SPATIO TEMPORAL FORECASTING OF DENGUE OUTRBREAKS USING MACHINE LEARNING

Manju Lasantha Fernando

168061F

Degree of Master of Science

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

July 2019

SPATIO TEMPORAL FORECASTING OF DENGUE OUTRBREAKS USING MACHINE LEARNING

Manju Lasantha Fernando

168061F

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

July 2019

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters Thesis/Dissertation under my supervision.

Signature of the Supervisor: Date:

ACKNOWLEDGEMENTS

I am sincerely grateful for the advice and guidance of my supervisor Dr. Amal Shehan Perera. Without his insight, encouragement, and his willingness to be available at all times for advice and direction, this project would have not been completed. His consistent positive outlook and resourcefulness guided me through each stumbling block and made sure I carried this study to its conclusion. It would be amiss of me if I did not give similar thanks to Sriganesh Lokanathan, my team lead of the big data team at LIRNEasia, for his advice, guidance, his words of encouragement and the insights provided by him regarding this project work.

I would like to extend my gratitude to my progress review committee, Prof. Nalin Wickremasinghe and Dr. Surangika Ranathunga. Their valuable insights and guidance helped me immensely in shaping the direction of this research.

I would like to thank the entire staff of the Department of Computer Science and Engineering and my colleagues at LIRNEasia for providing their help in various ways during the course of this work, as well as for providing me with the adequate resources to conduct this research.

This work was partially funded by LIRNEasia through a grant from IDRC-Canada as well as from a research grant awarded by the Senate Research Committee, University of Moratuwa, for which I am extremely grateful.

I would also like to express my gratitude to all my friends who lent their support in numerous ways by engaging in discussion, questioning my assumptions and providing knowledge on different domains related to this work. Finally, I would like to thank my family, and especially my parents, for the unconditional support and encouragement provided by them to complete this study to the best of my ability.

ABSTRACT

Spatio Temporal Forecasting of Dengue Outbreaks using Machine Learning

Dengue is one of the most critical public health concerns in Sri Lanka which imposes a severe economic and welfare burden on the nation annually. Prior work has shown that there are multiple factors that contribute to propagation of dengue, including sociological factors such as human mobility. Therefore, it is a non-trivial task to model the propagation of this disease accurately at a regional level. However, accurate quantitative modeling approaches that can predict dengue incidence for a public health administrative division would be invaluable in allocating valuable public health resources and preventing sudden disease outbreaks.

In this study, we make use of large-scale pseudonymized call detail records of approximately 10 million mobile phone subscribers to derive human mobility patterns that can contribute towards disease propagation. We develop 3 distinct proxy indicators for human mobility based on different assumptions and evaluate the suitability of each indicator to accurately model the disease transmission dynamics of dengue. Using the proxy measures developed by us, we go on to show that human mobility has a significant impact on the disease incidence at a regional level, even if the disease is already endemic to a given region.

Combining these proxy mobility indicators with other climatic factors that is known to affect dengue incidence, we build multiple predictive models using different machine learning methods to predict dengue incidence 2 weeks ahead of time for a given MOH division. By introducing an automated input feature selection method based on genetic algorithms, we show that we are able to improve the predictive accuracy of our models significantly, with predictive models based on XGBoost yielding the best performance, with an R^2 of 0.935 and RMSE of 7.688.

Keywords: disease outbreak forecasting; human mobility models; mobile network big data; machine learning applications;

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
ARIMA	Autoregressive Integrated Moving Average
BTS	Base Transceiver Station
CDR	Call Detail Record
DALY	Disability-Adjusted Life Years
DHF	Dengue Hemorrhaegic Fever
DSS	Dengue Shock Syndrome
GA	Genetic Algorithm
LASSO	Least Absolute Shrinkage and Selection Operator
LS-SVM	Least Squares - Support Vector Machines
MC	Municipal Council
MOH	Medical Officer of Health
NDVI	Normalized Difference Vegetation Index
NN	Neural Networks
RF	Random Forests
RMSE	Root Mean Squared Error
RNA	Ribonucleic Acid
SEI	Susceptiple-Exposed-Infected
SEIR	Susceptiple-Exposed-Infected-Recovered
SIR	Susceptiple-Infected-Recovered
SVM	Support Vector Machines
SVR	Support Vector Regression
WHO	World Health Organization

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Dengue fever is a mosquito-borne tropical infectious disease that is caused by the dengue virus, a positive-strand RNA virus of the family *Flaviviridae* and genus *Flavivirus* [1]. There are four main serotypes of the virus identified as DENV 1-4, while a fifth serotype was also reported in October of 2013 [2]. Dengue fever is spread by the day biting Aedes mosquito species. It is mostly observed in tropical as well as sub-tropical regions of the world, and reported to be endemic in over 100 countries. An estimated 390 million dengue infections occur each year across the world, out of which 96 million infections show symptoms of apparent manifestation [3]. In Sri Lanka, the focus region of our study, 55,150 dengue cases were officially reported in 2016, the highest ever number of cases recorded for the country at that time. However, in 2017, this number was eclipsed by a reported 186,101 cases, with severe dengue outbreaks in multiple regions of the country. This highlights the need for possible forewarnings to contain sudden outbreaks and effective utilization of limited public health resources as part of coordinated effort to execute dengue prevention measures at a national scale.

In addition to the mortality and morbidity caused by dengue fever, it is important to take the burden to the economy due to the prevalence of this disease as well. While early detection and access to proper medical care can reduce the rate of mortality for dengue to below 1%, there is a significant economic burden from the morbidity caused by dengue. Stanaway et al. estimates 1.14 million (95% confidence interval - 0.73 million - 1.98 million) disability-adjusted life years (DALYs) were lost in 2013 globally due to mortality and morbidity from dengue [4]. In the city of Colombo itself in Sri Lanka, the burden on the national health care system for preventive measures alone was approximately 1 million US dollars (US$ 971,360) for year 2012 [5].

In such a context, spatio-temporal forecasting of dengue outbreaks at a national

or even regional level would be of significant value to the ongoing efforts to control and prevent dengue outbreaks. Key among that would be the ability for authorities to react before the actual occurrence of an outbreak to execute targeted preventive measures that can potentially preempt a regional outbreak of dengue incidence.

1.1 Problem

Forecasting the propagation of vector-borne infectious diseases at a regional level has been done in prior studies by using multiple approaches. Purely mathematical models as well as models that incorporate statistical and machine learning techniques have yielded mixed results when it comes to modeling propagation of dengue outbreaks across regions [6, 7, 8]. With the increased availability of ubiquitous large-scale heterogeneous data sources and the tools to process such massive volumes of data, there is huge potential in combining these heterogeneous data sources to build more accurate predictive models. Mobile network big data (MNBD) is one such ubiquitous data source that can be used to model human mobility patterns. However, there are multiple research questions that need to be explored in the domains of large-scale data processing, data fusion, and applied machine learning when combining such datasets with traditional data sources such as survey data, official government statistics and aggregated census data.

Multiple studies make use of mobile network big data to model human mobility [9, 10, 11] at population-scale within different application domains. Several studies explore the applicability of mobile network big data for modeling propagation of infectious diseases as well [12, 13, 14, 15]. However, to the best of our knowledge, there does not exist any previous work that derives proxy indicators for human mobility that can be directly associated with a regional administrative district, thus enabling fusion with longitudinal data associated with other risk factors such as previous disease incidence, rainfall, temperature and total population. Additionally, even though there exists work that shows the impact of human mobility in propagating dengue to non-endemic regions, we could not find any

work that quantified the impact of human mobility on regions where dengue is endemic. Lastly, most of the work that forecast dengue outbreaks using machine learning does not provide a comparison of model accuracy between multiple machine learning methods, which would be of significant value to any future researcher needing to decide which machine learning method to utilize when developing forecasting models for infectious disease propagation. Our study focuses on addressing those 3 key questions mentioned above.

1.2 Proposed Solution

We propose methodologies to derive 3 proxy indicators for human mobility, that can be derived using mobile network big data and assigned to a regional administrative district. Some of the human mobility models proposed in our work have a specific focus on disease transmission characteristics of dengue, which we believe would result in higher predictive accuracy for disease incidence forecasting. We further go on to explore the question of whether human mobility has an impact on propagation of vector-borne infectious diseases such as dengue, even in regions where dengue is endemic. In addition to that, we build multiple machine learning models to predict dengue incidence in several regions of Sri Lanka for 2014 and provide a comparison of model accuracy for different machine learning methods and different input features. We go on to select the best performing model based on this comparative analysis and discuss further steps that need to be implemented in order for this model to be generalized to forecast other infectious diseases as well.

1.3 Contributions

We make the following contributions in this thesis:

- Provide the methodology to obtain several proxy indicators for human mobility that can be derived from processing population-wide longitudinal mobile phone call detail records (CDRs), and provide the resultant mobility proxy values for each medical administrative division of Sri Lanka

- Propose a methodology to explore and quantify the impact of human mobility on dengue propagation
- Provide results that establish the relationship between human mobility and dengue incidence in a Sri Lankan context where dengue is endemic to most regions of Sri Lanka
- Compare predictive accuracy between different machine learning methods on forecasting dengue outbreaks for selected administrative districts of Sri Lanka

1.4 Organization

The rest of this document is organized as follows. Chapter 2 presents summarizes past and contemporary work in multiple domains that relate to our work and details the insights we have drawn from each of these studies. Chapter 3 describes the data set used in detail and the pre-processing techniques applied to each of these data sources. In chapter 4, we describe the data driven models formulated by us for understanding the impact of human mobility on propagation of dengue, and present the formal definitions for each of these models. Chapter 5 describes the methodology for developing the forecasting models. Chapters 6, 7 present the results and discuss them in detail respectively. Finally, chapter 8 presents the conclusion of our research while chapter 9 discusses future directions that can be explored based on this research.

Chapter 2

LITERATURE SURVEY

In this chapter, we look at prior work done in the area of forecasting the propagation of infectious diseases. First of all, we look at literature on risk factors associated with transmission and propagation of dengue fever. Then we go on to look at work that has used human mobility models for modeling disease propagation characteristics. Finally, we examine prior research studies that predict infectious disease outbreaks using statistical modeling approaches as well as machine learning based approaches.

2.1 Factors affecting propagation of dengue

Dengue is a virus that is transmitted to humans mostly by the day biting Aedes aegypti and Aedes albopictus mosquito species [16, 17]. The distribution of these vector populations have been directly linked to the disease from multiple studies ranging back to the post world war II period [18, 19]. In light of these studies, a global study done in 2015 [20] which attempts to map the global distribution of these two mosquito species using reported samples and temperature constraints is of significant importance when attempting to model the spread of these globally.

Many studies discuss in detail the relationship between weather related parameters and dengue incidence [21, 22, 23]. In addition to the weather parameters, sociological features such as human movement patterns [14], availability of potential mosquito breeding sites, population density can also contribute towards propagation of dengue outbreaks.

2.1.1 Weather related features

Multiple studies have shown that various weather related parameters affect the dengue transmission risk factors such as the size of the vector population, incubation period of the virus, and the biting rate of the mosquitoes. In a study done by Yang

et al. in 2009, a series of temperature-controlled experiments were conducted to assess the effects of temperature on the population of Aedes aegypti [24]. In their model, they assessed different entomological parameters that would affect the population size for different temperatures and model Q_0, the basic offspring number as a function of temperature.A 2014 study by Liu-Helmerssion et al. document how the vectorial capacity of Ae. aegypti species change according to the variation in temperature, and go on to map the epidemic potential for dengue based on these findings [25]. Similarly, another study done in 2014 on temperature constraints on the persistence of the main vectors Ae. aegypti and Ae. albopictus show that temperature affects the length of the first gonotrophic cycle of both mosquito species, and the oviposition suitability. It also shows that the temperature affects the introduction suitability and the persistence suitability of the disease in the two mosquito species at different rates. Interestingly, this study highlights that Ae. albopictus, considered to be a secondary vector species when compared to Ae. aegypti, can be much more resilient to changes in temperature and have a higher vectorial capacity [26].

In addition to the temperature factor, there are multiple weather related parameters that contribute towards a regional outbreak of dengue. A systematic review of climatic factors that affect dengue incidence find that most studies cite temperature, rainfall and relative humidity as critical factors for dengue propagation [27]. It also highlights the importance of having long term climate data as well as socio ecological data, and integrating different quantitative modeling approaches with interdisciplinary research collaborations to advance the spatio temporal modeling of dengue propagation. A study done in Singapore[21] demonstrates that weekly mean temperature and cumulative precipitation at a time lag of 5-16 and 5-20 weeks respectively increases linearly with dengue incidence. Hu et al. showed for a study done in Australia that a change in the amount of rainfall, in conjunction with other climatic factors resulted in a change in dengue incidence [28]. Another study made use of genetic algorithms to automatically derive the best features that can predict dengue incidence after decomposing each input feature into multiple terms using wavelet transforms [29]. The results indicated

that cloudiness, maximum temperature, minimum temperature, humidity(max, mean, and min), and rainfall intensity were contributing factors for dengue incidence. Interestingly, mean temperature, a weather parameter that has figured significantly in other studies [30, 23], was not selected as a significant input feature. At the same time, several other contemporary studies conducted in Guadeloupe, French West Indies[31] and in Singapore[32] show temperature effects to have a higher influence on the dengue disease incidence rate when compared to rainfall.

In Sri Lanka, which is the focus of our study, prior work has highlighted the relationship between dengue incidence and climate factors. A study in 2013 that analyses the effect of climatic factors using timeseries data for 3 districts does not find any strong correlations between dengue incidence and average maximum temperature or total rainfall [22]. Another study done in 2016 for Kalutara district finds that there is a strong association for rainfall at different time lags and El Nino Southern Oscillation with dengue incidence [33].

2.1.2 Human Mobility

A study done by L.E. Muir and B.H. Kay in 1998 on Aedes aegypti report a maximum dispersal of 160m [34] for this mosquito species while a much older study in 1958 [35] cites 1150m. However, in general most of the work report that a majority of Ae. aegypti mosquitoes disperse less than 80m [34]. A related study done in Brazil by Honorio et al. [36] report a maximum dispersal range of 800m for both Ae. aegypti as well as Ae. albopictus, another Aedes mosquito species that can potentially transmit the dengue virus. Due to the limited spatial range of the mosquito vectors of dengue, the propagation of the disease across large regions is believed to be due to movement of infected humans. This has been confirmed by several experimental studies as well [37, 12, 14].

2.2 Modeling Human Mobility

Modeling human mobility for the purpose of understanding disease transmission dynamics has been done using various survey based, mathematical, computational

and hybrid approaches [12, 6, 14]. Traditional survey data has been instrumental in modeling human movement patterns in multiple domains such as transportation, disease transmission, migration and related domains in sociology. However, with the availability of other forms of large-scale heterogenous data sources such mobile phone CDRs, and the tools and computational power to process such forms of data, it has been found that combining such data with survey based data can yield to more accurate, granular insights that had hitherto not being possible [12, 13].

