previously (Stoddard et al., 2009). However, using mathematical models to derive human mobility patterns using regional characteristics in order to forecast disease outbreaks, such as the methodology adopted by Sarzynska et. al (2013), have not yielded good accuracy. With the advent of increased computational capabilities and big data processing techniques, multiple research studies have utilized MNBD and particularly mobile phone CDRs as a means of deriving large scale human mobility patterns (Isaacman et al., 2012; Bengtsson et al., 2015; Jiang, Ferreira, & González, 2015) and some of these studies have even explored the application of MNBD in disease outbreak prediction and epidemic modeling (Tizzoni et al., 2013; Wesolowski et al., 2012; Wesolowski et al., 2015). We did not come across any study that used MNBD to derive proxy indicators for human movement patterns to determine its impact on dengue endemic regions.

Also, we could find only a few examples of existing literature that provides comparison between the performance of different machine learning techniques when building disease forecasting models for a vector-borne infectious disease such as dengue. A Malaysian study (Yusof & Mustaffa, 2011) compared two techniques, namely Least Squares - Support Vector Machines vs. Neural Networks to determine the best technique to predict dengue incidence. There was another Malaysian study that made use of wavelet decomposition, SVMs and GAs to detect climatic factors that contribute towards dengue incidence (Wu, Lee, Fu, & Hung, 2008). The main differences between the above study and ours is on the approach to optimizing the output of the genetic algorithm and selection of input features. While our work focuses on optimizing the coefficient of determination of the final output model (R^2), the above study used the Root Mean Squared Error (RMSE) as the metric to optimize the model. Our study also differs in how we consider the input features of the best instance of the final generation from GA as the input features for the final model, whereas the above study takes a feature if it appears in 70% of its cross validation cycles.

In a study done by Lessler and Cummings (2016), the authors attribute the first use of mathematical and mechanistic models in epidemiology and public health to Lowell Reed and Wade Hampton Frost, when they first made use of these models to teach epidemic theory to students back in 1930-1940s. According to them, mechanistic models on infectious diseases have been used as an integral tool in planning and implementing public health responses from that time onwards. In the above study, the authors surmise that in this era of big data, the models would likely be directly integrated to data at both population scales and individual within hosts scales. Our work is a direct example to validate this trend becoming established in the domain of epidemiological modeling, where data is integrated at an individual scale and fed as an input for the model after being aggregated to a population scale. Metcalf, Edmunds, and Lessler (2015), identifies one of the key challenges in modeling for public health policy as reducing the divide between modelers and policy makers. In our work, we present several recommendations in addressing this challenge based on the collaborative partnership that was established to carry out the work described in this study.

Data Sources

This paper uses CDR data spanning more than 1 year for nearly 10 million SIMs from

multiple mobile operators in Sri Lanka to derive aggregate human mobility patterns for the dengue forecasting model. The data is completely pseudonym-ized by the operator and the researchers do not maintain any mapping information between the generated identifier and the original phone number.

We used weekly reported dengue cases for each Medical Officer of Health (MOH) division, which is the smallest health administrative unit for Sri Lanka, 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 (NCEI, 2016). The mean Normalized Difference Vegetation Index (NDVI) was derived using the MOD13Q1 dataset from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data (NASA Land Data Products, 2016). 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 (van Buuren & Groothuis-Oudshoorn, 2011).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.

Research Methodology

The research methodology was focused on iteratively improving the model on two key components of the study. One was the development of a human mobility model using MNBD that accurately models the aggregate human population movements in Sri Lanka for the study period. The other was determining which machine learning technique gives the best disease forecasting model. The predictive models itself incorporated the values derived from human mobility model as well as the pre-processed data from the other available data sources. An MOH division was taken as the spatial unit to base the models upon. Therefore all input features from different data sources were projected to its corresponding MOH division. The temporal unit for our predictions was taken as a week, and similar scaling was done to project the input features to its weekly values as well.

