Mobile Network Big Data for Modeling Spread of Infectious Disease

LIRNE*asia* Big Data for Development Team

LIRNEasia

Pro-poor. Pro-market

This work was carried out with the aid of a grant from the International Development Research Centre, Canada and the Department for International Development UK

IDRC CRDI

Canada

UKaid

from the British people

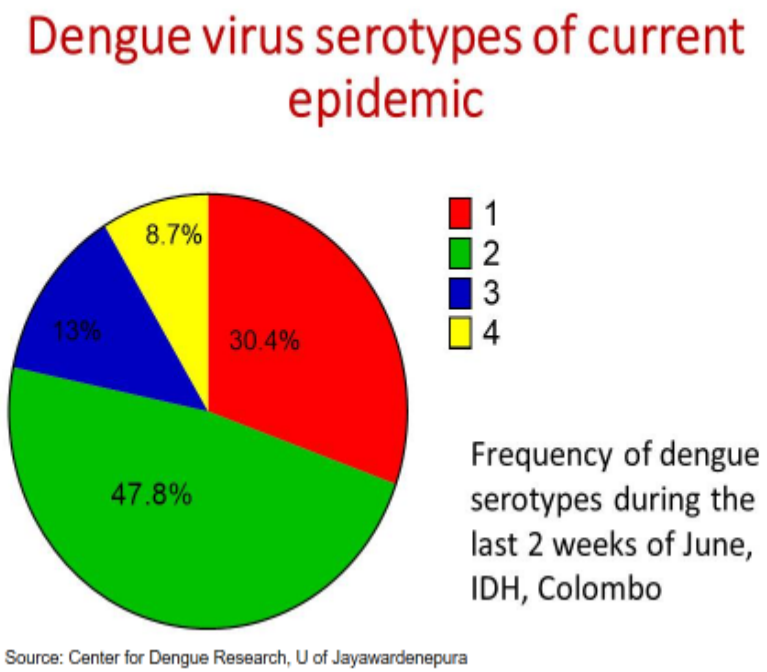
Dengue

- WHO estimates 50-100 million infections occur every year
- Endemic in over 100 countries
- A vector-borne disease caused by an RNA virus of family *Flaviviridae*; genus *Flavivirus* (Huhtamo, Eili et al.,2008)
- Main vectors (Monath, 1994)
  - Aedes aegypti
  - Aedes albopictus
- Four serotypes have been identified (DENV1-4)

Dengvaxia, a vaccine for dengue, developed by Sanofi Pasteur is currently undergoing clinical trials in multiple countries (Sirisena & Noordeen, 2016)

Dengue in Sri Lanka

- All 4 serotypes have been observed in Sri Lanka
- 42,194 cases reported in 2016 so far; likely to exceeed the 44,000 reported in 2012
  - 51.19% reported from Western Province
  - Increase of DENV-2 during June, 2016 outbreak



Role of human mobility in dengue propagation

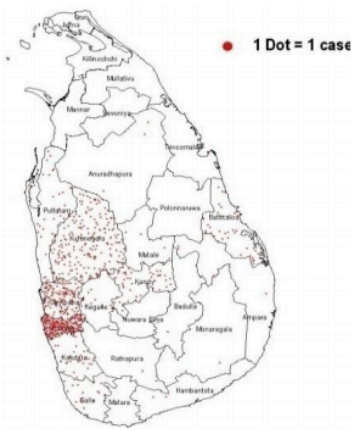
- The mosquito has a lifespan of 2-4 weeks and a range of around 100-800m (Muir & Kay 1998; Honório et al., 2003)
- Dengue is spread beyond the natural range of the mosquito by the movement of infected hosts

Knowledge of human mobility patterns can shed light on dengue propagation and the level of disease incidence in a region