Widely used mobility models such as gravity model[38], radiation model, disease transmission dynamics based SEIR, SIR [6] models, as well as more customized models[39] have been used in multiple studies to describe human movement patterns for the purpose of understanding the propagation of infectious diseases. A study conducted by Brockmann [37] on how a pre-identified set of dollar bills were circulated across the world is one of the first large-scale experiments that attempted to track human movement patterns at global scale.

Using mobile phone CDRs to model human movement patterns had become feasible with the advances in large-scale data processing techniques as well as availability of increased computational power for academic purposes. A research study done by Gonzalez et al. [40] is one of the earliest examples of using CDRs to track the trajectory of individual users. Research done by Isaacman et al. make use of CDRs to compare mobility patterns of users in different cities in [41]. The work of Isaacman and colleagues on identifying important places of users [9] describes one of the key algorithms that has been used in much of the subsequent work to identify home and work locations of a mobile phone subscriber. However, the pioneering body of work done by Wesolowski et al. [42, 13, 14] has been the most influential in understanding human mobility using pseudonymized mobile phone CDRs in the context of infectious disease propagation. Similar studies done by Bengtsson et al. [15] and Finger et al. [43] also make use of mobile phone data to disease dynamics of infectious diseases.

2.3 Disease Outbreak Forecasting

Modeling disease transmission dynamics to forecast the propagation of an infectious disease, is an open problem that had been of significant interest to researchers and practitioners in the health care industry for decades. Mathematical, statistical models that were developed initially have been advanced and extended with the availability of computational techniques to estimate parameters of these models. In the case of dengue, there is a large body of work that describes numerous mathematical, statistical, computational as well as hybrid techniques that had been used to model and forecast the propagation of this disease.

2.3.1 Mathematical Models

Many studies [44, 45, 46] that model dengue propagation dynamics make use of the SEIR-SEI model which is a variant of the SIR model, for which the early work was done by Kermack and McKendrick [47]. The meta population model introduced by Sarzynska et al. in 2013 [6] use ordinary differential equations to model the population changes of each compartment in SEIR-SEI model. The SEIR-SEI model divides the host and vector population to different compartments. Each host in the population is considered to fall into one of the four categories: susceptible (S), exposed (E), infected (I) or recovered (R). A vector is considered to be in one of susceptible (S), exposed (E), or infected (I) states. The recovered state is not considered for the vector since it is assumed that during the very short life cycle of vectors such as mosquitoes, there is not enough time for a vector to recover from the infection before it reaches the end of its life cycle. This study further goes on to consider gravitation and radiation models of mobility to model human movement patterns which is incorporated into the overall disease propagation model. Other SIR based models have been developed further in works such as [48] and [7]. The study in [7] is significant due to the fact that it proposes a compartmental model that also considers secondary infections for dengue. However, even though the theoretical model is proposed and justified in this work, we could not find any work that had evaluated experimental results to

validate whether the model is applicable in practical settings.

Apart from the compartmental SEIR models, many studies have used time series based analysis for forecasting dengue outbreaks as well. Hii et al. has conducted multiple studies that analyse dengue incidence as a time series to forecast dengue outbreaks in Singapore[21, 30, 49]. This body work uses time series Poisson regression to predict dengue incidence and also breaks down the time series data into its seasonal, periodic and residual components to model dengue. Auto-regressive Integrated Moving Average (ARIMA) models have also been used in multiple studies [50, 51] as well as wavelet based approaches [52, 53].

2.3.2 Machine Learning Models

In addition to using purely statistical methods to predict dengue outbreaks, recent studies have increasingly made use of hybrid computational approaches that aid those statistical methods as well as machine learning models. If you consider machine learning based approaches, techniques such as Neural Networks (NN), Support Vector Machines (SVM) and Random Forests (RF) have been used extensively. A study done in 2008 pre-processed dengue incidence data to derive an entropy value, which was then used to train a neural network instead of directly feeding the number of dengue cases to the model [54]. The model was trained to predict whether a given week should be considered as an outbreak of dengue hemorrhagic fever (DHF) or not. It was able to achieve an accuracy of 86% using this entropy based technique. Another study done in Singapore also used Neural Networks to perform a regression where the number of dengue cases were predicted, which achieved a correlation of 0.91 and an RMS error of 50.7 [8].

SVM based models were also used to predict dengue outbreaks in many studies with good results. [55] compares a neural networks based approach against a Least Squares - Support Vector Machine (LS-SVM) based approach, using dengue incidence data for Malayasia, to show that LS-SVM gives a higher accuracy of 87% when compared to the 66% level of accuracy given by the neural networks [55]. A more advanced support vector regression based approach broke down

the time series dengue incidence data using wavelet decomposition and used a genetic algorithm (GA) based approach to predict the exact number of dengue cases that will be reported [29]. Support vector machines have been widely used in prediction of other infectious diseases as well [56, 57].

A more recent study done in Pakistan had made use of random forests to predict dengue incidence [58]. One interesting feature of this study is that the authors have used the awareness level as also an input feature to the model. Awareness of the diseases, and awareness of best practices to prevent occurrence of mosquito breeding sites by the community would arguably one of the important features that can affect the spread of a disease such as dengue. However, we were not able to find any other studies that included awareness as an input for training the predictive model apart from this study.

A multitude of other machine learning techniques, as well as hybrid approaches have been used to predict dengue incidence in many research studies. In addition to the techniques described above, LASSO regression [59], models based on cellular automata [60, 45], agent based modeling techniques [61], process based models [23], fuzzy association rule mining [62] have been used to predict spatial spread of dengue and other infectious diseases. However, it should be noted that we could not find studies that compared between many techniques to establish the suitability of a given technique for this particular problem.

2.4 Summary

Multiple studies have already focused on predicting dengue outbreaks using machine learning and mathematical modeling, with varied levels of success. A careful review of contemporary studies indicate that the transmission dynamics of a complex disease such as dengue is highly dependent on the context, where a multitude of factors related to seasonal weather patterns, the level of urbanisation, human movement patterns and other sociological behaviours play a critical role. As such, in order to determine which modeling techniques work best for a tropical region where dengue is already endemic such as Sri Lanka, a comparison between

different approaches for the same data set is needed.

Previous studies had made use of ubuiquitous secondary data sources such as CDRs to model human movement which in turn was used to model disease propagation. However, it is difficult to abstract these techniques to derive mobility information that can be used as an input data set for any computational modeling approach. Additionally, whether human mobility is a significant factor for propagating dengue in an already endemic environment is a question that had not been answered in domain literature to the best of our knowledge.

The methodology for our study, explained in detail in the next three chapters, was focused mainly on addressing the gaps in literature identified above, and developing an efficient computational approach to model spatio-temporal propagation of dengue outbreaks accurately.

Chapter 3

DATA COLLECTION AND PRE-PROCESSING

Our work focused on predicting dengue outbreaks for selected Medical Officer of Health (MOH) divisions, which are administrative districts demarcated by the Ministry of Health. In order to develop the predictive models, we first needed to identify potential data sources that could be used as input for our models. Based on our literature survey, the following sources were identified for data collection and pre-processing.

- Weekly past dengue incidence data, reported for each MOH division
- Estimated population of each MOH division
- Pseudonmyized mobile phone CDRs
- Daily temperature measurements
- Daily rainfall measurements
- Vegetation cover

Different pre-processing techniques were used to clean and normalize the above data sources. A short description of the pre-processing steps undertaken as well as the nature of the data source itself is provided in the remainder of this chapter.

3.1 Dengue Incidence Data and Information on MOH Divisions

The number of confirmed dengue cases reported for each MOH division for each week was obtained from the Epidemiology Unit, Ministry of Health [63] for 3 years from 2012 to 2014. We were also able to obtain the 2014 population estimates for each MOH division, as well as the digital shapefile that denoted the demarcation boundaries for each MOH division.

Table 3.1: Structure of a Call Detail Record

Caller Party ID	Called Party ID	Cell ID	Call Time	Call Duration
A24BC1571X	B321SG141X	3134	13-04-2013 17:42:14	00:03:35

3.2 Pseudonymized Mobile Phone Call Detail Records

Pseudonymized mobile phone call detail records were obtained for a period spanning more than 1 year from mobile network operators in the period of 2012-2013 for the entire country. The mobile network operators assigned a unique identifier to each mobile subscriber replacing his/her actual phone number in the entire dataset, before the CDR data was shared with the researchers. The basic structure of a call detail record that was available for this study is given in table 3.1.

Daily CDRs of voice calls for almost 10 million mobile phone subscribers were analyzed to obtain the proxy indicators for mobility models described in chapter 4. The CDR dataset was filtered to get exactly 52 weeks of data from the entire time-period for which data was available.

3.3 Temperature and Rainfall Measurements

Initially, we obtained weather data for a single year from the Department of Meteorology, Sri Lanka for 112 weather stations for the year 2014 [64]. However, it was decided not to use this dataset due to data quality issues. We also explored the weather data sets available under the dark sky service which provides historical weather data for any location [65]. However, we decided not to use that as a source due to the unavailability of the methodology through which the data was obtained.

Ultimately, weather data was obtained from the National Oceanic and Atmosperic Administration (NOAA) Integreated Surface Database (ISD) [66] for 22 weather stations across Sri Lanka. In general, weather data was reported on a daily basis, while some measurements such as precipitation was reported in 12-hourly, 6-hourly or hourly intervals. This data was grouped according to the week of the year, and converted to weekly aggregate measurements to obtain mean

temperature, maximum temperature, minimum temperature and total precipitation for each week.

3.4 Vegetation cover

Data regarding the vegetation indices were derived using remote-sensing data available from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite. We used the MOD13Q1[67] data product, which has data available at spatial resolutions of 250m,500m,1km and 0.05 degrees. The temporal resolution would vary according to the orbit of the satellite and was not available at a granularity of 1 week. Due to the variations in temporal granularity, we derived a composite value for the greenness of a particular spatial area for a 16-day period. For the resolution used by us to derive the vegetation index, each pixel represents a 250m x 500m (125,000 m^2) area. The Normalized Difference Vegetation Index (NDVI) value, which is an indicator of the amount of chlorophyll present for a given resolution, was directly reported in the MOD13Q1 dataset. The 16-day composite NDVI value obtained for each pixel was averaged across all the pixels belonging to a particular MOH division to obtain the mean NDVI value for a particular MOH division. We used the 16-day based mean NDVI value for a given MOH division to assign weekly mean NDVI values, based on the assumption that the NDVI value does not change significantly during a 2-week period.

3.5 Imputation of missing values

The identified input features had data available for different time-periods and regions. Also, since we were considering time-series data, we used measured input values for previous weeks or lagged values as well. For certain datasets, the lagged values were not available. Due to these reasons, there were missing values for some of the data points. After carefully reviewing the availability of data and the quality of the data available, we decided to restrict our study to 20 MOH divisions where the weather data was also available. Even after applying these restrictions, there were multiple missing data points which were imputed.

For imputation of missing input values, we used the Multivariate Imputation via Chained Equations (MICE) package available for R, where Predictive Mean Matching (PMM) was used the imputation methodology [68].

3.6 Summary

The heterogenous nature of the different input data sources required significant data processing and cleaning to be transformed to be compatible as input for various computational models. The final cleaned data product had weekly dengue incidence data, population of the MOH division, 4 weather parameters (maximum temperature, minimum temperature, mean temperature, and total rainfall), and mean NDVI values in addition to the 6 mobility measurements yielded by the pseudonymized CDR data. All the easurements that were used as input features, except for population, were lagged up to 12 weeks, yielding 167 variables spanning for 3 years as the complete cleaned dataset. The process of deriving mobility measurements using CDRs, the motivation and assumptions behind each mobility model and the processing techniques are described in detail in the next chapter.

Chapter 4

HUMAN MOBILITY MODELS

We derive multiple proxy indicators for human movement patterns and mobility related risk, where some of our models consider specific dengue transmission characteristics as well. All models, except for the 2 trip based models, consider the location where a subscriber's home or work place is situated. We came up with 6 mobility models based on 3 conceptual approaches to model human movement patterns. Some of the models are extensions of similar work encountered in literature and make use of the same assumptions, while others have been based on our own intuition. In all of our models, it must be noted that we assume the behavior of the mobile phone subscribers in our data set to be representative of the entire population.

The 3 conceptual modeling approaches are based on the following high level observations: (a) If a subscriber spends more time at a given region, he/she is more likely to propagate or contract the disease from that region (b) The number of people coming in and out of a given region would influence the risk of a region having an outbreak (c) Risk of exposure would vary according to the time of day and place where a subscriber spends his/her time.

4.1 Identifying Home and Work Locations

The BTS corresponding to the home and work location of a subscriber was identified by considering the tower that appeared on most number of days during a night time or day time, given a particular subscribers CDRs. We used the methodology developed by Lokanathan et al. [69] to derive the home and work location. Time between 1000 to 1500 hours was considered to identify tower location that corresponded to work, while the time between 2100 and 0500 hours was considered to identify the tower location corresponding to a subscriber's home location. The home or work MOH division was selected by considering

MOH division on which the home/work BTS was located.

4.2 Probabilistic Mobility Model

We developed our initial mobility model based on the assumption that the amount of time a particular subscriber spends in a particular region is proportional to the number of calls he/she had made or received within that given region. This assumption, while quite generic, was used by researchers previously to highlight the role of human mobility in spreading of the disease in the 2005 cholera outbreak in Senegal [43]. We believe that based on the results of that study, we are justified in using that same assumption to develop a mobility model that assigns a mobility value to each subscriber based on the probability he/she might be found within a particular region for a given week. We go on to define two mobility values that are assigned to each coverage area of BTS based on this model - visitation based probabilistic mobility and exploration based probabilistic mobility. The formal definition of these two mobility indicators has been described below.

