Mining Malaria Health Data in Mozambique
From Bayesian Incidence Risk to Incidence Case Predictions

Orlando Zacarias
DSV

2013-Nov-12
1 Motivation

2 Research Objectives

3 Published Articles
   - Paper 1: Mapping Malaria incidence that accounts for environmental factors in Maputo province - Mozambique (malaria journal)
   - Paper 2: Spatial and temporal patterns of malaria incidence in Mozambique (malaria journal)
   - Paper 3: Comparison of infant malaria incidence in districts of Maputo province, Mozambique (malaria journal)
   - Paper 4: Predicting the Incidence of Malaria Cases in Mozambique Using Regression Trees and Forests (comp. eng. journal)
   - Paper 5: Strengthening the Health Information System in Mozambique through Malaria Incidence Prediction (IST Africa + IEEE)
   - Paper 7: Generalization of Malaria Incidence Prediction Models by Correcting Sample Selection Bias (to appear Springer)

4 Work At Near Future
Motivation I

- Malaria: **vector-born disease**, caused by a female mosquito

- Millions of people affected in the World

- among deadliest diseases in Mozambique major public health problem
  - 48% all hospitalized patients
  - 63% infants hospitalizations
  - 26.7% of country mortality (INE-Nat. Inst. Statistics 2007)

- Health Information System (HIS): Collects and stores disease data and some interventions
  - Provide statistics
Motivation II

- No information and knowledge:
  - Spatial and temporal RR estimates and mapping (where and when occurs?)
  - Estimates of possible disease predictor factors (how each contributes?)
  - Estimates of future disease cases (how much, where and when will occur?)

- Efforts required:
  - Improve disease management
  - Increment measures of prevention and vector control
  - Increase awareness
  - Accurate prediction of malaria incidence cases at district and national level
Research Objective I

- Design/develop models (patterns) of data extraction to improve decision-making in health sector (national malaria program)
  - Provision of malaria incidence relative risk estimates (space-time)
  - Determining important predictors and their relative contribution
  - Provide estimates of future malaria cases
    - give indications for better planning and disease control
    - contribute to disease reduction
Data: malaria cases and climatic factors (temperature and rainfall) 2001-2002

Bayesian spatial seasonal random effects model (no time dependency)

- Extended to include environmental covariates temperature and rainfall
- Seasonality: rainy (summer: Oct-March) and dry (winter: April-Sept)

Deviance Information Criteria (DIC) used for model comparison:

- **2001 best fitted model**: summer with temp only & winter all equal value
2002 best fitted model: with both predictors in winter & summer simplest model (spatial no covariates)

Causal relationship:

- **Summer**: both years rainfall not associated to malaria incidence in models with this covariate only

- **temperature** associated to malaria incidence in all models

- Winter 2001: **temperature** not associated to malaria incidence in model with both covariates

- Winter 2002: **Rainfall** not associated in model with both covariates
Mapping Malaria Incidence that accounts for environmental factors in Maputo province-Maputo: Paper 1 III

- Spatial pattern of malaria incidence risk in 2001:
  - Highest incidence rate ranging from 4.6 to 5.9 cases per 100 population year in Magude and Moamba

- Spatial pattern of RR in both seasons of 2002 is over the rate of 9.5 in district of Matola

- RR Very high compared to 2001
Spatial and temporal patterns of malaria incidence in Mozambique: Paper 2 I

- Data: malaria cases and climatic factors (temperature, rainfall & humidity) for 1999-2008

- Modeling: Malaria cases that evolve in space and time
  - Missing data multiple imputation employed to climatic data
  - Spatial misalignment data problem
  - Predictor variables: humidity, temperature (minimal and maximal) and rainfall
  -Introduced temperature variation
  -Humidity, Minimal and Maximum Temperature converted to categorical variables due to non-linear relationship to malaria cases

- CAR-model chooses from two different weighting matrices: binary and border length
Three different models compared for their fitness on available data: Full model with binary adjacency fitted best

No association to malaria incidence risk:
- rainfall
- minimum temperature in range 11 to 16°C
- maximum temperature for range 24.5 to 27°C

Associated to malaria incidence risk:
- minimum temperature between 17 and 21.1°C
- maximum temperature ranging 28 to 35°C
- relative humidity of 54.4% to 83%
**Trends of incidence** in districts of Matutuine and Namaacha are lower as compared to other districts except for August where it shows sharp increase in Namaacha.