Dengue in Jaffna

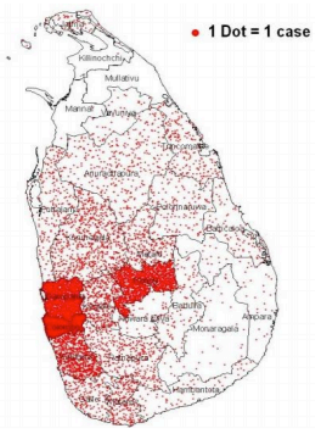
Ceasefire agreement may have contributed to dengue outbreaks in Jaffna after 2002

1996



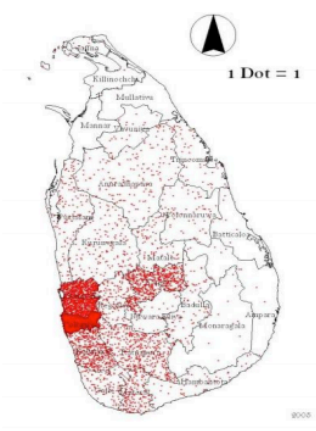
1294

2004



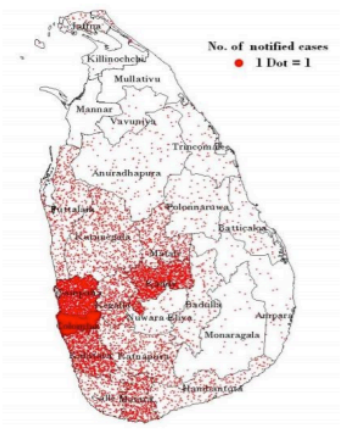
15463

2005



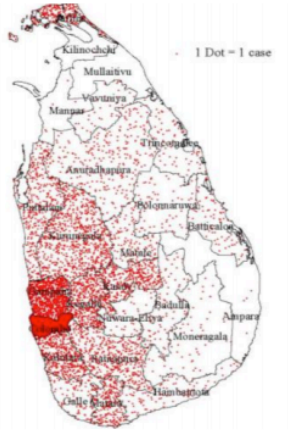
5994

2006



11980

2007



7327

Propagation of dengue in Sri Lanka over the years (Image Source: Epidemiology Unit, M. of H. (2009). CURRENT SITUATION & EPIDEMIOLOGY OF DENGUE IN SRI LANKA. In *Media Seminar for Dengue Week*. Retrieved from http://www.healthedu.gov.lk/web/images/pdf/msp/current_situation_epidemiology_of_dengue.pdf)

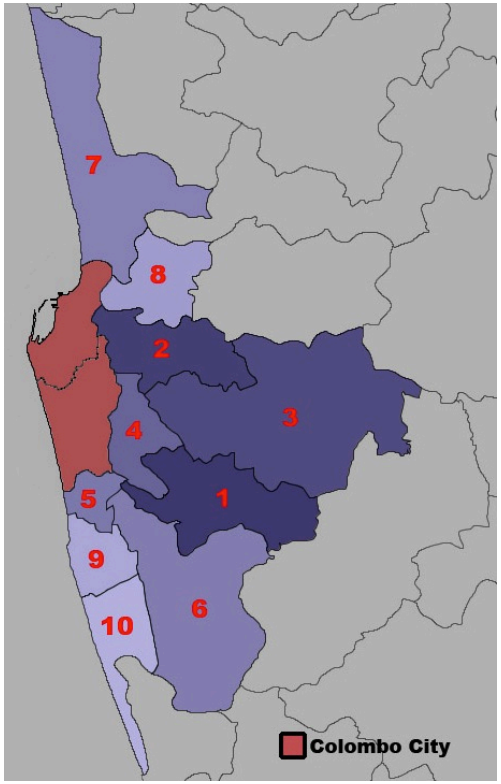
LIRNEasia
Pro-poor. Pro-market.