4.2.1 Formal Definition - Visitation Based Probabilistic Mobility

Let us define B - Set of all BTS coverage areas, S - Set of all mobile phone subscribers under consideration. Then, we define,

$$cdr(b_i, s_j, w_k) = Number\ of\ CDRs\ in\ BTS\ b_i,\ for\ subscriber\ s_j,$$
$$during\ week\ w_k,\ \forall b_i \in B, \forall s_j \in S$$

Visitation based mobility of subscriber s_j at BTS b_i, at week w_k

$$\forall b_i \in B, \forall s_j \in S: \quad visitation\_mob(b_i, s_j, w_k) = \frac{cdr(b_i, s_j, w_k)}{\sum_i^B cdr(b_i, s_j, w_k)} \tag{4.1}$$

Visitation based mobility for BTS b_i, at week w_k

$$\forall b_i \in B: \quad visitation\_mob(b_i, w_k) = \frac{\sum_j^N visitation\_mob(b_i, s_j, w_k)}{|N|} \tag{4.2}$$

$$where\ N = \{s_j \in S \mid home(s_j) \neq b_i,\ cdr(b_i, s_j, w_k) > 0\}$$

4.2.2 Formal Definition - Exploration Based Probabilistic Mobility

Similar to above, let us define B - Set of all BTS coverage areas, S - Set of all mobile phone subscribers under consideration. We define the other terms as follows.

$$outside\_cdr(s_j, w_k) = \sum_i^T cdr(b_i, s_j, w_k)$$
$$\forall s_j \in S,\ \forall b_i \in T,\ where\ T = \{b_i \in B \mid b_i \neq home(s_j)\}$$

Exploration based mobility of subscriber s_j at BTS b_i, at week w_k

$$\forall s_j \in S: \qquad exploration\_mob(s_j, w_k) = \frac{outside\_cdr(s_j, w_k)}{\sum_i^B cdr(b_i, s_j, w_k)} \tag{4.3}$$

Exploration based mobility for BTS b_i, at week w_k

$$\forall b_i \in B: \quad exploration\_mob(b_i, w_k) = \frac{\sum_j^Q exploration\_mob(s_j, w_k)}{|Q|} \tag{4.4}$$

$$where\ Q = \{s_j \in S \mid home(s_j) = b_i\}$$

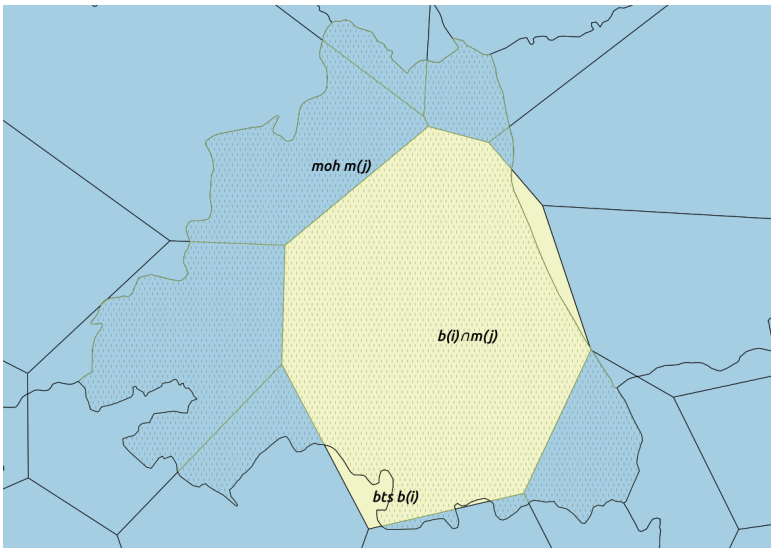


Figure 4-1: Mapping of BTS b(i) (b_i) to MOH m(j) (m_j)

4.2.3 Mapping Probabilistic Mobility Indicators to MOH divisions

In order to map a mobility value that was derived for a BTS tower coverage area to an MOH division, we use the following definition.

Consider BTS b_i and MOH m_j overlaps as in figure 4-1. Then,

$$A_{b_i} = Voronoi\ cell\ based\ coverage\ area\ of\ BTS\ b_i,$$
$$A_{m_j} = Area\ of\ MOH\ m_j,$$
$$A_{b_i \cap A_{m_j}} = Area\ of\ overlap\ between\ BTS\ b_i\ and\ MOH\ m_j,$$
$$x_i = Per\ person\ mobility\ value\ for\ BTS\ b_i,$$
$$x_j = Per\ person\ mobility\ value\ for\ MOH\ m_j$$

We define an MOH based overlap ratio, $R_{M_{ij}}$ as follows:

$$R_{M_{ij}} = \frac{A_{b_i \cap A_{m_j}}}{A_{m_j}} \tag{4.5}$$

Then, per person mobility value for MOH m_j, denoted by x_j is:

$$x_j = \sum_i^N x_i \times R_{M_{ij}} = \sum_i^N x_i \times \frac{A_{b_i \cap A_{m_j}}}{A_{m_j}} \tag{4.6}$$

4.3 Trip Based Mobility Model

For comparison, we utilized a mobility model that was based on the number of trips by a mobile phone subscriber from one BTS coverage area to another. A trip is defined as a consecutive pair of call detail records for a given subscriber where the corresponding BTS tower ID is different. We also impose an upper bound of 1 day and a lower bound of 10 minutes for the trip. An upper bound was introduced so that a trip can be time bound to avoid noise that might get introduced if trips overlapping multiple weeks are considered for the model. The lower bound was decided based on similar thresholds that were introduced in previous studies [70].

4.3.1 Formal Definition - Trip Mobility

Let us define M - Set of all MOH, S - Set of all subscribers of the network. Then we define,

$$trips(s_j, b_p, b_q, w_k) = Number\ of\ trips\ by\ subscriber\ s_j\ from\ b_p\ to\ b_q,$$
$$during\ week\ w_k,\ \forall\ b_p, b_q\ \in B \mid b \neq q,\ \forall\ s_j \in S$$

We define R_v as the set of subscribers residing in coverage area of BTS b_v. We denote *Per Person Trip Mobility* for BTS b_v from BTS b_p to BTS b_q during week w_k as $ppt(v, b_p, b_q, w_k)$. Then,

$$pptm(v, b_p, b_q, w_k) = \frac{\sum_j^{R_v} trips(s_j, b_p, b_q, w_k)}{|R_v|}$$

Per Person Trip Mobility from BTS b_p to BTS b_q during week w_k will be,

$$\forall b_p, b_q \in B, \forall s_j \in S: \qquad pptm(b_p, b_q, w_k) = \sum_v^B pptm(v, b_p, b_q, w_k) \qquad (4.7)$$

4.3.2 Mapping Trip Mobility Indicators to MOH Divisions

Mapping mobility values derived for a BTS coverage area was estimated for an MOH division using the following formula.

Consider that a trip originates from coverage area of BTS b_p, which overlaps MOH m_i, and destination lies in coverage area of BTS b_q, which overlaps MOH m_j. Then using the definition 4.5 above,

$$pptm(m_i, m_j, w_k) = \sum_p^B \sum_q^B pptm(b_p, b_q, w_k) \times R_{M_{pi}} \times R_{M_{qj}} \tag{4.8}$$

4.3.3 Normalized MOH based trip mobility

By using the per person trip mobility value define in equation 4.8 above, we derive Normalized Outgoing Trip Mobility and Normalized Incoming Trip Mobility as follows:

For a given MOH m_i, normalized outgoing trip mobility is defined as,

$$\forall m_i \in M : \qquad trip\_mob\_out(m_i, w_k) = \sum_j^M pptm(m_i, m_j, w_k) \tag{4.9}$$

Similarly, for a given MOH m_j, normalized incoming trip mobility is defined as,

$$\forall m_j \in M : \qquad trip\_mob\_in(m_j, w_k) = \sum_i^M pptm(m_i, m_j, w_k) \tag{4.10}$$

4.4 Mobility based Risk Model

We also developed a third model to derive a risk score based on CDR data that takes into account the location and the time of day of a particular subscriber. The motivation behind giving different risk scores based on location and time of day is due to the fact that Aedes aegypti, the primary vector for dengue, is a day biting mosquito, and the risk of getting bitten by an infected mosquito would depend on the time of day and the location. Based on the observation of a mobile phone subscriber at a particular location, inferred by the BTS location

of a single call detail record, we assign the following risk scores for each location and time band. A more theoretical approach to modeling risk has been discussed in a study done by Stoddard et. al in 2009 [71]. However, their model is not validated against actual data and is discussed by considering an example. Some of the entomological parameters needed for these models such as biting rate of mosquitoes for a given site, the proportion of vectors at a given site etc. will be difficult to estimate at a national scale in practice.

Table 4.1: Risk score based on time band and location type

Location Type	Time Band	Risk Score
Home	06:00 - 09:00	0.5
	17:00 - 19:00	0.5
	Other	0.3
Work	06:00 - 09:00	0.7
	17:00 - 19:00	0.7
	Other	0.4
Other	06:00 - 09:00	0.8
	17:00 - 19:00	0.8
	Other	0.6

The time bands are selected based on the fact that Aedes aegypti mosquito vector is mostly active just after sunrise, and just before sunset [72]. Using these initial risk scores, we derived two risk indicators representing a single MOH. Formal definition for these two risk indicators is described below.

4.4.1 Formal Definition - Mobility based Risk

Let us define M - Set of all MOH, S - Set of all subscribers of the network, C - Set of all CDRs, L - Set of all location types. T - Set of all time bands. Then we define,

$$c_p(l, t, s_j, b_i, w_k) = A\ call\ detail\ record\ at\ location\ type\ l,$$
$$within\ time\ band\ t,\ of\ subscriber\ s_j, at BTS\ b_i,$$
$$during\ week\ w_k, \forall l \in L, \forall t \in T, \forall\ b_i \in B, \forall\ s_j \in S$$

Then we define risk as $risk(s_j, w_k)$ for a given subscriber s_j, and the presence

of a subscriber in BTS b_i as $presence(s_j, b_i, w_k)$, during week w_k as follows:

$$risk(s_j, w_k) = \frac{\sum_p^{C_j} risk(c_p(l, t, s_j, b_i, w_k))}{|C_j|}$$
$$where\ c_p \in C_j \subset C | subscriber(c_p) = s_j$$
$$presence(s_j, b_i, w_k) = \frac{\sum_p^{C_{j_i}} c_p}{|C_j|}$$
$$where\ c_p \in C_{j_i} \subset C | subscriber(c_p) = s_j\ \&\ bts(c_p) = b_i$$

We additionally denote by S_{ik} as the set of subscribers seen in BTS b_i, during week w_k. Then we go on to define two risk indicators based on the above formulations. Direct risk score, denoted by $direct\_risk(b_i, w_k)$, is derived by only considering the average risk scores of each subscriber that visited a given BTS coverage area. Risk score based on the percentage of the presence of a subscriber, $percent\_risk(b_i, w_k)$, considers the fraction of presence of a user as well as his/her risk score for the week w_k.

$$\forall\ b_i \in B \qquad direct\_risk(b_i, w_k) = \frac{\sum_j^{S_{ik}} risk(s_j, w_k)}{|S_{ik}|} \tag{4.11}$$

$$\forall\ b_i \in B \quad percent\_risk(b_i, w_k) = \frac{\sum_j^{S_{ik}} risk(s_j, w_k) \times presence(s_j, b_i, w_k)}{|S_{ik}|} \tag{4.12}$$

4.4.2 Mapping Risk Scores to MOH divisions

In order to map risk scores to MOH division, equation 4.6 described in sub section 4.2.3 above is used.

4.5 Summary

The 6 mobility models were developed based on 3 different conceptual approaches because we wanted to experiment and understand which assumptions provide the best results for our models. In order to process such large volumes of data, we used a cluster of 10 machines based on commodity PCs to run Apache Hadoop and Spark, where an offline processing job using Apache Spark was run to perform the necessary aggregations for each model. After the output from the processing of CDRs was available, validations and experimental visualizations were performed to verify the accuracy of the resultant mobility proxy indicators. The mobility data, combined with other weather parameters were used to build the final prediction models. Our approach to develop those predictive models is described in detail in the next chapter.

Chapter 5

DEVELOPING FORECASTING MODELS

Since prior literature did not conclusively point towards a single computational technique to model spatio-temporal dengue propagation, and one of our research objectives was to identify which techniques work best in this context, our methodology conisted of multiple experimental steps before the final models were trained. Model performance was measured using both RMSE and R^2 with more weight given to improving R^2 since it accounts for the variance of the data. Initial exploration of the data set was limited to 6 MOH divisions where normalization and transformation techniques were applied to determine whether such transformations increased predictive accuracy. The machine learning methods used for the final round of training was determined based on the initial exploration performance. In addition to that, correlation between variables was measured using multiple techniques due to the fact that we wanted to capture non-linear relationships between variables as well. The initial traning of the machine learning models were done by feeding all available input variables, which affected the performance of some techniques such as neural networks. The genetic algorithm based feature selection was introduced later on which significantly improved the performance across all models.

5.1 Exploring the data set

We initially selected 6 MOH divisions and built preliminary models using data for that 6 MOH divisions. The MOH divisions were Nuwara Eliya, Galle, Kandy, Anuradhapura, Kurunegala and Moratuwa. In that phase, data from 2012 and 2013 were used as training data, but excluding the data for year 2013 from Moratuwa MOH division. Data from 2013 for Moratuwa MOH was used as the test set for the exploratory analysis conducted. For exploration, we utilized generalized linear modeling techniques to build predictive models after applying

various pre-processing steps such as log scaling, normalized response variables, using principal components of the predictors instead of using the predictors themselves. We also trained predictive models using the four machine learning techniques that were eventually used (neural networks, support vector regression, random forests and XGBoost), without the genetic algorithm based optimization. Hyper parameter tuning was done using different R packages available. For neural networks, hyper parameter tuning was done using the caret[73] package, while in-built tooling of the e1071 R package was used for tuning support vector regression models. For random forests and XGBoost, parameter tuning was done manually.

5.2 Correlation Analysis

Prior to training the final machine learning based models, we measured the correlation between different input features and the dengue incidence by using the pre-processed data. Pearson's correlation[74] measurement was used initially, which can verify the existence of a linear relationship between two given variables. However, this correlation metric is unable to capture non-linear relationships. Therefore, we made use of other correlation measures as well. For this purpose, we used distance correlation [75] and mutual information measure between variables [76].

5.3 Input data for the model

After initial data exploration, we separated the available data set of 3,120 data points for 20 MOH divisions spanning years 2012-2014 into two separate sub sets: a training data set and a test data set. We did not make use of a validation data set due to the limited number of data points available for training. For 5 MOH divisions, namely Colombo-MC, Galle, Kandy, Jaffna and Haputale, the data for the entire year of 2014 was considered as test data set. All remaining 2,860 data points were used for training. We did not consider the value of each input variable for a given week and for the week before that. Only the input feature values 2 to 12 weeks before was considered so that the model is forecasting 2 weeks ahead

using the current data.

In the Sri Lankan context, dengue outbreaks within the Colombo municipal council is of significant interest due to the fact that highest dengue incidence is reported each year from this MOH division. It is also the commercial capital of Sri Lanka and acts as a hub for human mobility. Therefore we wanted to see whether our models could accurately predict dengue incidence for a critical administrative region such as Colombo-MC. For this reason, model performance metrics for Colombo-MC was calculated separately for each machine learning method.