Developing a Human Mobility Model

Using big data techniques to process records of more than 10 million mobile SIMs, an aggregate human mobility value for an MOH division was derived using the CDR dataset. Initially, the home MOH location of a subscriber was identified by considering the most frequent number of CDR records for a given subscriber during the time period from 9.30 PM to 5.30 AM, which was based on previous literature (Lokanathan et. al., 2014). After identifying the home MOH of each subscriber, his/her mobility was calculated as the proportion of CDR records occurring outside the home MOH in comparison to the total number of CDR records for a given week. The underlying assumption was that the fraction of calls initiated or received outside of the home MOH division is proportional to the fraction of actual time spent outside, which is similar to the assumption used by Finger et. al. (2016) in a study in Senegal. The mobility of an MOH division was taken as the average mobility of all the visitors (subscribers who are not residents of that MOH) for an MOH division during a given week.

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:

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

Mobility of subscriber s_j at MOH m_i can be defined as

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

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

Mobility for MOH m_i can be defined as

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

$where\ N\ is\ the\ no.of\ subscribers\ travelled\ to\ m_i\ for\ that\ week$

Correlation Analysis

In order to determine the impact of human mobility on dengue incidence when compared to other parameters such as rainfall, temperature, past dengue cases and vegetation index, we performed a correlation analysis on the different input features. Since there is no evidence to conclude that the relationship between these variables and dengue incidence will be linear, we decided to make use of two correlation measurements that were known to account for non linear relationships as well. The measurements were distance correlation, which takes the Euclidean norm to calculate pair wise distance between observations to obtain variance, standard deviation and covariance and mutual information estimate, which makes use of information entropy to determine how much knowing of one variable reduces the uncertainty of the other variable under consideration. The results from these correlation analysis are presented in Table 1 under Results & Discussion section.

Prediction Models

The focus of this study was on using Machine Learning methods to develop the prediction models as opposed to purely statistical methods. Four machine learning techniques were selected for comparison, which were Support Vector Regression (SVR), Neural Networks (NN), Random Forests and XGBoost.

SVR uses an appropriate kernel function such as a Gaussian function or a polynomial function to transform the dataset to higher dimensional feature space where a good linear regression model can be obtained to fit the data points in that high dimension feature space. NNs are modeled by having interconnected layers of weighted nodes determining what the output should be where as RF is a decision tree based classification/regression technique. All these techniques were chosen for evaluation based on the success of its use in related literature while XGBoost (T. Chen & Guestrin, 2016), which is another decision tree based algorithm, was selected due to its recent popularity and success in many prediction problems.

In order to evaluate the performance of different techniques, initially, data from 6 MOH divisions from years 2012-2014 were used. The 6 MOH divisions were chosen due to the fact that all of them had high dengue incidence and mobility compared to other regions. The chosen MOH divisions for the evaluation phase were MC-Nuwara Eliya, MC-Galle, MC-Kandy, Anuradhapura, Kurunegala and Dehiwala. Data for the 156 weeks during 2012-2014

period was separated into a training set and a test set. A validation set was not prepared since only a limited number of data points were available for a single MOH division. Data from the first 117 weeks 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. The same starting set of input features were used to evaluate all the machine learning methods and the performance was measured for each technique with mobility as well as without mobility in order to quantify the improvement to model accuracy due to the introduction of human mobility as an input feature. Hyper parameter tuning was also done for each technique to ensure that no unfair advantage was given to a particular method due to improperly tuned models. The hyper parameter tuning information is also provided in Table 2 in the Results & Discussion section.

For the second phase of the study, 14 more MOH divisions were selected in addition to the initial 6 MOH divisions making a total of 20 MOH divisions. The test set was selected to be the data points for the complete year of 2014 for 5 MOH divisions that had different characteristics. They were MC-Colombo, MC-Galle, Trincomalee, Haputale and Batticaloa. The model was tuned using the tooling provided by the e1071 R package (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2015) in the case of SVR and by evaluating a range of feasible values for other techniques. The models developed in this phase were able to provide comparative RMSE and R^2 measures after tuning. However, some MOH divisions such as MC-Colombo provided much lower accuracy when compared to the overall average. Therefore, further optimization was done by introducing a GA based approach to the training phase of each model that specifically targeted improving the R^2 measure.