Mobile Network Big Data (MNBD) used in the research

- Multiple mobile operators in Sri Lanka have provided four different types of meta-data
  - Call Detail Records (CDRs)
    - Records of calls
    - SMS
    - Internet access
  - Airtime recharge records
- Data sets do not include any Personally Identifiable Information
  - All phone numbers are pseudonymized
  - LIRNEasia does not maintain any mappings of identifiers to original phone numbers
- Cover 50-60% of users; very high coverage in Western (where the capital city is located) & Northern (most affected by civil conflict) provinces, based on correlation with census data

Insights already used in urban planning in Colombo

46.9% of Colombo City's daytime population comes from the surrounding regions

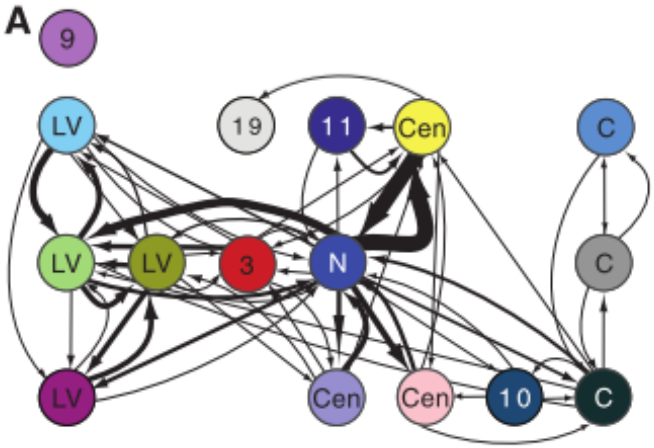


Home DSD	%age of Colombo's daytime population
Colombo city	53.1
1. Maharagama	3.7
2. Kolonnawa	3.5
3. Kaduwela	3.3
4. Sri Jayawardanapura Kotte	2.9
5. Dehiwala	2.6
6. Kesbewa	2.5
7. Wattala	2.5
8. Kelaniya	2.1
9. Ratmalana	2.0
10.Moratuwa	1.8

LIRNEasia
Pro-poor. Pro-market.

MNBD for modeling spread of disease

- Amy Wesolowski has pioneered use of MNBD for disease modeling
- Study on 2009 Malaria outbreak in Kenya (Wesolowski et. al, 2012) uses MNBD to identify regions with high Malaria outbreak risk
- Travel network is derived by calculating the average monthly trips by a subscriber to different regions



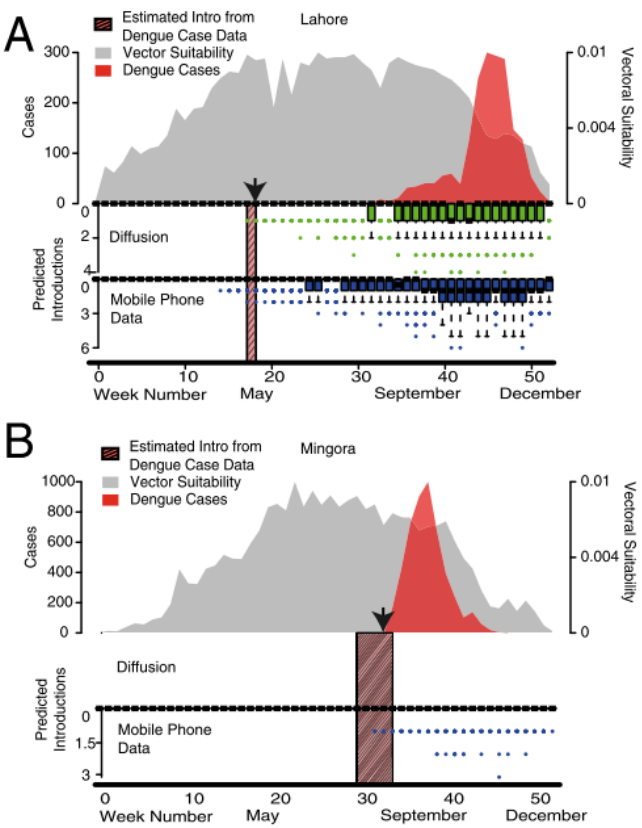
Travel network between nodes (Regions). Edges weighted by volume of traffic. Wesolowski et al. 2012