5.4 Measuring performance of the model

After getting the predicted dengue incidence values using the trained models and comparing the against the actual values that were reported, we used two separate metrics to estimate the performance of each model. Only predictions for the test data set were used to calculate the model performance. The two metrics used were Root Mean Squared Error (RMSE), which gives you an idea about how much your model deviates from the actual value on average, and R^2, the co-efficient of determination. The following standard formulas were used to calculate RMSE and R^2 measures.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{5.1}$$

$$y_i = i^{th} observation\ of\ the\ response\ variable$$
$$\hat{y}_i = i^{th} prediction\ of\ the\ response\ variable$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.2}$$

$\bar{y} =$ *Mean of response variable*

5.5 Genetic Algorithm based optimization

In order to select features that are most relevant for a model, we developed a genetic algorithm (GA) based approach. A similar approach was described by Wu et al. [29] for a study conducted in Singapore. However, the approach used for our study had some key differences from the approach described in that study. First of all, a cross validation approach was used in [29] to determine which features were used as input for the final model. In our approach, the selection of the features is based on the best models of a certain generation, which would be carried to the next generation automatically. Additionally, in [29], the feature selection was not constrained to have mandatory inclusion of features derived from a given data source. Therefore, there was a possibility for all derivative features of a particular data sources to be entirely excluded from the model, while the derivative features of another data source might be included in its entirety. Additionally, since we derive multiple time lagged features from a single data sources, there is a possibility for these derived features from the same data source to have significant auto correlation. Therefore, we constrained the feature selection process such that only a configurable minimum and maximum amount of features from a given data source is allowed.

We used the GA package available for R [77] and used binary chromosome based GA for feature selection. Custom R functions were written to evaluate the fitness of each model and constraining of the maximum number of features were done by overriding the default crossover function. We also made use of the parallel GA feature and used a cluster of machines running in parallel to speed up the training process. The configuration parameters for the genetic algorithm is summarized in table 5.1.

Table 5.1: Parameters for the genetic algorithm

Parameter	Value
Population size	100
Maximum iterations	50
Crossover Probability	0.8
Mutation Probability	0.1
Elitism	5
Parallel back-end	PSOCK
Number of worker machines	8
Number of cores per worker	6

5.6 Machine Learning Methods

We evaluate several machine learning techniques to predict dengue outbreaks for a given MOH division. The techniques we used and the reason for selecting each different technique is described below briefly.

5.6.1 Neural Networks

We selected neural networks as one of the techniques to be evaluated because of its popularity in literature, which makes it ideal to be used for comparing our work against similar contemporary studies. We use the R package 'neuralnet'[78] and trained with time lagged input features after pre-processing was done as described in chapter 3. We experimented with different algorithms, activation functions, error functions and different architectures with multiple hidden layers of neurons. Based on experimental results and manual tuning of parameters, the neural network for the final results used sum of squared errors (SSE) as its error function, a threshold of 0.01 in the difference in error to stop training iterations. Resilient back propagation with weighted backtracking was used as the algorithm for adjusting weights. Activation function used was a logistic function.

5.6.2 Support Vector Machines

Support vector machines were also selected since it was had performed quite well in disease outbreak prediction according to multiple studies [29, 55, 57]. In broad terms, support vector machines consider a data point to lie in n-dimensional

space and attempts to find the maximum margin hyperplane that classifies the data according to the class labels. The theory behind support vector machines was developed by Cortes and Vapnik in early 1990s [79]. For our work, we use support vector regression (SVR), which was developed on the same theoretical basis as the classification algorithm [80].

We used the 'e1071' package in R [81] to train our models and initial hyper parameter tuning was done using the 'caret' package [73]. For the final SVR model, we use a radial basis function kernel, with ν-regression and a cost of 3, where the ν value is set to 0.35, γ value of 0.004. 4-fold cross validation was used when training the model.

5.6.3 Random Forests

We used random forests [82] also as a method to develop forecasting models because of its suitability for a wide-ranging set of problems in machine learning, as well as the fact that its decision tree based methodology provides information on which input variables contributed most for the output generated by the model. The 'randomForest' package in R [83] was used with the number of trees being set to 120 after experimenting with different values.

5.6.4 XGBoost

XGBoost is a distributed gradient boosting technique that was developed by Chen and Guestrin [84] and initially released in 2014. It has gained significant popularity in machine learning platforms such as Kaggle due to the fact that it has performed better than other well known machine learning algorithms []. We selected XGBoost as one of the machine learning methods to be evaluated because of it was giving better results than the above mentioned algorithms in multiple domains and the time required to train models using XGBoost was comparatively less when compared to a method such as neural networks. The R implementation in [85] was used with linear regression as the objective function. Step size η was set at 0.05 with the maximum depth of a tree limited to 4. 4-fold cross-validation

was used to train the models with maximum number of rounds set at 10,000.

5.7 Summary

We took careful consideration to ensure that a same amount of effort was expended when tuning each of the different machine learning techniques so as not to give any unfair advantage to any single technique. After final tuning parameters were determined, the genetic algorithm based selection and optimization was scripted and run automatically without any intervention on an Intel Xeon Quad Core E5-1603 CPU at 2.8 GHz, with 64 GB of RAM. The model performance for each machine learning method was recorded along with the model itself for further comparison and evaluation. The accuracy of the prediction results and the analysis of the results is presented in chapters 6 and 7.

Chapter 6

RESULTS

In this chapter, we present results for all the experiments and evaluations described in chapters 3, 4, and 5. First of all, we present the visualizations from the mobility models, which we use to intuitively verify whether the underlying assumptions hold for this dataset. Then we go on to describe the results on correlation analysis to show that there is high correlation between human mobility and dengue incidence. Finally, we present the model performance for each machine learning technique. We also present the accuracy of predictions when considering only Colombo-MC MOH division, since Colombo is an outlier for our model due to its high disease incidence, and it is also a critical commercial and administrative hub through which dengue can propagate to the rest of the country quite easily.

6.1 Mobility Models

Mobility indicator values for each mobility model were visualized both temporally as well as spatially to verify that the results conformed to patterns observed in previous studies [86, 87, 69]. For temporal visualization of results, 4 MOH divisions were considered: Colombo-MC, Kandy, Galle and Jaffna. For spatial visualization, mobility data for the 32^{nd} week was selected, which falls in the beginning of August, which is approximately 2-3 months before a peak in dengue incidence observed that year.

For all of the derived mobility based models, we could see a significant decrease in mobility during week 14 to 16. This coincides with the traditional Sinhalese and Tamil new year celebrated in Sri Lanka, where it is customary for people to return to their ancestral homes to participate in the traditional celebrations.

In the probabilistic models described in section 4.2, the reduced mobility is evident for all 4 MOH divisions that were considered. However, it is interesting to note that for the visitation based model, which was based on the number

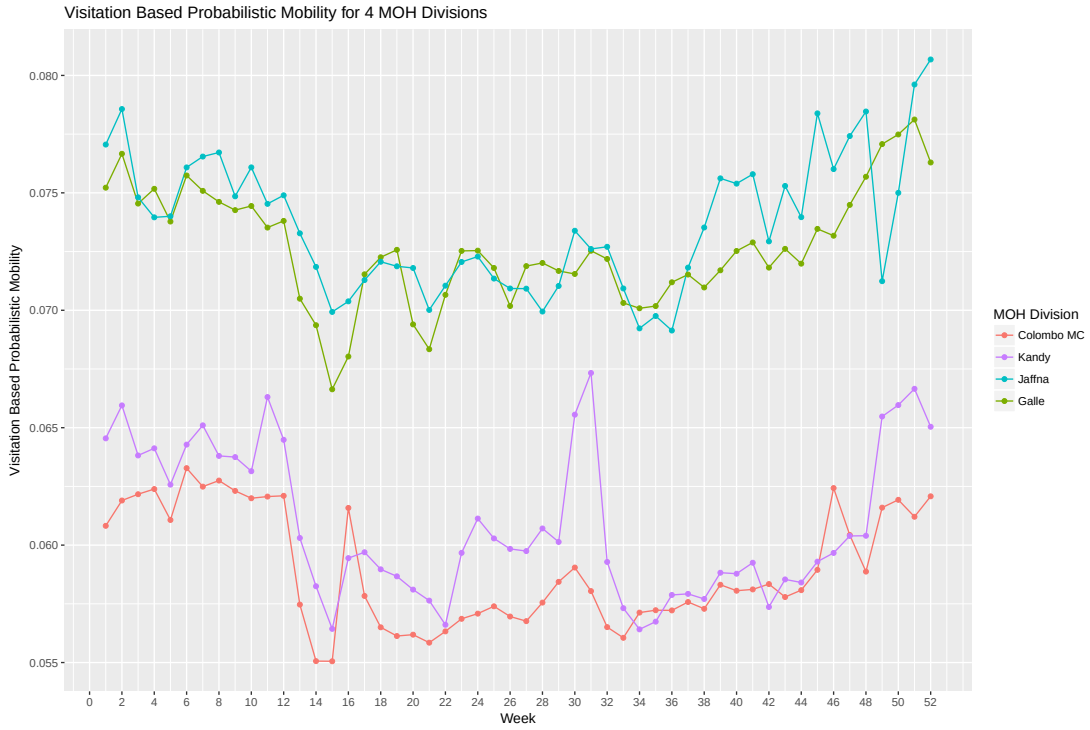


Figure 6-1: Visitation based probabilistic mobility for 4 MOH divisions

of people residing in other MOH divisions visiting a given MOH, there is a significant drop in mobility during Week 14-16 for Colombo-MC (Fig. 6-1), while for the exploration based model, which was based on residents of a particular MOH visiting other MOH divisions, there is an increase in mobility in week 16 in Colombo (Fig. 6-2. This confirms the expected pattern where people are moving out of Colombo, the urbanized commercial capital in Sri Lanka. It is noteworthy that even with the simplistic assumptions used to model human movement patterns in this case, our data driven models were able to capture this large-scale mobility dynamics quite well. In Fig. 6-1, we can also see an increase in visitation based mobility in Kandy during the weeks 31 and 32, which coincides with an international sporting event held during the same period.

When probabilistic mobility models are visualized for a single week for the entire country, we are able to observe significantly higher levels of mobility in urban regions such as Colombo and Kandy. Both exploration and visitation based models show a similar spread in intensity for the given week (See Figs. 6-3a, 6-3b).

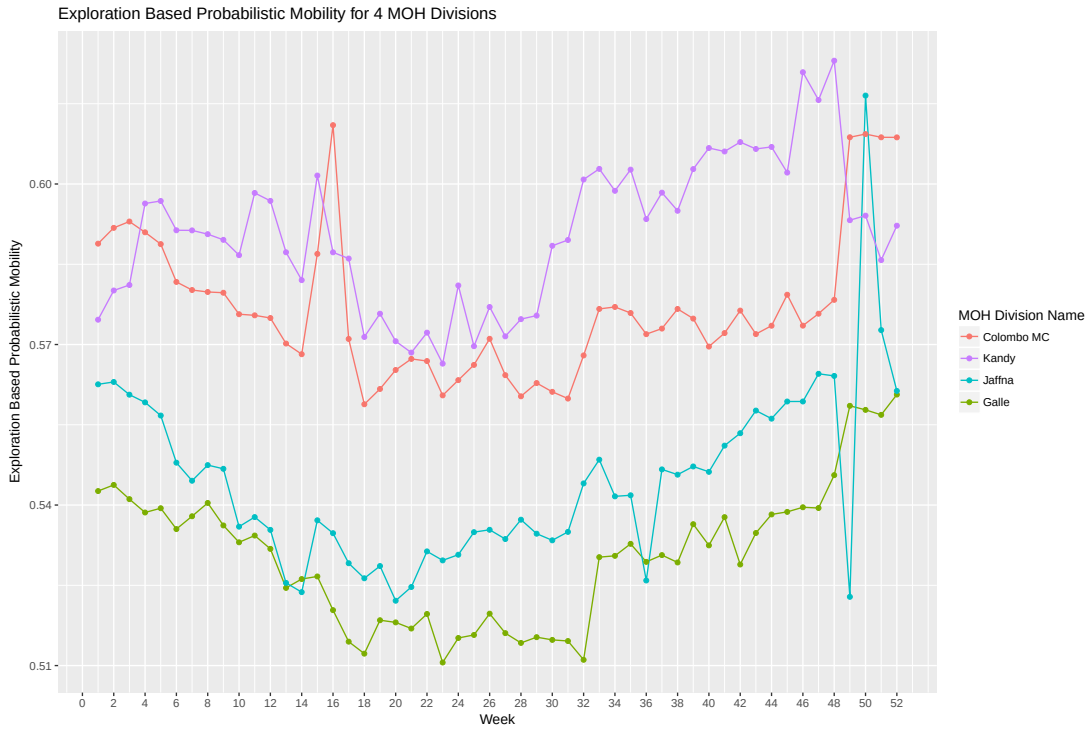
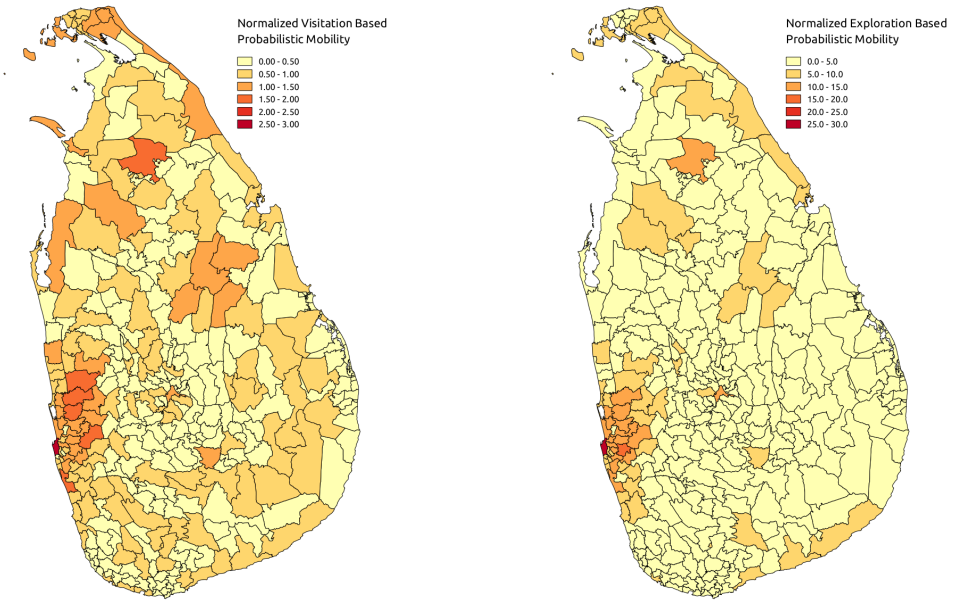