For the GA based optimization, input features were represented as a binary chromosome and the fitness function was designed to minimize the R^2 of the model. Since all the machine learning techniques were trained to optimize the RMSE measure, this effectively created a training process that optimizes RMSE and then R^2 in an alternating fashion. The population size of a single generation was selected to be 100 after experimenting with different population sizes and the GA model was trained for 50 generations. Crossover probability was set at 0.8 while the mutation probability was set at 0.1.

Results & Discussion

The human mobility model developed in our work provides a single representative value for an entire MOH division for a given week, which is needed for the output from the mobility model to be integrated directly into different machine learning methods. In contrast, if mobility was derived as a parameter that represents movement from one location to another, which is might be more intuitive when considering human movement patterns, further processing would be needed to apply that output to different forecasting models.

After getting the output from the human mobility model, the results from the correlation analysis is able to provide us with an idea about how much the derived mobility value is correlated with dengue incidence. It also provides a mechanism to validate the approach taken to model human mobility. The distance correlation and mutual information of derived mobility values and other input features against dengue incidence is shown in Table 1. The values show significant correlation between the mobility value derived using our model and dengue incidence. Also high correlation is shown between mean NDVI and dengue incidence.

Input Variable	Distance Correlation	Mutual Information
case.lag.1.week	0.913	0.419
mobility.no.lag	0.406	0.198
mean.ndvi.no.lag	0.395	0.128
precip.lag.11.week	0.180	0.051
mean.temp.lag.10.week	0.175	0.124
max.temp.lag.4.week	0.168	0.108
min.temp.lag.12.week	0.150	0.089

Table 1 - Correlation of Input Variables vs Dengue Incidence

The initial results from the first phase where training models incorporated 6 MOH divisions from Sri Lanka, with and without mobility as an input feature, are shown in Table 2. RMSE and R^2 were both selected as metrics for measuring model performance because we needed our models to forecast the number of dengue cases accurately as well as identify the general trends and peaks of the epidemic curve. RMSE gives a good idea about the accuracy of individual predictions, while R^2 gives an idea about how well the prediction fits the actual epidemic curve. In a region where the number of dengue cases are relatively low, we might get a low RMSE value, but still predict incorrect peaks when compared to the actual epidemic trend. A model that performs as such will be captured by using R^2 as an additional metric to measure model performance.

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

Table 2 - RMSE, R^2 Comparison for different Machine Learning Methods

Values given in Table 2 show the show us that SVR had the best performance for the initial 6 MOH divisions when considering both the metrics, RMSE and R^2. Additionally, it is evident that the introduction of mobility as an input feature improves the performance of the model, even though the improvement is marginal in most cases.

However, as mentioned in the previous section, the models did not perform well for some MOH divisions when the dataset was expanded to 20 MOH divisions. With the introduction of GA based optimization, we were able to obtain significant improvements to model performance in all of the machine learning methods. After performing hyper parameter tuning in order to build a more generic predictive model in addition to the GA based feature selection, we were able to obtain an RMSE value of 7.852 and an R^2 value of 0.933 for overall predictions of the 5 MOH divisions in the test set of the second phase. The best performance for the final dataset was obtained by using XGBoost as the machine learning

method. The improvement in predictive accuracy for each technique due to the GA optimization is shown in Table 3 while the details of the final model are given in Table 4.

Machine Learning Technique	Without GA Optimization		With GA Optimization	
	RMSE	R^2	RMSE	R^2
Random Forests	9.746	0.896	8.258	0.926
NN	26.495	0.235	12.154	0.839
XGBoost	9.556	0.9	7.852	0.933
SVR	10.191	0.887	8.618	0.919