Using MNBD to predict Dengue Epidemics

- Model developed by Wesolowski et al. (2015) model on emergence of dengue epidemics in Pakistan corresponds to actual dengue propagation
- Existing ento-epidemiological model was used to predict the number of dengue cases within an administrative unit known as a tehsil
  - Model was fitted to the endemic region of Karachi starting at Day 1, to estimate daily infected number of people over time based on reported cases
    - Used estimate likelihood of dengue importation from Karachi to other regions on a given day
  - Model was fitted (in reverse) to naïve regions to estimate the time for introduction of dengue from outside solely with reported cases
    - Used to validate estimates of dengue introduction based on human mobility patterns
- Mobility models developed from CDR and gravity models were compared to determine which model was more accurate

Conclusions from Pakistan study

- CDR Mobility model can predict the time of introduction of dengue more accurately than the diffusion model
- CDR predicted the first infected cases coming to Lahore some time before the actual outbreak
  - This delay in the outbreak in Lahore is attributed to the immunology of the population in Lahore
  - There was a previous outbreak in 2011
- On the other hand, the outbreak occurred immediately after the introduction in Mingora
  - Mingora population is immunologically naive



The timeline of introduction of Dengue to A) Lahore, B) Mingora according to the Pakistan study (Wesolowski et. al, 2015)

LIRNEasia
Pro-poor. Pro-market.

Data needed to predict dengue outbreaks

- Mobility is one among many factors that affect disease propagation
  - Weather, seasonality
  - Vector population dynamics
  - Incubation periods of the disease
  - Vector control mechanisms
  - Asymptomatic transmission
  - Immunity of the population to different serotypes of dengue

How do we model these different parameters?

- Different approaches are possible
  - Machine learning techniques
  - Statistical modelling techniques
- Need to derive mathematical models with good assumptions
- Need to understand how different parameters behave to apply machine learning models
- Epidemiological and entomological expertise are essential

$$\lambda^{V\to h} = \frac{a\,\phi^{V\to h} I_V S_h}{N_h}$$

$$\lambda^{h\to V} = \frac{a\,\phi^{h\to V} I_h S_V}{N_h},$$

Estimating vector to human ($\lambda^{V\text{->}h}$) and human to vector ($\lambda^{h\text{->}V}$) dengue incidence rate requires knowledge on the populations (I_h,S_h,N_h,I_v,S_v,N_h) and biting rate (a) and transmission probabilities ($\Phi^{V\text{->}h}$,$\Phi^{h\text{->}V}$) (Wesolowski et. al., 2015)

LIRNEasia
Pro-poor. Pro-market.

Sri Lanka: Multi-disciplinary, multi-stakeholder research effort

- Epidemiology Unit, Ministry of Health
  - Case data, ground truth data
  - LIRNEasia & U of Moratuwa lack expertise on epidemiology and entomology needed to understand the disease dynamics
  - Key collaborators: Dr. Hasitha A. Tissera, Dr. Azhar Ghouse

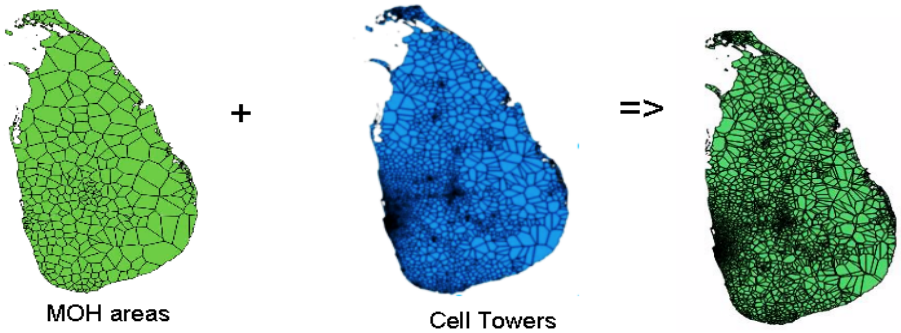
- University of Moratuwa
  - Provides expertise on
    - Statistics & Machine learning
    - Computational modeling
  - Key collaborators : Dr. Shehan Perera