Figure 6-2: Exploration based probabilistic mobility for 4 MOH divisions



(a) Visitation based (b) Exploration based

Figure 6-3: Normalized probabilistic mobility - Week 32

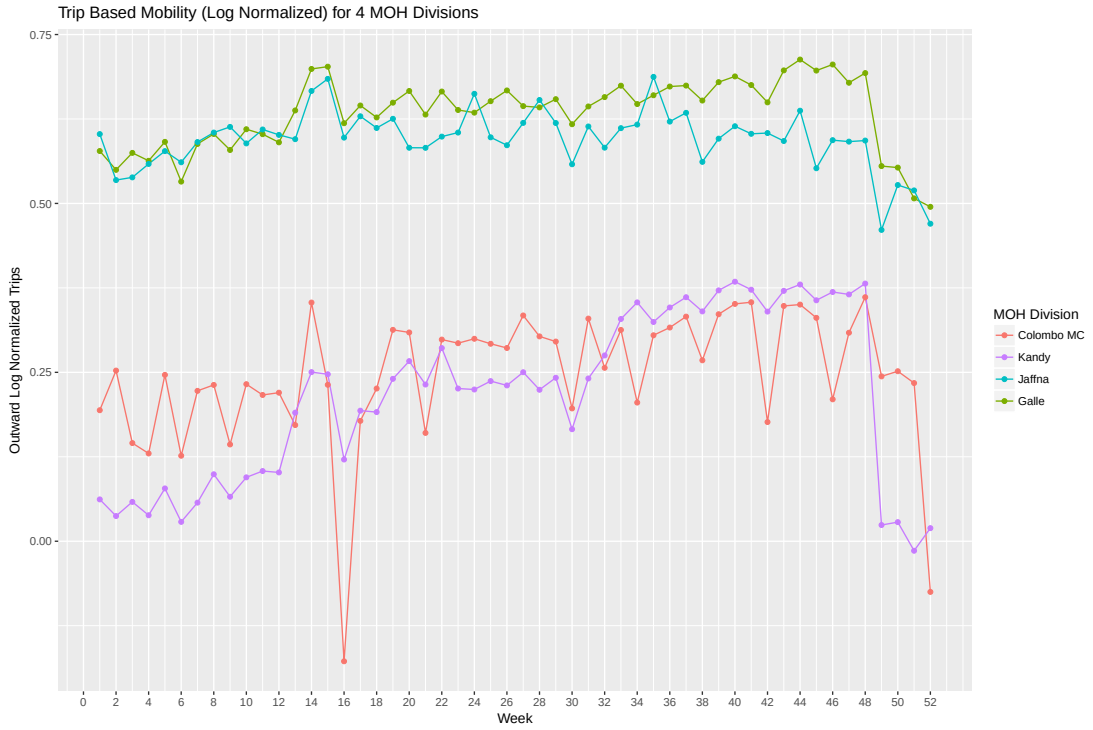


Figure 6-4: Trip based outward mobility for 4 MOH divisions

For trip based mobility, we visualize after getting the natural logarithm of the reported value. The significant decrease in mobility for week 16 is quite evident in this model as well (Figs. 6-4, 6-5). However, this model does not seem to capture the increase in presence at Kandy during week 31-32. However, it shows an increasing amount of mobility towards week 32-34 with a peak in week 34. Week 32-34 coincide with the Kandy Perehara, another traditional festival that is held in the month of August. Interestingly, the pattern for outward trips and inward trips is very much similar temporally (Figs. 6-5, 6-4) as well as spatially (Figs. 6-6a, 6-6b).

In the mobility based risk models as well, we see the decrease in mobility during the traditional new year period. Once again, we can see that both the models have similar values both temporally and spatially (See Figs. 6-7, 6-8, 6-9a, 6-9b)

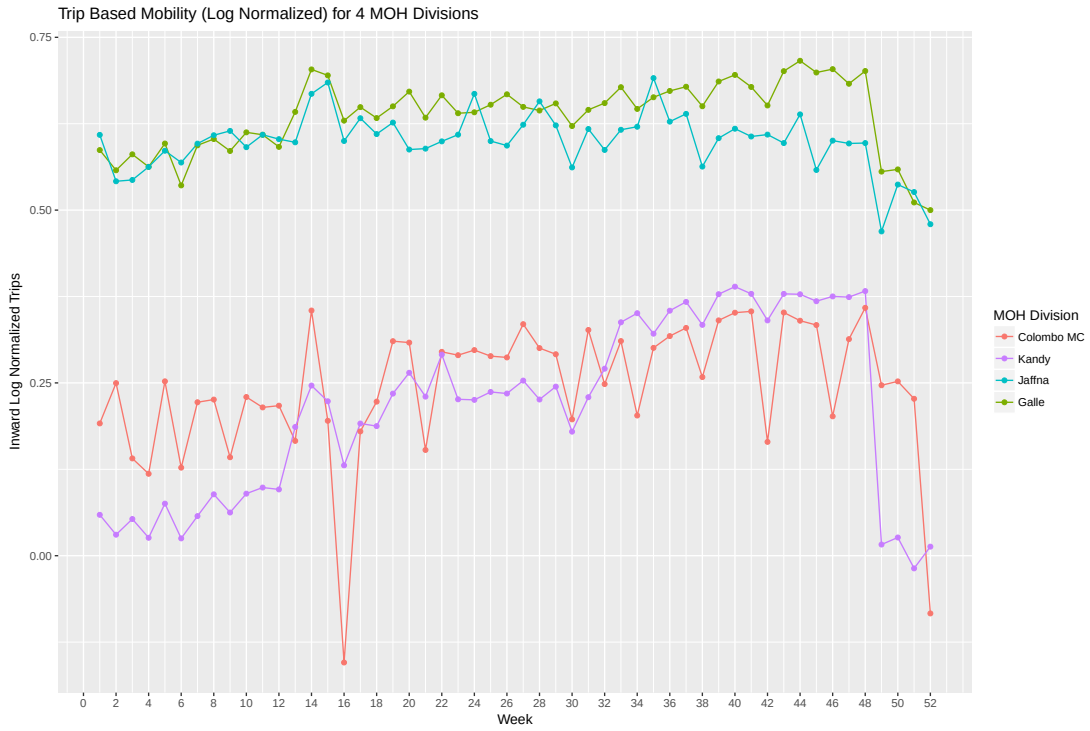
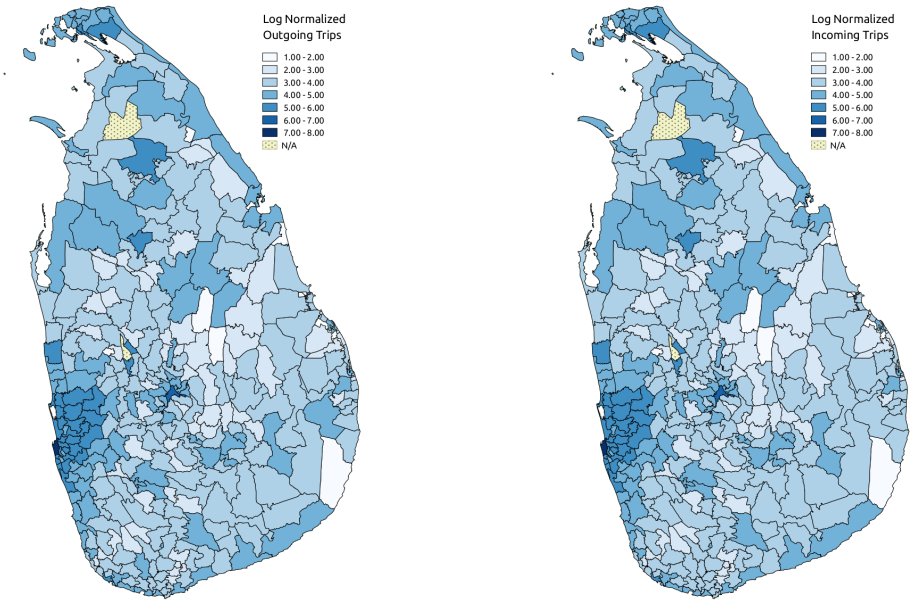


Figure 6-5: Trip based inward mobility for 4 MOH divisions



(a) Outward trips (b) Inward trips

Figure 6-6: Normalized log scaled trip mobility - Week 32

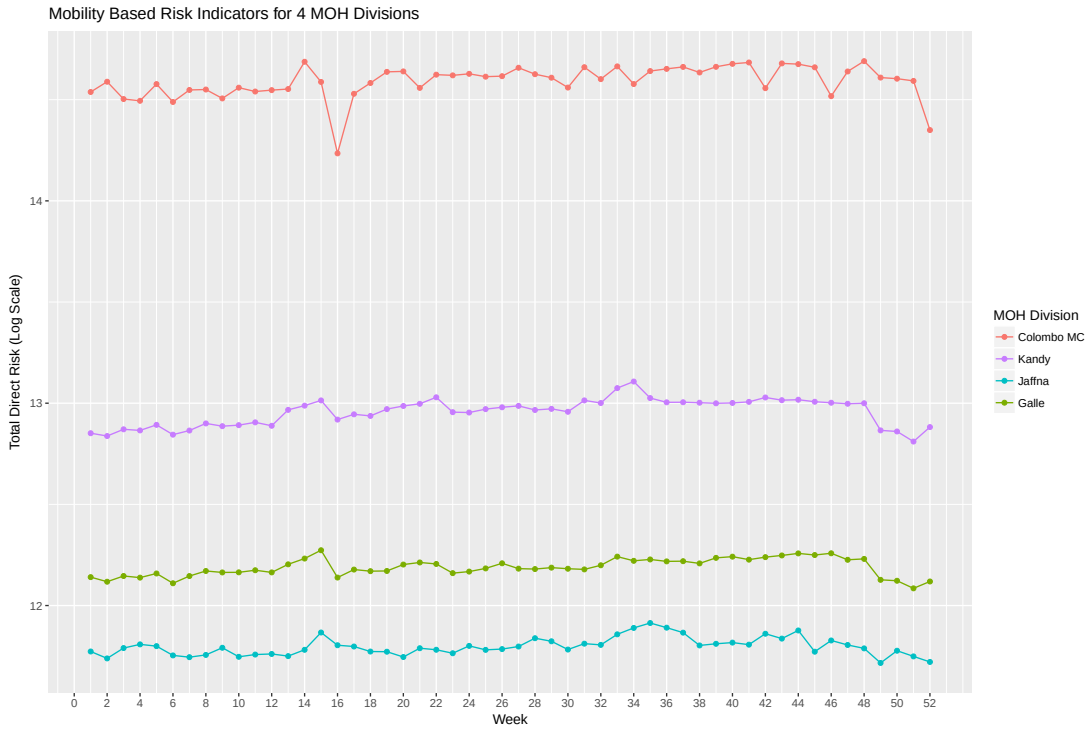


Figure 6-7: Mobility based total direct risk (log scale) for 4 MOH divisions

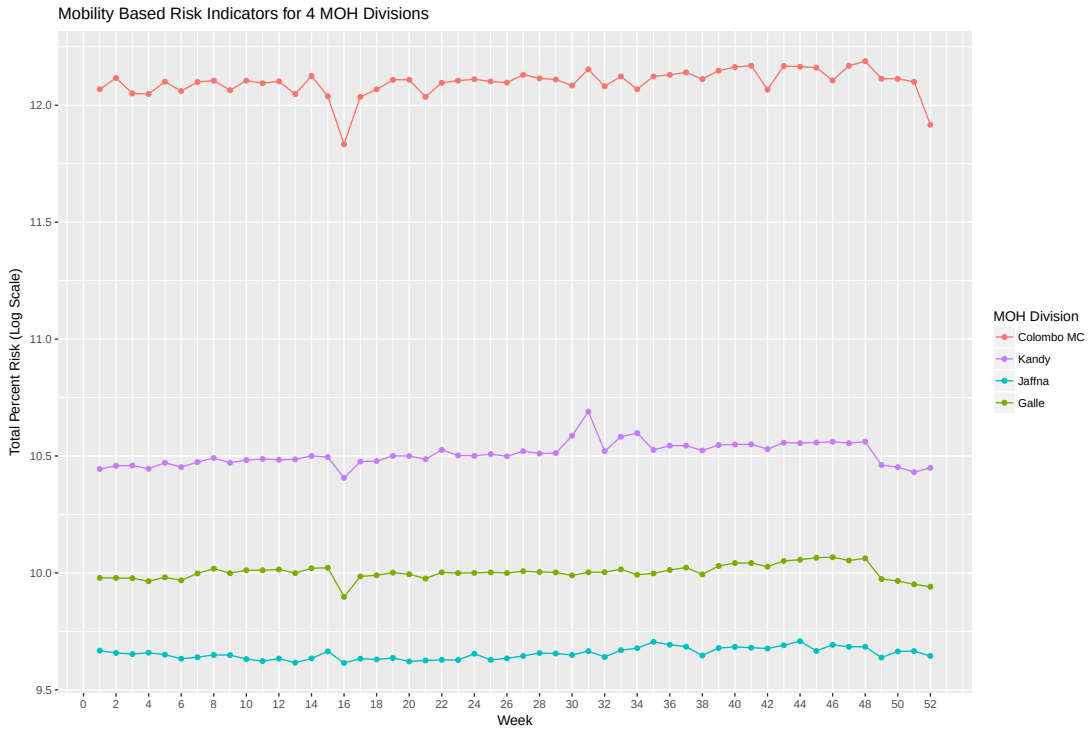
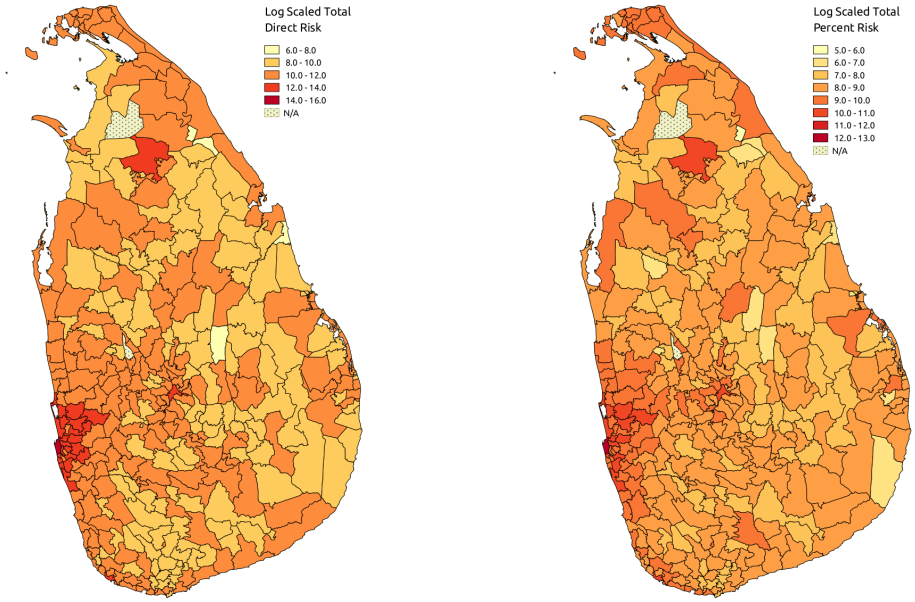


Figure 6-8: Mobility based total percent risk (log scale) for 4 MOH divisions



(a) Direct risk (b) Percent risk

Figure 6-9: Log scaled mobility based total risk - Week 32

6.2 Correlation Analysis

Pre-processed data for 20 MOH divisions was used to measure correlation between variables. Initially, no time-lagged features were considered and correlation between all variables was visualized. We could observe high correlation between different mobility features, as well as between related weather features such as mean temperature, min. temperature and max.temperature (Fig. 6-10.)