Table 3 - RMSE, R^2 Comparison with & without GA optimization

Parameter	Value
Machine Learning Technique	XGBoost
Study Period	2012 to 2014
Training Set	2012-2014 data of Anuradhapura, Badulla, Seeduwa, Rathnapura, Vavuniya, Dehiwala, Hambantota, MC-Jaffna, MC-Kandy, Kurunegala, Thambuttegama, Mannar, MC-Nuwara Eliya, Pothuvil and Puttalam 2012-2013 data for MC-Colombo, MC-Galle, Trincomalee, Haputale and Batticaloa
Test Set	2014 data for MC-Colombo, MC-Galle, Trincomalee, Haputale and Batticaloa
Best RMSE (Overall Model)	7.852
Best R^2 (Overall Model)	0.933

Table 4 - Parameter values for the final model

The application of the GA based optimization resulted in improved predictions for critical MOH divisions such as MC-Colombo (where the highest dengue incidence is seen annually) as well. The RMSE and R^2 value for the year 2014 for MC-Colombo was 20.401 and 0.571 respectively before applying the optimization. After GA optimization was applied, the RMSE value was reduced to 16.476 while R^2 was increased to 0.72 for XGBoost. The better fit of the predicted curve compared to the actual epidemic curve after GA optimization was applied can be observed in Fig. 1 and Fig. 2 below.