Work done so far

- Data on weekly new dengue cases obtained and preprocessed for 2013 and 2014
- Data from Sri Lanka's weather stations (total 414) being analyzed
  - Obtained rainfall data from 112 stations for 2013
    - 70 stations had data for the entire year
    - 24 stations had missing data for less than 2 months (filled by imputation)
    - 18 stations had missing data for more than 2 months (discarded)
  - Temperature data from 23 stations
    - One station was missing 6 months of data (discarded)
    - Other missing values were imputed
- Spatial boundaries of cell towers, MOH divisions, and GN divisions differed
  - We could not map population to MOH division or mobility to MOH division because of these different boundaries
  - An undergraduate group from University of Moratuwa was able to solve the issue of overlapping spatial boundaries using Voronoi tessellation and developed their own algorithm to derive mobility

Student work in detail

- Voronoi tessellation was employed to estimate cell tower coverage and MOH divisional boundaries
- The fraction of coverage overlapping an MOH division was used to obtain fractional values for mobility from each cell coverage



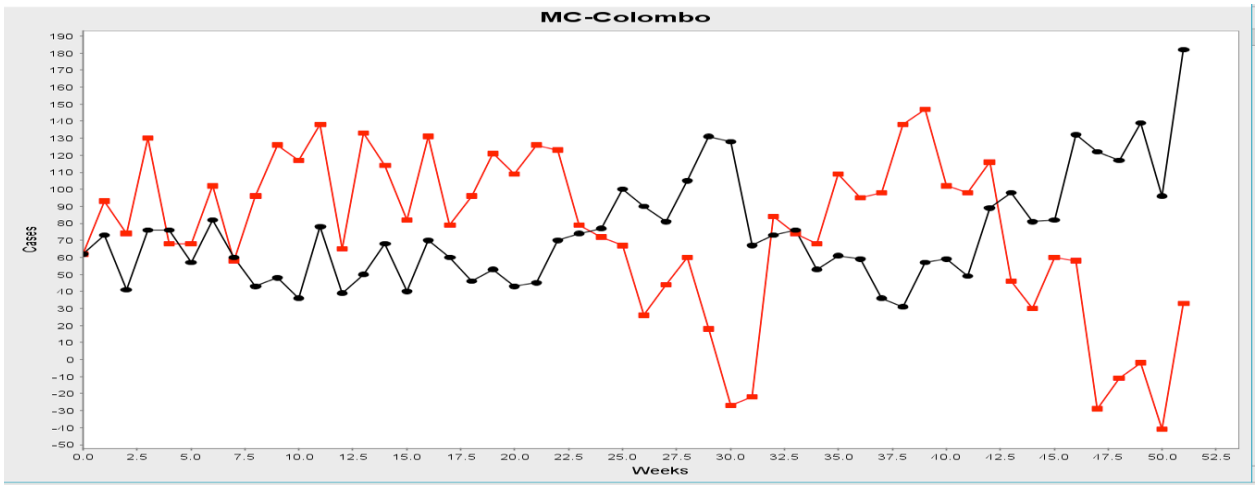
$$f_{ij} = \frac{\text{Intersection of cell tower j for MOH i}}{\text{Complete area covered by cell tower j}}$$

f_{ij} - fractional coverage of a cell tower j on MOH area i

- Students were later able to obtain the MOH boundaries and the populations with assistance from the Epidemiology Unit
- Multiple statistical models as well as different machine learning techniques were tried out on the data

Meta-population model for Colombo City

- A statistical model proposed for dengue by Sarzynska, Udiani & Zhang, 2013 in a study done for Peru was adapted for Sri Lanka, but using parameter values from Peru
- Results were not encouraging (RMSE-70.29)



- We are working on estimating parameters for Sri Lanka using machine learning techniques

LIRNEasia
Pro-poor. Pro-market.