Since we are mainly interested in the correlation between dengue incidence for a given week against all the input features that can be used in building a model, correlation of a all possible input features (including time-lagged features) with dengue incidence was measured. We calculated Pearson's correlation, distance correlation and mutual information estimate for 133 potential predictor variables, which included multiple time-lagged features for the same data source as well. The highest correlated value from each separate input data source, as well as the time-lag in weeks for which the correlation value was reported against is presented in table 6.1 for all the 3 methods of correlation.

Pearson's correlation measured against dengue incidence did not show any

Table 6.1: Highest correlation with dengue incidence for each input data source using multiple methods of correlation

Input data source	**Pearson's Correlation**		**Distance Correlation**		**Mutual Information**	
	Time lag	**Abs. Value**	**Time lag**	**Value**	**Time lag**	**Value**
Past Dengue Incidence	2	0.889	2	0.880	2	0.443
Population	N/A	0.769	N/A	0.727	N/A	0.341
Visitation Based Probabilistic Mobility	11	0.311	12	0.415	11	0.237
Exploration Based Proabilistic Mobility	4	0.216	4	0.274	4	0.140
Outward Trip Mobility	9	0.061	11	0.213	2	0.148
Inward Trip Mobility	9	0.059	12	0.212	3	0.145
Mobility Based Direct Risk	6	0.130	12	0.182	8	0.120
Mobility Based Percent Risk	11	0.345	12	0.405	11	0.228
Mean NDVI	12	0.371	12	0.382	11	0.118
Mean Temperature	12	0.124	12	0.191	10	0.108
Maximum Temperature	12	0.058	4	0.187	6	0.115
Minimum Temperature	12	0.141	10	0.176	10	0.079
Total Precipitation	8	0.131	8	0.176	6	0.037

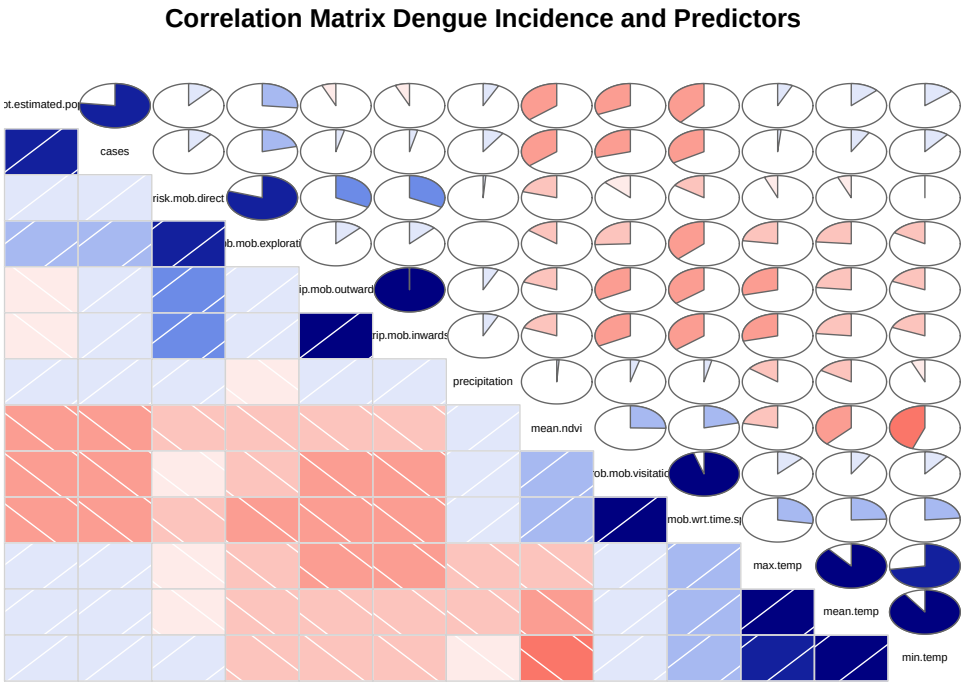


Figure 6-10: Pearson's correlation between variables (without time-lagged data)

features with high correlation except for the past dengue incidence. However, for distance correlation and mutual information estimate, derived mobility indicators showed significant correlation (Fig. 6-11).

6.3 Predictive Models

The results from training the data using the four machine learning techniques is summarized in table 6.2. From those results, XGBoost is the machine learning method with best performance, providing an overall RMSE value of 7.688 and an R^2 of 0.935. Performance of neural network models was significantly lower. But SVM and random forests had comparable performance when compared to the best performing XGBoost method.

For Colombo-MC, the RMSE and R^2 values were calculated separately and reported for each machine learning method. The results are presented in Table 6.3. Even for a single MOH division, we see that the best performing model is generated by XGBoost when the risk based mobility model is being used.

Table 6.2: Model Performance for 20 MOH divisions (GA - Genetic Algorithms, NFC - Without feature classes, FC - With feature classes)

Machine Learning Method	**Mobility Model**	**Without GA**		**GA NFC**		**GA FC**	
		RMSE	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
Support Vector Regression	N/A	10.034	0.889	8.961	0.911	9.037	0.91
	Probabilistic	10.082	0.888	9.061	0.909	9.044	0.91
	Trip based	10.161	0.886	9.074	0.909	9.076	0.909
	Risk based	10.078	0.888	9.050	0.909	9.018	0.91
	All	10.223	0.884	8.969	0.911	9.009	0.91
Random Forests	N/A	10.553	0.877	9.808	0.894	9.738	0.895
	Probabilistic	10.153	0.886	10.452	0.879	10.486	0.878
	Trip based	10.570	0.876	9.927	0.891	9.729	0.895
	Risk based	10.060	0.888	9.582	0.898	9.966	0.890
	All	10.658	0.874	10.472	0.879	10.114	0.887
XGBoost	N/A	10.077	0.888	8.035	0.929	8.153	0.926
	Probabilistic	10.102	0.887	8.158	0.926	8.06	0.928
	Trip based	9.977	0.89	8.089	0.928	7.791	0.933
	Risk based	9.678	0.896	7.910	0.931	**7.688**	**0.935**
	All	9.980	0.89	7.956	0.93	8.079	0.928
Neural Networks	N/A	33.006	-0.205	13.203	0.807	12.838	0.818
	Probabilistic	33.897	-0.271	14.153	0.778	17.214	0.672
	Trip based	33.959	-0.276	11.667	0.849	12.227	0.835
	Risk based	34.772	-0.338	15.373	0.739	13.808	0.789
	All	32.513	-0.169	12.816	0.818	13.245	0.806

Table 6.3: Model Performance for Colombo-MC MOH division (GA - Genetic Algorithms, NFC - Without feature classes, FC - With feature classes)

Machine Learning Method	**Mobility Model**	**Without GA**		**GA NFC**		**GA FC**	
		RMSE	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
Support Vector Regression	N/A	20.677	0.559	18.193	0.659	18.355	0.653
	Probabilistic	20.798	0.554	18.518	0.647	18.24	0.657
	Trip based	20.914	0.549	18.206	0.658	18.49	0.648
	Risk based	20.813	0.554	18.455	0.649	18.148	0.661
	All	21.002	0.546	18.234	0.657	18.214	0.658
Random Forests	N/A	21.997	0.501	19.956	0.59	19.841	0.594
	Probabilistic	20.944	0.548	21.393	0.528	21.437	0.526
	Trip based	21.817	0.510	20.002	0.588	19.562	0.606
	Risk based	20.491	0.567	19.159	0.622	20.136	0.582
	All	21.848	0.508	21.375	0.529	20.335	0.574
XGBoost	N/A	20.724	0.557	15.458	0.754	15.800	0.743
	Probabilistic	20.456	0.569	15.467	0.754	15.147	0.764
	Trip based	20.403	0.571	15.224	0.761	14.518	0.783
	Risk based	19.217	0.619	14.995	0.768	**14.241**	**0.791**
	All	19.914	0.591	14.781	0.775	14.903	0.771
Neural Networks	N/A	72.832	-4.466	27.293	0.233	25.509	0.33
	Probabilistic	74.476	-4.715	27.579	0.216	33.193	-0.135
	Trip based	74.907	-4.782	23.983	0.407	25.872	0.31
	Risk based	33.791	-0.177	30.989	0.010	26.918	0.253
	All	71.623	-4.286	26.435	0.280	27.671	0.211

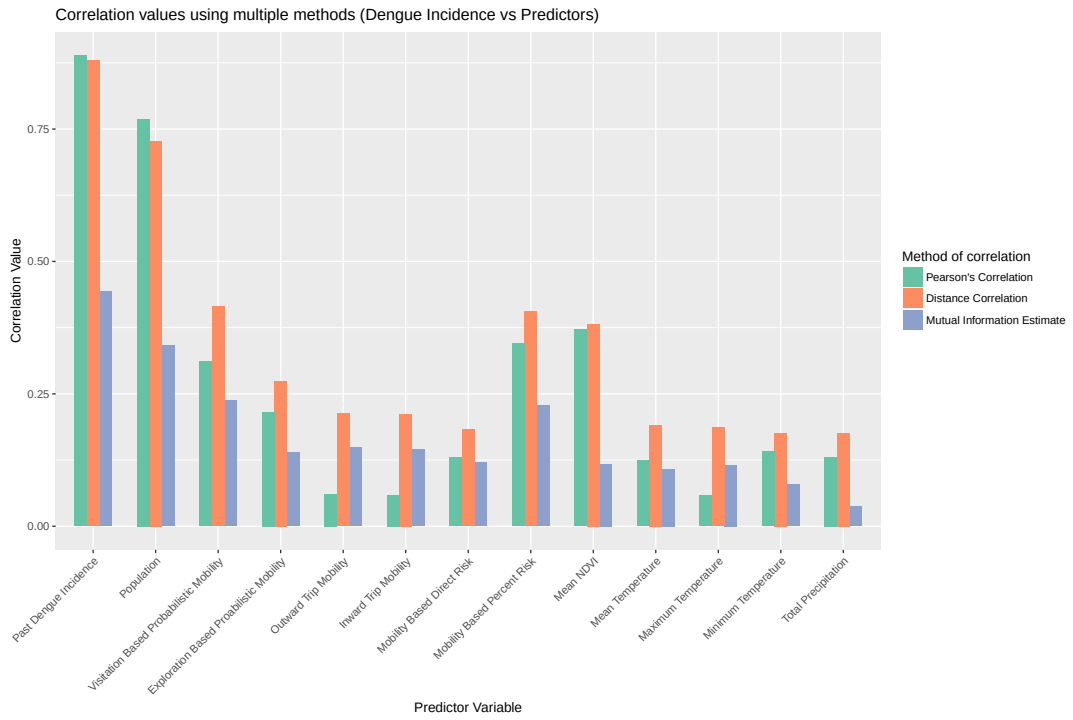


Figure 6-11: Correlation against dengue incidence using different methods

The predicted dengue incidence vs actual dengue incidence is visualized in Fig. 6-12, which shows that the prediction curve is able to detect the changes in the trend and closely follows the actual epidemic curve.

6.4 Summary

The results presented in this chapter demonstrate that the mobility models derived using CDRs are able to reflect the fluctuation in mobility patterns due to regional events (e.g. international sport events, traditional festivals) as well as national events (e.g. national festivals, holidays). The spatial visualization of these models demonstrate that the indicative values derived from our models are able to capture the regional hotspots where risk of dengue transmission is highest. In the correlation analysis, we see that mobility of the previous weeks is highly correlated with dengue incidence, while it is more pronounced in distance correlation and mutual information measures. The machine learning models were trained with and without mobility as an input because we needed to determine the impact of mobility in the accuracy of our results. The results show that introducing

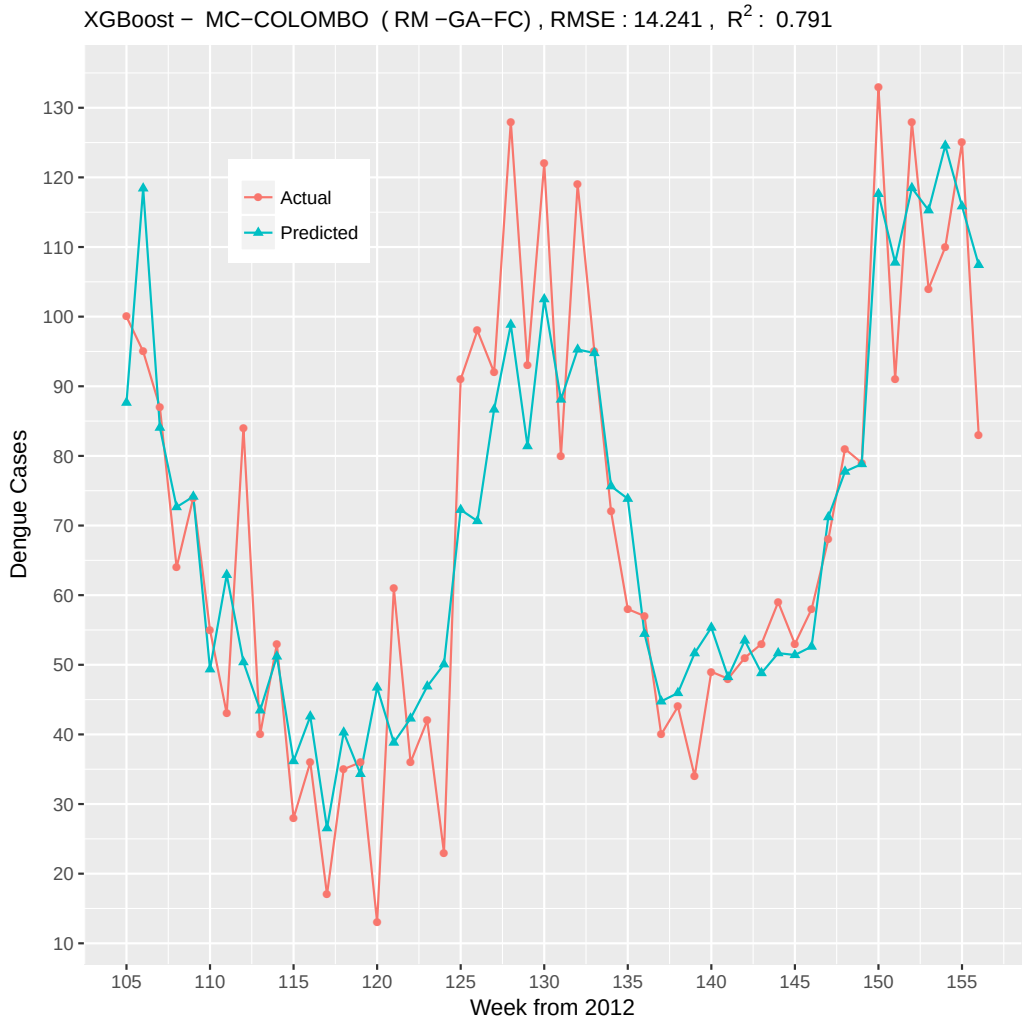


Figure 6-12: Dengue Incidence - Predicted vs Actual for year 2014 - Colombo MC

mobility provides an increase in accuracy, even though the increase is marginal in some cases. It is also interesting to note that the introduction of GA based feature selection significantly improved the prediction accuracy, while the introduction of feature classes while using GA based feature selection provided a further marginal improvement. Possible reasons behind these observations, as well as further interpretation of these results is discussed in the next chapter.