**Smoothed maps** of monthly averaged malaria incidence risk allow visualization of initial and peak months of transmission.
Comparison of infants malaria incidence in districts of Maputo province, Mozambique: Paper 3 I

  - Use of explanatory variables rainfall, temperature and humidity
  - Four Bayesian models were fitted

- **Temporal and space-time effects** (no spatial random effect) determined as to best fit the data

- **Malaria incidence risk** found to be higher in districts of Manhiça, Matola and Magude
Rainfall and humidity were significant predictors of malaria incidence risk

- 1% increase on humidity levels lead to 5% increase on malaria incidence risk
- Additionally 1mm of rain lead to same percentage unity (1%) increase of incidence risk
- Humidity played major role in malaria transmission and incidence risk for this demographic group and study period

Temperature covariates determined no association to malaria incidence in the region

Significant variation in space-time pattern of incidence in the region for months January-April
Data: 1999-2008 malaria (all ages), climatic factors and indoor-residual spraying

Method: **Decision Trees** and **Random Forests (RFs)**
- 960 instances generated
- Training set: 864 examples (nine years: 1999-2007)
- Test set: 96 instances (one year: 2008)
- Parameter tuned for RFs: # splitting attributes and # trees to grow
- Nine models generated from different time-windows trained and compared (mean squared error)
Predictive performance

- Trees: one-year time window
- RFs: two-years time window (best model, i.e., fewer errors committed)
- Results were statistically significant

Predictors relative importance ranking

- Indoor-residual spraying
- Climatic factors: min. temperature and rainfall
- Regional (districts or spatial): Manhiça, Matola and Marracuene
- Months (time): January, July, June, September, August and October
Data: 2007-2009 malaria (infants, < 5 years), climatic factors and indoor-residual spraying

Method: **Support Vector Machines (SVMs)**
- 288 instances generated
- Training: 192 examples (two years: 2007-2008)
- Test set: 96 instances (one year: 2009)
- Grid optimization tuning parameters of four kernels: *linear, polynomial, radial basis function and sigmoid*
- Radial Basis Function (RBF) kernel selected
Relative importance of predictors: **RFs approach, value each feature randomly permuted & measure its effect on predictive performance using test set**

**Importance Ranking**
- temperature variation with highest predictive power
- Regional (districts or spatial): Matutuine, Namaacha, Moamba and Marracuene
- Months (time): April, February, September and August
Comparing Predictive Strategies & Models - Paper 6 I

- Data: 1999-2008 malaria (all ages), climatic factors and indoor-residual spraying

- Method: **Support Vector Machines (SVMs) and Random Forests (RFs)**
  - 960 instances generated
  - Training set: 864 examples (nine years: 1999-2007)
  - Test set: 96 instances (one year: 2008)
  - RFs and DT models developed in paper 4
  - Missing values 2 strategies: mean (svm1) and multiple imputation (svm2)
  - Grid optimization tuning parameters of four kernels and RBF kernel chosen
  - Nine models generated from different time-windows trained and compared (mean squared error)
Comparing Predictive Strategies & Models - Paper 6 II

- Predictive performance
  - SVMs models over performed RFs and baseline DTs
  - Most effective, svm1 for 2 years time-windows
  - Results were statistically significant

- Relative importance of predictors: RFs approach

- Predictors relative importance ranking
  - Indoor-residual spraying (both svm1 and RFs)
  - Climatic factors: humidity, min. temperature and temperature variation
  - Regional (districts or spatial): Manhiça and Matola (both svm1 and RFs)
  - Months for svm1 (time): February, March, May and June
Complementary data: 2009-2012 malaria (all ages), climatic factors and indoor-residual spraying

Test SVMs Infants and All Ages Models

Two provinces: Zambézia and Cabo Delgado

Infants dataset Zambézia. All ages both provinces

**Sample selection bias problem:** training and test set drawn from different iid - Nonrandom data sampling problem

Comparing performance of models after **correcting bias** with standard models
Results:

- Performance of infant model increases in all districts when applying **correcting bias** compared to direct use of standard model.

- Decreased performance for all ages after **correcting bias** in some districts, while slightly increases in others (less than 4%).

Application of sample selection approach improves performance of malaria incidence prediction but, **this cannot be completely guaranteed**.
Work at Near Future

- Drafting and correcting thesis
- Pre-Doc seminar
- Redesign predictive models aiming at reducing prediction time to weeks or months (requires more involvement from Ministry of Health-MoH)
  - draft an article and propose a tool to MoH (not to include in thesis)
THANK YOU!
AND or OCH
TACKLE SAMYCKET!