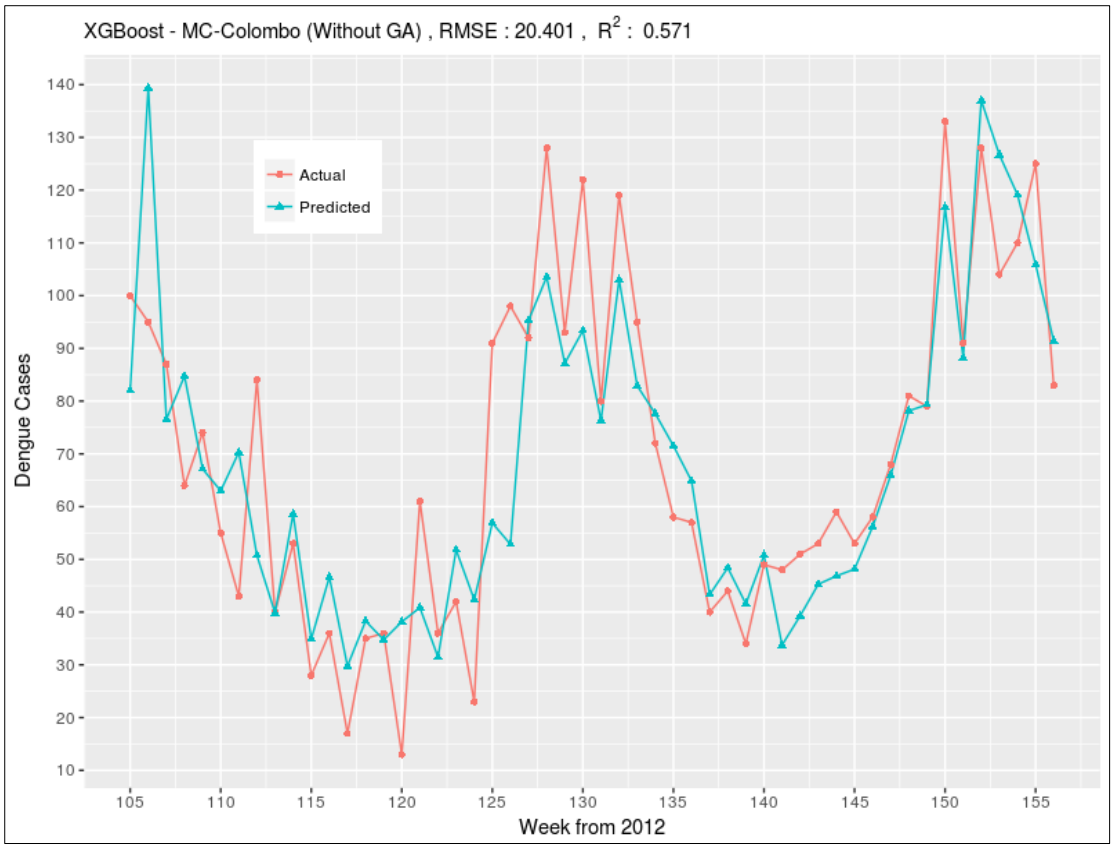

Figure 1. Dengue Incidence (Predicted vs. Actual) without GA optimization

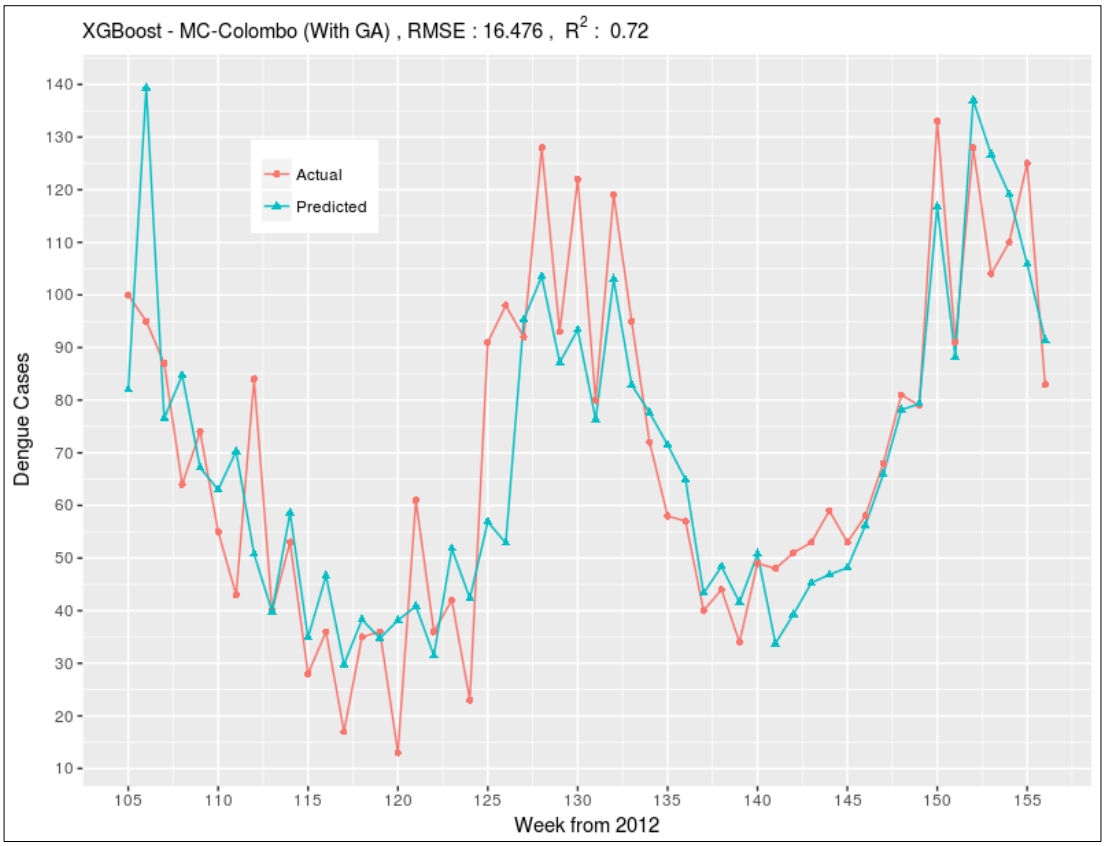

Figure 2. Dengue Incidence (Predicted vs. Actual) using GA optimization

Considering the impact on public health sector from this research work, significant reduction

in financial costs, mortality and morbidity of the disease can be achieved by integrating this forecasting model to public health sector resource allocation. With a per capita health expenditure of LKR 7,497 in 2014 and a population of 1,179 per medical officer, the Sri Lankan public health system would be extremely resource constrained to deal with a large dengue epidemic that occurs annually or biannually. Sri Lanka uses an integrated vector control approach that makes use of biological control, chemical control in addition to other environmental measures. However, chemical control is applied after an outbreak occurs due to cost, potential toxicity and death to other organisms. If we can predict outbreaks in advance, vector control mechanisms can be applied beforehand to prevent or reduce the potential for an outbreak.