Machine learning: From 1 to many MOH divisions

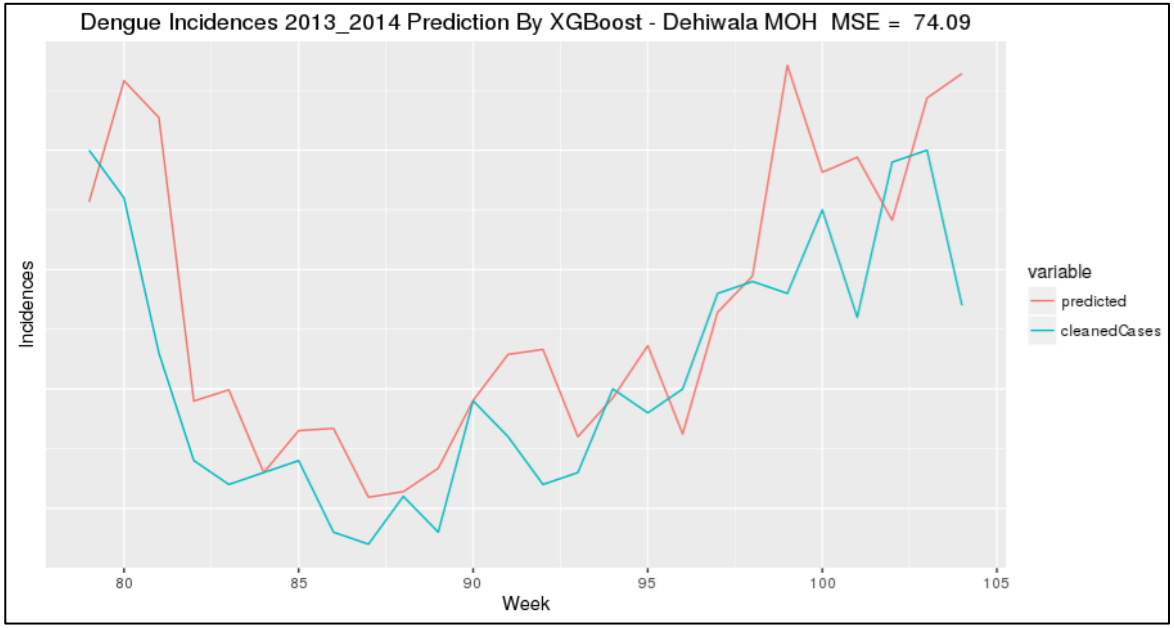
- For a single MOH division, only 52 data points for a year
  - Only 70% (36 data points) left for training
- Due to small training set, models trained are overfitted and perform badly when predicting unseen data
- Training for a single MOH division was done to identify suitability of different techniques
  - Final model will incorporate spatial aspects and predict for the entire country
- Since training for a single MOH was not yielding good results, 6 similar MOH areas were considered for training
  - Provides more data points for training
  - Provides an idea on model behaviour under different conditions

Multiple approaches

- Random Forest method used in Pakistan (Rehman et al., 2016) tried in Dehiwala MOH Division
  - Low error, but does not predict trend yet
- Neural networks have been tried in endemic countries such as Singapore (Aburas et al., 2010); was tried for Kotte MOH Division
  - Sri Lanka lacks data for a long enough period; nine years of data in SE Asia cases

