Improving Disease Outbreak Forecasting Models for efficient targeting of Public Health Resources

CPR*south* 2017
Yangon, Myanmar
Sep 01st, 2017

Lasantha Fernando, Sriganesh Lokanathan, Shehan Perera, Azhar Ghouse, Hasitha Tissera

LIRNEasia
Pro-poor. Pro-market

IDRC CRDI
Canada

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.

UKaid
from the British people

A multi-disciplinary, multi-stakeholder collaborative effort

- Tripartite collaboration with Epidemiology Unit of the Ministry of Health, University of Moratuwa, and LIRNEasia
  - Epidemiology Unit provides expertise on epidemiology/entomology as well as case data (Key Collaborators: Dr. Hasitha Tissera, Dr. Azhar Ghouse)
  - University of Moratuwa provides expertise on computational modeling (Key Collaborator: Dr. Shehan Perera)
  - LIRNEasia provides CDR data as well as research expertise
- Research is funded by IDRC, Canada and the Senate Research Committee of University of Moratuwa

Dengue - A global menace & the rising trend in Sri Lanka

- WHO estimates 50-100 million infections globally every year
  - Endemic in over 100 countries (including Sri Lanka)
- Main vectors: Aedes aegypti and Aedes albopictus (Monath, 1994)
  - Aedes mosquitoes have a limited spatial range (Muir & Kay, 1998)
  - Human mobility plays a critical role in introducing dengue across regions (Stoddard et. al, 2009; Wesolowski et. al 2015)
- 2017 saw the worst ever dengue epidemic in Sri Lanka
  - 125,387 cases reported up to August
  - Over 200 deaths
  - ~ 29k cases in 2015
  - ~ 55k in 2016, a record at the time
  - Official reported statistics from Epidemiology Unit - Ministry of Health (2017)

- Need better predictions for efficient prevention & control

What can we do with better predictions?

- Developing countries have limited resources to effectively prevent or control an outbreak
  - With our predictive models, we can predict where the next outbreak will most likely occur

Limited Public Health Sector Resources: In Sri Lanka, during an outbreak, security forces are called in to help with dengue prevention and control due to resource shortages - Source: http://www.army.lk/files_eng/Centraldengue_1.jpg

LIRNEasia
Pro-poor. Pro-market.

What are the policy questions we are trying to answer?

- Can we use Call Detail Records (CDR) to derive human mobility models & apply for disease outbreak predictions?
- Can disease outbreak forecasts be used for
    a. allocating public health sector resources efficiently?
    b. formulating epidemic disease policy?
- Can we extend this work to predict other infectious disease outbreaks as well?
    - The severe dengue epidemic in 2017 would not have benefited much from forecasting beforehand
    - What if we get hit by a different infectious disease like Zika or Chikungunya?

Methodology: Forecasting models using big data & machine learning

- Different human mobility models derived from CDRs
  - Tried a probabilistic model, a trip based model & a risk based model
  - We wanted to know which mobility model predicts best
- Evaluated multiple machine learning methods as well
  - Literature did not point towards a conclusive single technique
  - Evaluated Support Vector Regression, Neural Networks, XGBoost and Random Forests
  - Predict dengue incidence 2 weeks ahead
  - Lot of feature engineering and tuning in between data collection & prediction

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

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

- All models were run with/without mobility as an input
- Used evolutionary algorithms to improve feature selection
- RMSE and R^2 to measure model performance

How can we use big data to infer human movement patterns?

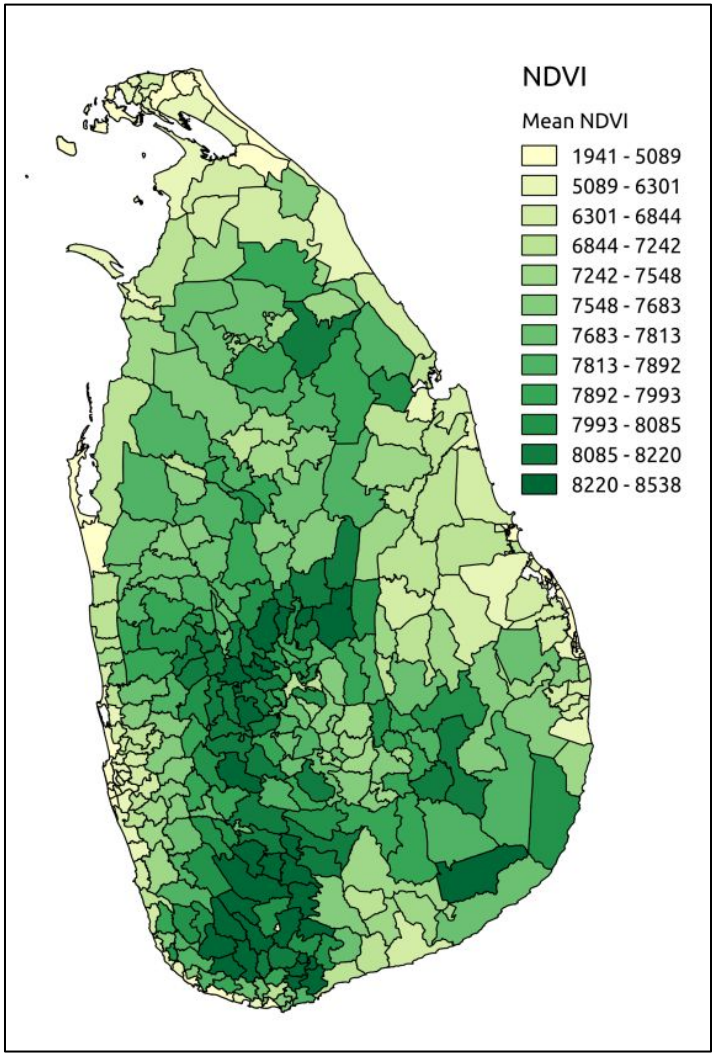
- Mobile Network Big Data (MNBD) can provide detailed information on human mobility patterns
- Structure of a Call Detail Record