Additionally, if high risk regions can be identified beforehand, public awareness campaigns can be launched, which can significantly change the dynamics of the spread of the disease given that proper action is taken by the public. Also, the epidemiology unit of Sri Lanka conducts routine campaigns to ensure that potential mosquito breeding sites are cleaned up and maintained. These campaigns are conducted in targeted areas with the participation of security forces in the country as well due to limited resource availability. In such a context, information on where to target these campaigns would be immensely beneficial, and the forecasting model described in this paper can be developed further to provide this required information with reasonable accuracy.

Challenges & Future Work

One of the biggest challenges in developing the prediction models for this work lied in the fact that different data sources were in different temporal and spatial scales and it was not always possible to project these values to the units chosen for this study. For example, mobility data was available at a granularity that corresponds to the coverage area of a Base Transceiver Station (BTS) of a mobile operator. In order to get the corresponding mobility value for an MOH division, the values needed to be recalculated based on the area of overlap between an MOH division and a BTS coverage area.

Additionally, the MOH divisional boundaries changed during the study period. For example, the number of MOH divisions in Colombo district increased from 12 to 15 where larger MOH divisions were split into smaller multiple MOH divisions. These changes had to be factored in when preparing the dataset for the study.

High dimensionality of the data after feature engineering was another issues that would have affected the performance of some models. This can be one of the reasons for the relatively poor performance of NNs along with its inherent difficulty in determining the exact number of nodes and layers needed for an accurate model. A GA based optimization for selecting the number of layers and nodes in NNs similar to what was applied for SVR can be considered in future work as a technique to reduce the errors due to dimensionality.

While the study focuses on building regression models for each MOH division, when implementing it practically, the public health sector would benefit from a prediction of probable risk bands for a given MOH division rather than predicting the exact number of cases that will be reported. Therefore, we plan to develop a classification model that forecasts the risk band at the implementation phase of this work. A risk prediction that gives a disease outbreak risk value from a scale of 1-5 would be easier to act upon than an estimate of the number of cases that might be reported for purposes of public health sector resource allocation.

We are also working on developing a human mobility model that considers the risk of

infection according to the time and place of a CDR record of an individual mobile subscriber and use that information to come up with a human movement pattern model that correlates even better with the dengue incidence for a given time and region.

Conclusion & Policy Recommendations

This work introduces a human mobility model derived from CDR data that can be applied to multiple machine learning techniques directly. While our work shows that the introduction of mobility improves the prediction accuracy, the improvement varies according to the machine learning technique that is applied. Neural Networks showed an improvement of more than 4 times for R^2, while only a 1.59 % increase was shown for XGBoost. However, the improvement is consistent, and considering the high correlation of mobility with dengue incidence as well, we can conclude that human movement patterns influence dengue incidence, even when the disease is already endemic. We have also shown that the methodology introduced in building our mobility model can be used to provide disease outbreak forecasts, validating MNBD as a viable and influential data source in disease outbreak prediction.

In addition to introducing a methodology that can be used to predict dengue outbreaks for different MOH divisions in Sri Lanka 1 to 2 weeks in advance, our work also provides a comparison of the suitability of different machine learning techniques when applied to the specific domain of disease outbreak prediction. The methods discussed in this paper can be utilized directly to predict where the next outbreak will occur and apply control mechanisms appropriately. If the techniques introduced in this study are implemented by the relevant policy makers in the public health sector, it should yield significantly more efficient resource allocation in the disease prevention and control strategies at not just the national level but more importantly at the regional level.

Also, it is imperative that researchers provide outcomes in an easily actionable form that helps public health sector officials act upon it with little turnaround time as possible. It is in this regard that a risk classification model is also being developed as part of future work where the output will be one of predefined risk levels. Such an output can be translated directly to a decision to execute preventive measures or not, allowing the public health sector officials to consume and utilize the prediction output more effectively.