XGBoost for Dehiwala MOH

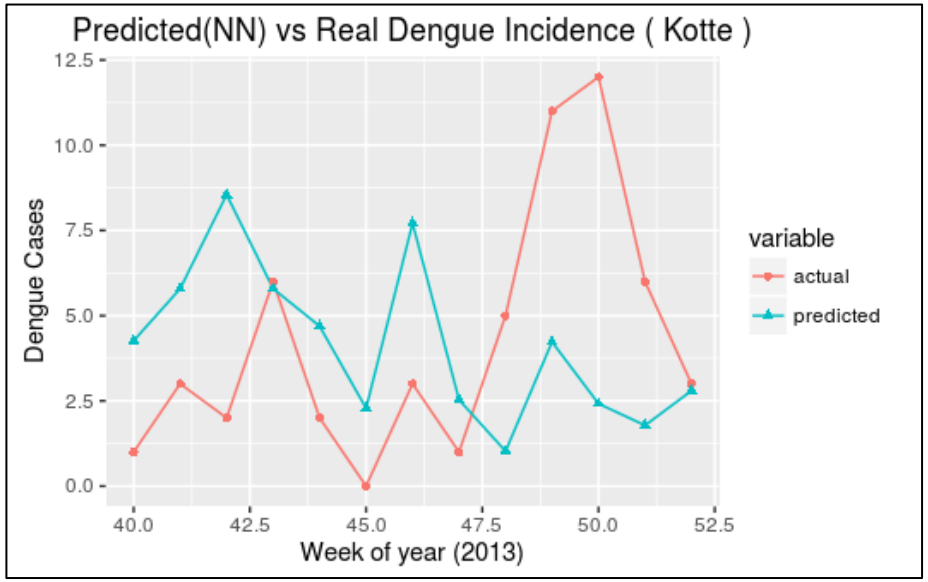
- XGBoost is short for "Extreme Gradient Boosting". Can be considered as an ensemble decision tree algorithm
- RMSE of 8.61 after cleaning the data
- Temperature, weather data extrapolated for 2014 due to unavailability of real data
- Top features
  - temperature lag for 2 weeks
  - cases lag for 2 weeks
  - temperature lagged for 1 week
- Our predictions are getting closer to the actual curve



LIRNEasia
Pro-poor. Pro-market.

Does Mobility improve our model?

- Applied the naive mobility model to neural networks
- Small improvement in RMSE value from 9.62 to 8.75
- Improvement in the prediction curve
- Word of caution: These are preliminary results; In some iterations we did not see significant improvements
- However, high correlation with mobility was consistent



Observed higher correlation for mobility than rainfall during our preliminary analysis

- Analysed with 15 variables
- Correlation of rainfall : 0.10497
- Correlation with mobility: 0.14514

LIRNEasia
Pro-poor. Pro-market.

What do the preliminary results say?

- There is not much difference in RMSE values between the machine learning techniques attempted so far
  - Not enough of a difference to discard any technique outright
- Even though RMSE value is good, models without mobility do not predict the peaks, troughs and the general trend of dengue incidence
- Mobility has a definite correlation with dengue incidence and improved RMSE as well as the prediction curve
  - Room for improvement in two aspects when considering mobility
    - Estimating the time spent by the travelling population in an MOH area
    - Methodology of fusing mobility to our models
- We haven't figured out the finer details of incorporating spatial properties to the model, in which case, introducing multiple MOH divisions might actually increase the error
- Quality of weather data, presence of missing values can also affect model accuracy

Generalized disease propagation prediction model

- If our model can incorporate the essential factors (including human mobility) that affect spatio temporal disease propagation and accurately predict for dengue, we can generalize it for other diseases
- Can use the same model for Zika and predict an outbreak
  - Can execute vector control mechanisms, awareness campaigns pre-emptively
- Mobility will play an even larger role in non-vector borne diseases
  - If we can come up with an accurate model for vector borne diseases, modelling other types of infectious diseases will be easier