Chapter 7

DISCUSSION

This chapter attempts to highlight some of the challenges encountered in building these models for our study. Then we go on to interpret the results obtained from our mobility models and attempt to engage in a nuanced discussion regarding the possibility of using ubiquitous data sources such as CDRs to help disease outbreak prediction. After that, we discuss on the different performance levels obtained by different machine learning models, after which, we go on to detail the motivation and reasoning behind some of the optimizations introduced in our methodology. We also argue that there are 3 main contributions that are delivered as outcomes of this study. The first of them is the use of large-scale data processing techniques to develop CDR based multiple regional mobility measures that can be directly as an input for machine learning and statistical models. We also attempt to answer the question on whether human mobility has an impact on propagation of dengue in endemic regions. Finally, our study provides a comparison of performance for multiple machine learning methods for the same dataset, which we hope will aid in deciding which techniques are most suitable to be used for future research studies in a similar context.

7.1 Data collection and pre-processing

One of the biggest challenges for our study was in collecting data from multiple heterogenous data sources and transforming those datasets to a compatible format. Availability of weather data that covers a larger region would have allowed us to train using more data points, which in turn would have resulted in more robust prediction models. Another source of data that could have been used is remote sensing data, which can then be used to derive weather data. However, freely available existing data products need further processing that would require significant additional effort beyond the scope of our study. New research that

develop comprehensive methodologies to process such remote sensor data can alleviate this concern making such high resolution weather data available for future studies.

Vector surveillance data, and vector population data are other useful data sources that were difficult to obtain in a Sri Lankan context. Additionally, up to date and granular socio economic data sets would have also been useful as potential input features if it was made available publicly for academic purposes in Sri Lanka.

7.2 Mobility Models

Our study introduces 6 mobility models (3 main approaches with 2 sub models for each approach) that can be used as an input feature for any machine learning model that predicts propagation of infectious disease outbreaks. Each model has different assumptions which would affect its performance.

The probabilistic mobility models, while having a broad assumption of the amount of time spent by a subscriber at a given location being proportional to the number of calls initiated or recieved at that location, has performed relatively well when considering the results of predictive models and correlation analysis. In fact, visitation based probabilitic mobility has the highest distance correlation and mutual information score against dengue incidence, suggesting that people visiting a given region and spending more time in the preceding months can affect dengue incidence. Similarly, risk based mobility models also have high correlation, where as trip based mobility do not show significant correlation.

There can be multiple reasons for the trip based mobility models having only weak correlation with dengue incidence. In our model, a trip is simply defined as a change in tower location between two consecutive CDRs. This definition can capture locations that a subscriber might have only passed through, which ideally should not be taken into consideration when assessing risk of disease transmission. It might also indicate that the time spent by a visitor or the time band during which the visit happened contributes more towards the risk of transmission rather

Table 7.1: t-test on improvement of predictive accuracy due to mobility (X = set of error terms with mobility, Y = set of error terms without mobility)

t-value	Degrees of freedom	p-value	95% confidence interval	Mean of X	Mean of Y
-2.001	14379	0.0454	-0.6605 to -0.0068	5.624	5.9577

than the act of simply visiting a given location.

From the results, we can also see that there is not much of a difference in the predictive accuracy when comparing the risk based model vs the probabilistic model. This might be due to the fact that we failed to pick the optimal values for the time bands and the weights assigned to the different time bands since this was done only using our intuition. However, the fact that the probabilistic models also performed comparatively as well as the risk based models suggest that complex models do not necessarily translate into better predictive accuracy. However, the highest predictive accuracy was observed when risk based mobility was used, making it the best model that captures risk of disease transmission due to mobility.

7.3 Impact of mobility on predictive accuracy

The results suggest a consistent improvement in RMSE and R^2 values when mobility is introduced into the model. However, this improvement is not very pronounced. This might be due to the fact that dengue is endemic in most regions of Sri Lanka, and human mobility is not the primary factor in introducing dengue to a given region. However, the consistent improvement in accuracy when mobility was introduced, as well as the comparatively high correlation observed between dengue incidence and various mobility measures, suggest that mobility does contribute to dengue incidence even in an endemic setting. In order to determine whether the impact on the predictive accuracy of disease incidence was significant due to mobility, we ran Welch's two sample t-test to measure the p-value at a significane level of 0.05 for a two-sided hypothesis. The results of this test are available in table 7.1.

7.4 Comparison of machine learning methods

Out of the different machine learning methods we used to predict dengue incidence, XGBoost provided the best RMSE and R^2 values. It is interesting to note that there was a distinctive difference in predictive performance when considering each of the techniques in the final round of results. For the final results XGBoost had R^2 values in the range of 0.926 to 0.935, while SVR had a range of 0.909 to 0.91, and random forests had a range between 0.878 to 0.895. However, this difference in performance became apparent after the genetic algorithm based optimization was applied. Before the optimization, the difference in performance was not that pronounced, as evidenced by R^2 measures of 0.887 to 0.896 for XGBoost, 0.884 to 0.889 for SVR, and 0.874 to 0.888 for RF. Under each of these conditions, neural networks consistently provided poorer performance, and was not able to match the performance of other 3 techniques even after the GA based optimization was applied.

7.5 Genetic algorithm based optimization

The GA based feature selection technique that was based off prior work by Wu et. al [29] provided significant improvement in predictive accuracy in all of our models. We did several modifications to the original technique by optimizing R^2 measure and RMSE measures simultaneously, introducing the concept of feature classes to make sure that no feature is completely dropped out of the model, and automating the process of training and selecting the final model without manually selecting the features that appeared most frequently in multiple runs (as was the case in the study by Wu et. al). Applying this technique yielded significant improvements to model accuracy across all 4 machine learning methods.

7.6 Summary

The discussion above highlights multiple areas where our study could have benefited from if certain data sources were made available. It also points to potential

research directions that can help to bridge the gap in some of the data products required to build more robust and large-scale forecasting models. However, using the available data products, we were able to demonstrate that accurate predictive models can be developed while also contributing to multiple application areas that cut across several academic disciplines. Mainly, our study demonstrates the feasibility of using large-scale ubiquitous datasets to model human mobility at regional level for disease outbreak predictions. Our results also indicate that human mobility has a significant impact on dengue propagation, even in an endemic setting. Finally, the results provide a comparison between machine learning methods for predicting infectious disease outbreaks in Sri Lanka. Using the points raised in this discussion, the main conclusions of this study are detailed in Chapter 8.

Chapter 8

CONCLUSION

Our models are able to forecast weekly number of dengue incidence 2 weeks ahead of time, with very good accuracy (RMSE - 7.688, R^2 - 0.935) for the selected MOH divisions. This shows that the methodology established in this work is feasible and can be applied practically. Additionally, we explored 3 models of human mobility, derived using pseudonymized CDRs and demonstrated the feasibility of such novel sources of data in improving the accuracy of disease forecasting models. We also showed that the inclusion of human mobility improved the accuracy of the models significantly (p-value = 0.0454, 0.95 % confidence interval), suggesting that human mobility has an impact on dengue incidence, even when the disease is already endemic to a specific region.

Out of the different mobility models and machine learning methods utilized, our study shows that the risk based mobility model developed by us performs best, and XGBoost has the best accuracy for our dataset. The study was done using 3 years of worth of dengue incidence data and other data sources were fused with modifications as necessary, depending on the time span for which the data was available. With more recent disease incidence data and CDR data covering a larger timespan, we should be able to increase the predictive accuracy and get good results at an individual MOH level. For this, a partnership with relevant government health agencies and research organizations should be established to utilize this work practically. Such a partnership would provide an opportunity for public health officials to provide details on how the output of our models should be structured and also allow them to give input on weighting of risk scores, which can ultimately result in a near real time system that provides risk of dengue outbreaks at national scale.

Chapter 9

FUTURE WORK

There are several aspects that could have been explored to further improve the accuracy of our predictions. Our study did not consider any deep learning techniques, which might have provided better predictions. However, the disease incidence data, CDRs and weather data was available only for a limited timespan, limiting the applicability of deep learning techniques, which in general need larger datasets. If a data sharing partnership could be established between data owners and research insitutes, with the availability of more data, deep learning would definitely be a viable option and a valuable future research direction.

Additionally, an optimization technique such as genetic algorithms or other auto tuning methods can be used to set the hyperparameters. The training time for each of these models would increase if such an approach is adapted, but would be expected to yield better results. Another research direction that can be explored is in determining the optimal weights to be assigned to different time-bands when assigning risk scores for the mobility models. These risk scores can be optimally assigned by performing a sensitivity analysis to determine which weights yield an aggregate risk measure for a given MOH division that correlates best with its dengue incidence history.

References

[1] Izabela A. Rodenhuis-Zybert, Jan Wilschut, and Jolanda M. Smit. Dengue virus life cycle: Viral and host factors modulating infectivity. *Cellular and Molecular Life Sciences*, 67(16):2773–2786, 2010.

[2] Dennis Normile. Surprising new dengue virus throws a spanner in disease control efforts. *American Association for the Advancement of Science*, 2013.

[3] Samir Bhatt, Peter W Gething, Oliver J Brady, Jane P Messina, Andrew W Farlow, Catherine L Moyes, John M Drake, John S Brownstein, Anne G Hoen, Osman Sankoh, Monica F Myers, Dylan B George, and Thomas Jaenisch. The global distribution and burden of dengue. *Nature*, 496(7446):504–507, 2013.

[4] Jeffrey D Stanaway, Donald S Shepard, Eduardo A Undurraga, Yara A Halasa, Luc E Coffeng, Oliver J Brady, Simon I Hay, Neeraj Bedi, Isabela M Bensenor, Carlos A Castañeda-Orjuela, et al. The global burden of dengue: an analysis from the global burden of disease study 2013. *The Lancet infectious diseases*, 16(6):712–723, 2016.

[5] Neil Thalagala, Hasitha Tissera, Paba Palihawadana, Ananda Amarasinghe, Anuradha Ambagahawita, Annelies Wilder-Smith, Donald S Shepard, and Yeşim Tozan. 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):e0004466, 2016.

[6] Marta Sarzynska, Oyita Udiani, and Na Zhang. A study of gravity-linked metapopulation models for the spatial spread of dengue fever. *arXiv preprint arXiv:1308.4589*, 2008:1–32, 2013.

[7] Raúl Isea and Karl E Lonngren. A preliminary mathematical model for the dynamic transmission of dengue, chikungunya and zika. *American Journal of Modern Physics and Application*, 3(2):11–15, 2016.

[8] Hani M. Aburas, B. Gultekin Cetiner, and Murat Sari. Dengue confirmed-cases prediction: A neural network model. *Expert Systems with Applications*, 37(6):4256–4260, 2010.

[9] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6696 LNCS(June):133–151, 2011.

[10] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. Human mobility modeling at metropolitan scales. *Proceedings of the 10th international conference on Mobile systems, applications, and services - MobiSys '12*, page 239, 2012.

[11] Yuzuru Tanahashi, James R. Rowland, Stephen North, and Kwan-Liu Ma. Inferring human mobility patterns from anonymized mobile communication usage. *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia - MoMM '12*, page 151, 2012.

[12] Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C González, and Vittoria Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS Computational Biology*, 10(7):e1003716, 2014.

[13] Amy Wesolowski, Gillian Stresman, Nathan Eagle, Jennifer Stevenson, Chrispin Owaga, Elizabeth Marube, Teun Bousema, Christopher Drakeley, Jonathan Cox, and Caroline O Buckee. Quantifying travel behavior for

infectious disease research: a comparison of data from surveys and mobile phones. *Scientific reports*, 4:5678, 2014.

[14] Amy Wesolowski, Taimur Qureshi, Maciej F Boni, Pål Roe Sundsøy, Michael A Johansson, Syed Basit Rasheed, Kenth Engø-Monsen, and Caroline O Buckee. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*, 112(38):11887–11892, 2015.

[15] Linus Bengtsson, Jean Gaudart, Xin Lu, Sandra Moore, Erik Wetter, Kankoe Sallah, Stanislas Rebaudet, and Renaud Piarroux. Using mobile phone data to predict the spatial spread of cholera. *Scientific reports*, 5:8923, 2015.

[16] A Rudnick. Aedes aegypti and haemorrhagic fever. *Bulletin of the World Health Organization*, 36(4):528, 1967.

[17] Scott B Halstead. Mosquito-borne haemorrhagic fevers of south and south-east asia. *Bulletin of the World Health Organization*, 35(1):3, 1966.

[18] PM Sheppard, WW Macdonald, RJ Tonn, and B Grab. The dynamics of an adult population of aedes aegypti in relation to dengue haemorrhagic fever in bangkok. *The journal of animal ecology*, pages 661–702, 1969.

[19] Albert Rudnick and YC Chan. Dengue type 2 virus in naturally infected aedes albopictus mosquitoes in singapore. *Science*, 149(3684):638–639, 1965.

[20] Moritz U G Kraemer, Marianne E. Sinka, Kirsten A. Duda, Adrian Q N Mylne, Freya M. Shearer, Christopher M. Barker, Chester G. Moore, Roberta G. Carvalho, Giovanini E. Coelho, Wim Van Bortel, Guy Hendrickx, Francis Schaffner, Iqbal Rf Elyazar, Hwa Jen Teng, Oliver J. Brady, Jane P. Messina, David M. Pigott, Thomas W. Scott, David L. Smith, G. R. William Wint, Nick Golding, and Simon I. Hay. The global distribution of the arbovirus vectors Aedes aegypti and Ae. Albopictus. *eLife*, 4(JUNE2015):1–18, 2015.

[21] Yien Ling Hii, Joacim RocklÃűv, Nawi Ng, Choon Siang Tang, Fung Yin Pang, and Rainer Sauerborn. Climate variability and increase in intensity and magnitude of dengue incidence in Singapore. *Global Health Action*, 2(1):1–9, 2009.