This study is very much a product of different skills and requires expertise from several different domains such as computer science, statistics, epidemiology, and also entomology. The co-authors come from three different organizations, LIRNEasia, University of Moratuwa, and the Epidemiology Unit of the Ministry of Health, reflecting the cross-disciplinary nature of this work. Effective mainstreaming of such techniques to bring efficient allocation of public health resources therefore will require multi-disciplinary and multi-stakeholder teams. This is because all these skills are often not available amongst a single stakeholder. Obviously capacity development is a priority, but these are specialized skills and it will take time for these to be mainstreamed or available within any one single organization. Therefore practical and timely applications of such techniques requires innovative partnerships.

References

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., … Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports*, *5*, 8923.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., … Jaenisch, T. (2013). The global distribution and burden of dengue. *Nature*, *496*(7446), 504–507.

Chen, C. C., & Chang, H. C. (2013). Predicting dengue outbreaks using approximate entropy

algorithm and pattern recognition. *Journal of Infection*, *67*(1), 65–71.
Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD*, 1–10.
Finger, F., Genolet, T., Mari, L., de Magny, G. C., Manga, N. M., Rinaldo, A., & Bertuzzo, E. (2016). Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. *Proceedings of the National Academy of Sciences*, *113*(23), 6421-6426.
Hales, S., de Wet, N., Maindonald, J., & Woodward, A. (2002). Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. *Lancet*, *360*, 830–834.
Herath, P. H. M. N., Perera, A. A. I., & Wijekoon, H. P. (2014). Prediction of dengue outbreaks in Sri Lanka using artificial neural networks. *International Journal of Computer Applications*, *101*(15), 1–5.
NOAA/NCEI Integrated Surface Database (ISD). (2016). Retrieved April 20, 2017, from https://www.ncdc.noaa.gov/isd
Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., & Willinger, W. (2012). Human mobility modeling at metropolitan scales. *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services - MobiSys '12*, 239.
Jiang, S., Ferreira, J., & González, M. C. (2015). Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data : A Case Study of Singapore. *ACM KDD UrbComp'15*, 1–13.
Lessler, J., & Cummings, D. A. T. (2016). Commentary Mechanistic Models of Infectious Disease and Their Impact on Public Health, *183*(5), 415–422.
Lokanathan, S., Kreindler, G. E., Silva, N. H. N. De, Miyauchi, Y., Dhananjaya, D., & Samarajiva, R. (2014). The Potential of Mobile Network Big Data as a Tool in Colombo ' s Transportation and Urban Planning. *Information Technologies & International Development*, *12*(2), 63–73.
Metcalf, C. J. E., Edmunds, W. J., & Lessler, J. (2015). Six challenges in modelling for public health policy. *Epidemics*, *10*, 93–96.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. Retrieved from https://cran.r-project.org/package=e1071
Myers, M. F., Rogers, D. J., Cox, J., Flahault, A., & Hay, S. I. (2000). Forecasting disease risk for increased epidemic preparedness in public health, 309–330.
Nasa Land Data Products. (2016). Vegetation Indices 16 - Day L3 Global 250m, 1–4. Retrieved from https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod13q1
Rachata, N., Charoenkwan, P., Yooyativong, T., Chamnongthai, K., Lursinsap, C., & Higuchi, K. (2008). 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.
Sarzynska, M., Udiani, O., & Zhang, N. (2013). A study of gravity-linked metapopulation models for the spatial spread of dengue fever. *arXiv Preprint arXiv:1308.4589*, *2008*, 1–32. Retrieved from http://arxiv.org/abs/1308.4589
Stoddard, S. T., Morrison, A. C., Vazquez-Prokopec, G. M., Soldan, V. P., Kochel, T. J., Kitron, U., … Scott, T. W. (2009). The Role of Human Movement in the Transmission of Vector-Borne Pathogens. *PLoS Negl Trop Dis*, *3*(7).
Thalagala, N., Tissera, H., Palihawadana, P., Amarasinghe, A., Ambagahawita, A., Wilder-Smith, A., … Tozan, Y. (2016). Costs of Dengue Control Activities and Hospitalizations in the Public Health Sector during an Epidemic Year in Urban Sri Lanka. *PLoS Neglected Tropical Diseases*, *10*(2), 1–13.