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

  - Records of all calls made and received by a person created mainly for the purposes of billing
  - The Cell ID in turn has a lat-lon position associated with it
- We used CDR data for more than one year in 2012-2014
  - Covers under 10 million SIMs
  - Nearly 1.5 billion records

Other data sources for the forecasting models

- Weekly dengue cases for a Medical Officer of Health (MOH) division (2012 to 2014)
- Temperature & rainfall data (22 stations)
  - From NOAA Integrated Surface Data (ISD)
  - Projected to weekly average estimate for an MOH
- Mean Normalized Difference Vegetation index (NDVI)
  - Using MODIS satellite data from NASA
  - Done by a colleague at U. of Moratuwa



Right: Mean Vegetation Index for given MOH

LIRNEasia
Pro-poor. Pro-market.

CDR based human mobility models improve predictive performance

- Verified correlation of each input vs dengue incidence
  - Mobility has very high correlation - Second most highly correlated after past dengue cases
- In preliminary evaluations, mobility improved model performance consistently, even if marginal in some cases

Model (Before GA Optimization)	R^2		RMSE	
	Without Mobility	With Mobility	Without Mobility	With Mobility
Random Forests	0.628	0.639	6.907	6.812
Neural Networks	0.063	0.335	10.966	9.239
XGBoost	0.630	0.640	6.892	6.794
Support Vector Regression (SVR)	0.680	0.704	6.408	6.170

Prediction curve matches trend, good prediction accuracy

Prediction for MC-Colombo for year 2014

Machine Learning Technique	Overall RMSE	Overall R^2
Random Forests	8.258	0.926
Neural Networks	12.154	0.839
XGBoost	7.852	0.933
SVR	8.618	0.919

Final predictions done for 20 MOH divisions

- Used genetic algorithms to improve results further
- XGBoost showed best results

LIRNEasia
Pro-poor. Pro-market.

Can repurpose our models with minimum modifications

- Dengue case data is used as a response variable for these models, while past case data is provided as an input for the model
- For vector-borne infectious diseases like zika or chikungunya, we simply need to replace dengue data with other disease data and retrain with minimal effort
  - Zika, chikungunya are transmitted by the same aedes mosquitoes with very similar characteristics
- For other infectious diseases, we would have to modify the some aspects of the methodology
  - Different time lags for input features due to different incubation periods of the diseases
  - Risk scores assigned in the mobility model would change

Policy findings & recommendations

- CDR is a rich source of data to model human mobility for disease outbreak prediction
  - CDR might have issues of representativity compared to one time surveys, but still highly useful (Consider high resolution photo vs. low resolution video)
  - Even in regions where the disease is endemic, human mobility is critical for dengue propagation
  - Mobility models should be consumed by the Ministry of Health for formulating public policy on infectious diseases
- Use the disease outbreak forecasts to
  - allocate public health sector resources efficiently
  - formulate epidemic disease policies

Policy findings & recommendations: Cntd.

- Repurpose these models to predict other infectious disease outbreaks
  - Sri Lanka is an island which has a single point of entry for most international travellers
  - Easier to track and predict outbreaks if an entirely new infectious disease is introduced to the country
- Negotiate data access from mobile operators with the assistance of government organisations to establish a sustainable model to continuously predict outbreaks

Next steps: Risk maps & predictive classification models

- Our work focuses on regression models that attempt to predict the exact number of dengue cases
- But In order to improve public service delivery, we need risk maps
- To generate risk maps, we need classification models
- Next step: Identify risk bands and give our predictions as a risk classification for an MOH division
  - Simply need to retrain the machine learning models to do classification instead of regression
  - With risk classification, our models should be able to classify with higher confidence
  - Easier to visualize and communicate
  - Easier for public health sector officials to act upon such an output

References

1. Epidemiology Unit - Ministry of Health. (2017). Distribution of Notification(H399) Dengue Cases by Month. Retrieved from http://epid.gov.lk/web/index.php?option=com_casesanddeaths&Itemid=448&lang=en
2. Monath, T. P. (1994). Dengue: The Risk to Developed and Developing Countries. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(7), 2395–2400.
3. Muir, L. E., & Kay, B. H. (1998). 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.
4. Stoddard, S. T., Morrison, A. C., Vazquez-Prokopec, G. M., Soldan, V. P., Kochel, T. J., Kitron, U., ... & Scott, T. W. (2009). The role of human movement in the transmission of vector-borne pathogens. PLoS neglected tropical diseases, 3(7), e481.
5. 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.

Thank you