[22] Kensuke Goto, Balachandran Kumarendran, Sachith Mettananda, Deepa Gunasekara, Yoshito Fujii, Satoshi Kaneko, S Rajapakse, C Rodrigo, A Rajapakse, JL Rasgon, OJ Brady, PW Gething, S Bhatt, JP Messina, JS Brownstein, C Chastel, SC Weaver, N Vasilakis, HP Mohammed, MM Ramos, A Rivera, M Johansson, JL Munoz-Jordan, A Egbendewe-Mondzozo, M Musumba, BA McCarl, X Wu, F Huang, S Zhou, S Zhang, H Wang, L Tang, U Haque, M Hashizume, GE Glass, AM Dewan, HJ Overgaard, G Constantin de Magny, W Thiaw, V Kumar, NM Manga, BM Diop, SL Traerup, RA Ortiz, A Markandya, G Constantin de Magny, R Murtugudde, MR Sapiano, A Nizam, CW Brown, T Ben-Ari, S Neerinckx, KL Gage, K Kreppel, A Laudisoit, L Xu, Q Liu, LC Stige, T Ben Ari, X Fang, TB Ari, A Gershunov, R Tristan, B Cazelles, K Gage, GC de Magny, W Thiaw, V Kumar, NM Manga, BM Diop, ME Reller, C Bodinayake, A Nagahawatte, V Devasiri, W Kodikara-Arachichi, KG Weerakoon, SA Kularatne, DH Edussuriya, SK Kodikara, LP Gunatilake, HA Tissera, EE Ooi, DJ Gubler, Y Tan, B Logendra, N Kanakaratne, WM Wahala, WB Messer, HA Tissera, A Shahani, SA Kularatne, MM Pathirage, PV Kumarasiri, S Gunasena, SI Mahindawanse, S Kumar, S Managi, A Matsuda, R Opgen-Rhein, K Strimmer, OA Akinboade, LA Braimoh, H Pesaran, Y Shin, X-j Ji, Y-q Zhang, L-y Hao, M Gharbi, P Quenel, J Gustave, S Cassadou, G La Ruche, E Descloux, M Mangeas, CE Menkes, M Lengaigne, A Leroy, E Pinto, M Coelho, L Oliver, E Massad, YL Hii, J Rocklov, N Ng, CS Tang, and FY Pang. Analysis of Effects of Meteorological Factors on Dengue Incidence in Sri Lanka Using Time Series Data. *PLoS ONE*, 8(5):e63717, 2013.

[23] Melanie Bannister-Tyrrell, Craig Williams, Scott A. Ritchie, Gina Rau,

Janette Lindesay, Geoff Mercer, and David Harley. Weather-driven variation in dengue activity in Australia examined using a process-based modeling approach. *American Journal of Tropical Medicine and Hygiene*, 88(1):65–72, 2013.

[24] H M Yang, M L G Macoris, K C Galvani, M T M Andrighetti, and D M V Wanderley. Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue. *Epidemiology and infection*, 137(8):1188–1202, 2009.

[25] Jing Liu-Helmersson, Hans Stenlund, Annelies Wilder-Smith, and Joacim RocklÃűv. Vectorial capacity of Aedes aegypti: Effects of temperature and implications for global dengue epidemic potential. *PLoS ONE*, 9(3), 2014.

[26] Oliver J Brady, Nick Golding, David M Pigott, Moritz U G Kraemer, Jane P Messina, Robert C Reiner Jr, Thomas W Scott, David L Smith, Peter W Gething, and Simon I Hay. Global temperature constraints on Aedes aegypti and Ae. albopictus persistence and competence for dengue virus transmission. *Parasites & Vectors*, 7(1):338, 2014.

[27] Suchithra Naish, Pat Dale, John S Mackenzie, John McBride, Kerrie Mengersen, and Shilu Tong. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *BMC infectious diseases*, 14(1):167, 2014.

[28] Wenbiao Hu, Archie Clements, Gail Williams, and Shilu Tong. Dengue fever and el nino/southern oscillation in queensland, australia: a time series predictive model. *Occupational and environmental medicine*, 67(5):307–311, 2010.

[29] Yan Wu, Gary Lee, Xiuju J Fu, and Terence Hung. 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.

[30] Yien Ling Hii, Huaiping Zhu, Nawi Ng, Lee Ching Ng, and Joacim RocklÃűv. Forecast of Dengue Incidence Using Temperature and Rainfall. *PLoS Neglected Tropical Diseases*, 6(11), 2012.

[31] Myriam Gharbi, Philippe Quenel, Joël Gustave, Sylvie Cassadou, Guy La Ruche, Laurent Girdary, and Laurence Marrama. Time series analysis of dengue incidence in guadeloupe, french west indies: forecasting models using climate variables as predictors. *BMC infectious diseases*, 11(1):166, 2011.

[32] Edna Pinto, Micheline Coelho, Leuda Oliver, and Eduardo Massad. The influence of climate variables on dengue in singapore. *International journal of environmental health research*, 21(6):415–426, 2011.

[33] Prasad Liyanage, Hasitha Tissera, Maquins Sewe, Mikkel Quam, Ananda Amarasinghe, Paba Palihawadana, Annelies Wilder-Smith, Valérie Louis, Yesim Tozan, and Joacim Rocklöv. A spatial hierarchical analysis of the temporal influences of the el nino-southern oscillation and weather on dengue in kalutara district, sri lanka. *International journal of environmental research and public health*, 13(11):1087, 2016.

[34] L E Muir and B H Kay. Aedes aegypti survival and dispersal estimated by mark-release-recapture in northern Australia. *The American journal of tropical medicine and hygiene*, 58(3):277–82, 1998.

[35] Harvey B Morlan and Richard O Hayes. Urban dispersal and activity of aedes aegypti. *Mosq News*, 18:137–144, 1958.

[36] Nildimar Alves Honório, Wellington da Costa Silva, Paulo José Leite, Jaylei Monteiro Gonçalves, Leon Philip Lounibos, and Ricardo Lourenço-de Oliveira. Dispersal of Aedes aegypti and Aedes albopictus (Diptera: Culicidae) in an urban endemic dengue area in the State of Rio de Janeiro, Brazil. *Memórias do Instituto Oswaldo Cruz*, 98(2):191–198, 2003.

[37] Dirk Brockmann. Human Mobility and Spatial Disease Dynamics. *Reviews of Nonlinear Dynamics and Complexity*, 2:1–24, 2010.

[38] VV Belik, Theo Geisel, and Dirk Brockmann. The impact of human mobility on spatial disease dynamics. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 932–935. IEEE, 2009.

[39] Vitaly Belik, Theo Geisel, and Dirk Brockmann. Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases. *Physical Review X*, 1(1):1–5, 2011.

[40] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779, 2008.

[41] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, James Rowland, and Alexander Varshavsky. A tale of two cities. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, pages 19–24. ACM, 2010.

[42] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science (New York, N.Y.)*, 338(6104):267–70, 2012.

[43] Flavio Finger, Tina Genolet, Lorenzo Mari, Guillaume Constantin de Magny, Noël Magloire Manga, Andrea Rinaldo, and Enrico Bertuzzo. 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.

[44] José Lourenço and Mario Recker. The 2012 Madeira Dengue Outbreak: Epidemiological Determinants and Future Epidemic Potential. *PLoS Neglected Tropical Diseases*, 8(8), 2014.

[45] Murali Krishna Enduri and Shivakumar Jolad. Spatial Patterns of Spread of Dengue with Human and Vector Mobility. *arXiv preprint arXiv:1409.0965v1*, 2014.

[46] Líliam César de Castro Medeiros, Cesar Augusto Rodrigues Castilho, Cynthia Braga, Wayner Vieira de Souza, Leda Regis, and Antonio Miguel Vieira Monteiro. Modeling the dynamic transmission of dengue fever: investigating disease persistence. *PLOS Neglected Tropical Diseases*, 5(1):e942, 2011.

[47] WO Kermack and McKendrick AG. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, 1927.

[48] Diana Knipl. A new approach for designing disease intervention strategies in metapopulation models. *Journal of biological dynamics*, 10(1):71–94, 2016.

[49] Yien Ling Hii, Joacim Rocklov, Stig Wall, Lee Ching Ng, Choon Siang Tang, and Nawi Ng. Optimal Lead Time for Dengue Forecast. *PLoS Neglected Tropical Diseases*, 2012.

[50] S. Promprou, M. Jaroensutasinee, and K. Jaroensutasinee. Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA Models. *Dengue Bulletin*, 30:99–106, 2006.

[51] S. Wongkoon, M. Jaroensutasinee, and K. Jaroensutasinee. Assessing the temporal modelling for prediction of dengue infection in northern and northeastern, Thailand. *Tropical Biomedicine*, 29(3):339–348, 2012.

[52] W G van Panhuis, M Choisy, X Xiong, N S Chok, P Akarasewi, S Iamsirithaworn, S K Lam, C K Chong, F C Lam, B Phommasak, P Vongphrachanh, K Bouaphanh, H Rekol, N T Hien, P Q Thai, T N Duong, J H Chuang, Y L Liu, L C Ng, Y Shi, E A Tayag, V G Roque Jr., L L Lee Suy, R G Jarman, R V Gibbons, J M Velasco, I K Yoon, D S Burke, and D A Cummings. Region-wide synchrony and traveling waves of

dengue across eight countries in Southeast Asia. *Proc Natl Acad Sci U S A*, 112(42):13069–13074, 2015.

[53] W P T M Wickramaarachchi, S. S N Perera, and S. Jayasinghe. 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.

[54] Napa Rachata, Phasit Charoenkwan, Thongchai Yooyativong, Kosin Chamnongthai, Chidchanok Lursinsap, and Kohji Higuchi. 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*, pages 210–214, 2008.

[55] Yuhanis Yusof and Zuriani Mustaffa. Dengue Outbreak Prediction : A Least Squares Support Vector Machines Approach. *International Journal of Computer Theory and Engineering*, 3(4):489–493, 2011.

[56] Rakesh Kaundal, Amar S Kapoor, and Gajendra P S Raghava. Machine learning techniques in disease forecasting: a case study on rice blast prediction. *BMC bioinformatics*, 7:485, 2006.

[57] Vijeta Sharma, Ajai Kumar, Dr Lakshmi Panat, Ganesh Karajkhede, et al. Malaria outbreak prediction model using machine learning. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 4(12), 2015.

[58] Nabeel Abdur Rehman, Shankar Kalyanaraman, Talal Ahmad, Fahad Pervaiz, Umar Saif, and Lakshminarayanan Subramanian. Fine-grained dengue forecasting using telephone triage services. *Science Advances*, 2(7):1–10, 2016.

[59] Yuan Shi, Xu Liu, Suet-Yheng Kok, Jayanthi Rajarethinam, Shaohong Liang, Grace Yap, Chee-Seng Chong, Kim-Sung Lee, Sharon Sy Tan,

Christopher Kuan Yew Chin, Andrew Lo, Waiming Kong, Lee Ching Ng, and Alex R Cook. Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts and Policy Decision Support in Singapore. *Environmental health perspectives*, 124(November):1369–1375, 2015.

[60] DO Gerardi and LHA Monteiro. System identification and prediction of dengue fever incidence in rio de janeiro. *Mathematical Problems in Engineering*, 2011, 2011.

[61] Enrique Frias-Martinez, Graham Williamson, and Vanessa Frias-Martinez. An Agent-Based Model of Epidemic Spread using Human Mobility and Social Network Information. *3rd International Conference on Social Computing (SocialCom'11)*, pages 49–56, 2011.

[62] Anna L Buczak, Phillip T Koshute, Steven M Babin, Brian H Feighner, and Sheryl H Lewis. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC medical informatics and decision making*, 12:124, 2012.

[63] Epidemiology Unit - Ministry of Health. 'Distribution of Notification(H399) Dengue Cases by Month', 2017. [Online]. Available: `http://epid.gov.lk/web/index.php?option=com{\_}casesanddeaths{\&}Itemid=448{\&}lang=en`. [Accessed: 07-Jul-2019].

[64] Department of Meteorology, Sri Lanka. Meteorological Information for Sri Lanka, 2014, 2015.

[65] darksky.net. 'Dark Sky API', 2016. [Online]. Available: `https://darksky.net/dev`. [Accessed: 07-Jul-2019].

[66] Stephen A Del Greco, Neal Lott, Kathy Hawkins, Rich Baldwin, Dee Dee Anders, Ron Ray, Dan Dellinger, Pete Jones, and Fred Smith. Surface data integration at NOAA's National Climatic Data Center: data format, processing, QC, and product generation. 2006.

[67] K Didan. MOD13Q1 MODIS/Terra vegetation indices 16-day L3 global 250m SIN grid V006. *NASA EOSDIS Land Processes DAAC*, 2015.

[68] Stef van Buuren and Karin Groothuis-Oudshoorn. Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

[69] Sriganesh Lokanathan, Gabriel E Kreindler, NH Nisana de Silva, Yuhei Miyauchi, Dedunu Dhananjaya, and Rohan Samarajiva. The potential of mobile network big data as a tool in colombo's transportation and urban planning. *Information Technologies & International Development*, 12(2):pp–63, 2016.

[70] Danaja Maldeniya, Sriganesh Lokanathan, and Amal Kumarage. Origin-Destination Matrix Estimation for Sri Lanka Using the Four Step Model. *Proceedings of the 13th International Conference on Social Implications of Computers in Developing Countries*, pages 785–794, 2015.

[71] Steven T Stoddard, Amy C Morrison, Gonzalo M Vazquez-Prokopec, Valerie Paz Soldan, Tadeusz J Kochel, Uriel Kitron, John P Elder, and Thomas W Scott. The Role of Human Movement in the Transmission of Vector-Borne Pathogens. *PLoS Neglected Tropical Diseases*, 3(7), 2009.

[72] denguevirusnet.com. 'Life Cycle of Dengue Mosquito Aedes aegypti'. [Online]. Available: `http://www.denguevirusnet.com/life-cycle-of-aedes-aegypti.html` [Accessed: 2018-09-28].

[73] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2017. R package version 6.0-77.

[74] K Pearson. Notes on Regression and Inheritance in the Case of Two Parents Proceedings of the Royal Society of London, 58, 240-242, 1895.

[75] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

[76] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[77] Luca Scrucca. GA: A package for genetic algorithms in R. *Journal of Statistical Software*, 53(4):1–37, 2013.

[78] Stefan Fritsch and Frauke Guenther. *neuralnet: Training of Neural Networks*, 2016. R package version 1.33.

[79] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[80] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

[81] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017. R package version 1.6-8.

[82] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[83] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[84] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[85] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. *xgboost: Extreme Gradient Boosting*, 2017. R package version 0.6.4.6.

[86] Danaja Maldeniya. 'What Big Data tells: Where is everybody at Avurudu?', 2016. [Online]. Available: https://lirneasia.net/2016/04/what-big-data-tells-where-is-everybody-at-avurudu/. [Accessed: 07-Jul-2019].

[87] Keshan De Silva and Yudhanjaya Wijeratne. 'Using Call Data Records to analyze event attendance', 2017. [Online]. Available: https://lirneasia.net/2017/11/using-call-data-records-analyze-event-attendance/. [Accessed: 07-Jul-2019].