References

1. Huhtamo, Eili et al. “Molecular Epidemiology of Dengue Virus Strains from Finnish Travelers.” *Emerging Infectious Diseases* 14.1 (2008): 80–83. *PMC*. Web. 10 Oct. 2016.
2. Monath, T. P. (1994). Dengue: the risk to developed and developing countries. *Proceedings of the National Academy of Sciences*, *91*(7), 2395-2400.
3. Centre for Dengue Research, U. of S. J. (2016). Dengue Virus Serotypes of Current Epidemic. Retrieved October 6, 2016, from https://www.facebook.com/206869486156896/photos/a.206973839479794.1073741828.206869486156896/598635470313627/?type=3&theater
4. Brockmann, D. (2010). Human Mobility and Spatial Disease Dynamics. *Reviews of Nonlinear Dynamics and Complexity*, *2*, 1–24. http://doi.org/10.1002/9783527628001.ch1
5. Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, *439*(7075), 462–465. http://doi.org/10.1038/nature04292
6. Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science (New York, N.Y.)*, *338*(6104), 267–70. http://doi.org/10.1126/science.1223467
7. Wesolowski, A., Buckee, C. O., Bengtsson, L., Wetter, E., Lu, X., & Tatem, A. J. (2014). Commentary: Containing the Ebola Outbreak – the Potential and Challenge of Mobile Network Data. *PLoS Currents Outbreaks*, 1–17. http://doi.org/10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e.Funding
8. Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., … Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*, *112*(38), 11887–11892. http://doi.org/10.1073/pnas.1504964112
9. Sarzynska, M., Udiani, O., & Zhang, N. (2013). A study of gravity-linked metapopulation models for the spatial spread of dengue fever. *arXiv Preprint arXiv:1308.4589*, *2008*, 1–32. Retrieved from http://arxiv.org/abs/1308.4589
10. Rehman, N. A., Kalyanaraman, S., Ahmad, T., Pervaiz, F., Saif, U., & Subramanian, L. (2016). Fine-grained dengue forecasting using telephone triage services. *Science Advances*, *2*(7), 1–10. http://doi.org/10.1126/sciadv.1501215

References

11. Aburas, H. M., Cetiner, B. G., & Sari, M. (2010). Dengue confirmed-cases prediction: A neural network model. *Expert Systems with Applications*, *37*(6), 4256–4260. http://doi.org/10.1016/j.eswa.2009.11.077
12. Husin, N. A., Salim, N., & Ahmad, A. R. (2008). Modeling of dengue outbreak prediction in Malaysia: A comparison of neural network and nonlinear regression model. *Proceedings - International Symposium on Information Technology 2008, ITSim*, *4*, 6–9. http://doi.org/10.1109/ITSIM.2008.4632022
13. Rachata, N., Charoenkwan, P., Yooyativong, T., Chamnongthai, K., Lursinsap, C., & Higuchi, K. (2008). Automatic prediction system of dengue haemorrhagic-fever outbreak risk by using entropy and artificial neural network. *2008 International Symposium on Communications and Information Technologies, ISCIT 2008*, (Iscit), 210–214. http://doi.org/10.1109/ISCIT.2008.4700184
14. WHO Scientific Group on Arthropod-Borne and Rodent-Borne Viral Diseases. (1985). Arthropod-borne and rodent-borne viral diseases : report of a WHO scientific group [meeting held in Geneva from 28 February to 4 March 1983].
15. Vitarana, T., Jayakuru, W., & Withane, N. (1997). Historical account of Dengue Haemorrhagic Fever in Sri Lanka.
16. Yang, H. M., Macoris, M. L. G., Galvani, K. C., Andrighetti, M. T. M., & Wanderley, D. M. V. (2009). Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue. *Epidemiology and Infection*, *137*(8), 1188–1202. http://doi.org/10.1017/S0950268809002040
17. P. H. D. Kusumawathie, & R. R. M. L. R. Siyambalagoda. (2005). Distribution and breeding sites of potential dengue vectors in Kandy and Nuwara Eliya districts of Sri Lanka. *The Ceyion Journal of Medical Science* .