

# SPATIO TEMPORAL FORECASTING OF DENGUE OUTBREAKS 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 OUTBREAKS 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

4-1	Mapping of BTS $b(i)$ ( $b_i$ ) to MOH $m(j)$ ( $m_j$ ) . . . . .	20
6-1	Visitation based probabilistic mobility for 4 MOH divisions . . . .	34
6-2	Exploration based probabilistic mobility for 4 MOH divisions . . .	35
6-3	Normalized probabilistic mobility - Week 32 . . . . .	35
6-4	Trip based outward mobility for 4 MOH divisions . . . . .	36
6-5	Trip based inward mobility for 4 MOH divisions . . . . .	37
6-6	Normalized log scaled trip mobility - Week 32 . . . . .	37
6-7	Mobility based total direct risk (log scale) for 4 MOH divisions . .	38
6-8	Mobility based total percent risk (log scale) for 4 MOH divisions .	38
6-9	Log scaled mobility based total risk - Week 32 . . . . .	39
6-10	Pearson's correlation between variables (without time-lagged data)	41
6-11	Correlation against dengue incidence using different methods . . .	44
6-12	Dengue Incidence - Predicted vs Actual for year 2014 - Colombo MC	45

## LIST OF TABLES

3.1	Structure of a Call Detail Record . . . . .	14
4.1	Risk score based on time band and location type . . . . .	23
5.1	Parameters for the genetic algorithm . . . . .	30
6.1	Highest correlation with dengue incidence for each input data source using multiple methods of correlation . . . . .	40
6.2	Model Performance for 20 MOH divisions (GA - Genetic Algorithms, NFC - Without feature classes, FC - With feature classes) . . . . .	42
6.3	Model Performance for Colombo-MC MOH division (GA - Genetic Algorithms, NFC - Without feature classes, FC - With feature classes) . . . . .	43
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) . . . . .	48

## 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 Hemorrhagic 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	Susceptible-Exposed-Infected
SEIR	Susceptible-Exposed-Infected-Recovered
SIR	Susceptible-Infected-Recovered
SVM	Support Vector Machines
SVR	Support Vector Regression
WHO	World Health Organization



# TABLE OF CONTENTS

<b>Declaration of the Candidate &amp; Supervisor</b>	<b>i</b>
<b>Acknowledgement</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>List of Figures</b>	<b>iv</b>
<b>List of Tables</b>	<b>v</b>
<b>List of Abbreviations</b>	<b>vi</b>
<b>Table of Contents</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem . . . . .	2
1.2 Proposed Solution . . . . .	3
1.3 Contributions . . . . .	3
1.4 Organization . . . . .	4
<b>2 Literature Survey</b>	<b>5</b>
2.1 Factors affecting propagation of dengue . . . . .	5
2.1.1 Weather related features . . . . .	5
2.1.2 Human Mobility . . . . .	7
2.2 Modeling Human Mobility . . . . .	7
2.3 Disease Outbreak Forecasting . . . . .	9
2.3.1 Mathematical Models . . . . .	9
2.3.2 Machine Learning Models . . . . .	10
2.4 Summary . . . . .	11
<b>3 Data collection and pre-processing</b>	<b>13</b>
3.1 Dengue Incidence Data and Information on MOH Divisions . . . . .	13

3.2	Pseudonymized Mobile Phone Call Detail Records . . . . .	14
3.3	Temperature and Rainfall Measurements . . . . .	14
3.4	Vegetation cover . . . . .	15
3.5	Imputation of missing values . . . . .	15
3.6	Summary . . . . .	16
<b>4</b>	<b>Human Mobility Models</b>	<b>17</b>
4.1	Identifying Home and Work Locations . . . . .	17
4.2	Probabilistic Mobility Model . . . . .	18
4.2.1	Formal Definition - Visitation Based Probabilistic Mobility	18
4.2.2	Formal Definition - Exploration Based Probabilistic Mobility	19
4.2.3	Mapping Probabilistic Mobility Indicators to MOH divisions	20
4.3	Trip Based Mobility Model . . . . .	21
4.3.1	Formal Definition - Trip Mobility . . . . .	21
4.3.2	Mapping Trip Mobility Indicators to MOH Divisions . . .	22
4.3.3	Normalized MOH based trip mobility . . . . .	22
4.4	Mobility based Risk Model . . . . .	22
4.4.1	Formal Definition - Mobility based Risk . . . . .	23
4.4.2	Mapping Risk Scores to MOH divisions . . . . .	24
4.5	Summary . . . . .	25
<b>5</b>	<b>Developing Forecasting Models</b>	<b>26</b>
5.1	Exploring the data set . . . . .	26
5.2	Correlation Analysis . . . . .	27
5.3	Input data for the model . . . . .	27
5.4	Measuring performance of the model . . . . .	28
5.5	Genetic Algorithm based optimization . . . . .	29
5.6	Machine Learning Methods . . . . .	30
5.6.1	Neural Networks . . . . .	30
5.6.2	Support Vector Machines . . . . .	30
5.6.3	Random Forests . . . . .	31

5.6.4	XGBoost . . . . .	31
5.7	Summary . . . . .	32
<b>6</b>	<b>Results</b>	<b>33</b>
6.1	Mobility Models . . . . .	33
6.2	Correlation Analysis . . . . .	39
6.3	Predictive Models . . . . .	41
6.4	Summary . . . . .	44
<b>7</b>	<b>Discussion</b>	<b>46</b>
7.1	Data collection and pre-processing . . . . .	46
7.2	Mobility Models . . . . .	47
7.3	Impact of mobility on predictive accuracy . . . . .	48
7.4	Comparison of machine learning methods . . . . .	49
7.5	Genetic algorithm based optimization . . . . .	49
7.6	Summary . . . . .	49
<b>8</b>	<b>Conclusion</b>	<b>51</b>
<b>9</b>	<b>Future Work</b>	<b>52</b>
	<b>References</b>	<b>53</b>