Predicting the Incidence of Malaria Cases in Mozambique Using Regression Trees and Forests

Orlando P. Zacarias and Henrik Boström

***Abstract*—**Malaria remains a significant public health concern in Mozambique with disease cases reported in almost every province. This study investigates the prediction models of the number of malaria cases in districts of Maputo province. Used data include administrative districts, malaria cases, indoor residual spray and climatic variables temperature, rainfall and humidity. Regression trees and random forest models were developed using the statistical tool R, and applied to predict the number of malaria cases during one year, based on observations from preceding years. Models were compared with respect to the mean squared error (MSE) and correlation coefficient. Indoor Residual Spray (IRS), month of January, minimal temperature and rainfall variables were found to be the most important factors when predicting the number of malaria cases, with some districts showing high malaria incidence. Additionally, by reducing the time window for what historical data to take into account, predictive performance can be increased substantially.

***Keywords*—**Malaria, Regression trees, Regression forests.

I. Introduction

MALARIA has long been the most deadly disease in Mozambique and other tropical countries, especially for those in *South of the Sahara*. This situation has caused this disease to turn into a global health problem. The disease is usually transmitted by the bite of female anopheles mosquito - malaria vector [1]. Surface water pools are the best breading locations of the vector. Malaria accounts for 48% of all hospital visits and 63% of all pediatric hospitalizations in rural and general hospitals, and is responsible for approximately 26.7% of the total mortality in the country. Plasmodium falciparum is responsible for about 90% of all malarial infections and also the species associated with more severe cases of the disease [1]. Malaria is also the leading cause of death among children admitted to pediatric services in Mozambique. Having a population of approximately 21.5 million inhabitants, most people can be considered at risk of contracting malaria, especially for children younger than 5 years of age and pregnant women [1]. Taking into consideration that many people with malaria do not seek the formal health services, the reported number of cases are hence much lower than the true number.

In Mozambique, malaria is endemic, with some peaks in the rainy season (especially in December to March, when most cases occur). Under favorable conditions for transmission, malaria may well become epidemic with increasing number of cases and deaths. The epidemic is usually triggered by factors such as regular rainfall and high temperatures, which increase the breeding sites of mosquitoes that cause malaria. Natural disasters such as floods, droughts and cyclones have in recent years contributed to the increase of malaria transmission, particularly in coastal low-lying areas and along major country rivers. Until now there are unfortunately still no vaccines available for malaria. Hence, efforts are required to implement measures for prevention and control activities.

In [2], a model was proposed for predicting the malaria incidence in Burundi in an area of unstable transmission, studying the association between disease dynamics and environmental variables. That model used time series of quarterly notified malaria cases, normalized vegetation index, temperature and rain records. By employing an autoregressive integrated moving average methodology; they obtained a model capturing the relationship between the malaria cases and environmental variables. It was shown in [3] that adding rainfall as a covariate to autoregressive integrated moving average (ARIMA) models, the prediction accuracy improved slightly in some districts of Sri Lanka but worsened in others.

Data mining techniques have been applied in the analysis of Plasmodium falciparum malaria parasites to define organellar function [4]. A computational approach for mining malaria transcriptome was highlighted in a study [5], while yet another application for text mining of Chinese traditional medicine was employed in [6]. Work in [7] implements algorithms for mining large scale patient records to discover interesting relationships in malaria related cases, which may assist in crucial decision making and new policy formulation. Some of these studies perform prediction of malaria incidence while employing ARIMA models with very few climatic explanatory variables. Research on a similar mosquito caused (vector-borne) disease is found in [8]-[10], where the prediction of dengue hemorrhagic fever (DHF) in Malaysia and Thailand was studied, applying different architectures such as Artificial Neural Network, Nonlinear Regression and

Orlando P. Zacarias is with the Department of Computer and Systems Sciences, Stockholm University, Forum 100 SE-164 40 Kista, Sweden and Department of Mathematics and Informatics, Eduardo Mondlane University, Main Campus, POBox 250, Maputo, Mozambique (phone: +46-8-16 49 93 ; fax: +46-8-703 90 25 ; e-mail: si-opz@dsv.su.se or ozacas@uem.mz).

Henrik Boström is with the Department of Computer and Systems Sciences, Stockholm University, Forum 100 SE-164 40 Kista, Sweden (e-mail: henrik.bostrom@dsv.su.se).

Least Squares Support Vector Machines models. These studies used DHF datasets and climatic variables.

This study goes further by using the four basic climatic variables and IRS, to predict the incidence of malaria using machine learning. IRS is given as the percentage of homes subjected to anti-mosquitoes fumigation campaign, i.e. homes that allowed the indoor residual spray campaigns in their premises. The paper employ regression trees and ensembles of such trees (forests) using the R-statistical tool [11], through the application of *Rpart* [12] and *Random Forest* [13] routines. The prediction models use a data set comprising: number of malaria cases, IRS, climatic factors (rainfall, minimal and maximal temperature, and relative humidity) for ten years (1999-2008), and nominal district and month variables. The models are applied to nine different data subsets in order to advise for:

- Expected malaria cases
- Districts with low or high incidence of malaria cases
- Months of possible outbreak
- Relevance of Indoor Residual Spray (IRS) campaigns

The paper is organized as follows: a methodology description including employed datasets and learning algorithms is provided in Section 2. Results are presented in Section 3 and in Section 4, conclusions and directions for future work are given.

II. Methods

A. Data

The data set used in this study includes:

- Number of malaria cases and IRS campaigns (1999-2008).
- Set of climatic factors (rainfall, minimal and maximal temperature, and relative humidity, for years 1999 to 2008).
- Eight administrative districts of Maputo province.

Malaria, IRS and climatic factors data sets are monthly aggregated for a ten year period (1999-2008), yielding 960 examples (records) divided in a training set of 864 (nine years, 1999-2007) and a test set of 96 (one year, 2008) instances respectively. Data were initially arranged in eight attributes with two nominal (month and district) and the remaining numerical. Preprocessing of data was performed in two stages (using Weka 3-7-5 software):

- Analysis of missing predictor variables amounting for 16.7% of data. Though trees [12] are unaffected, random forests [13] cannot handle missing predictor values. Thus, missing variable replacement with global mean of each attribute (predictor) was performed in order to allow model comparisons based on similar datasets.
- Conversion of nominal data into binary. This was useful in order to better investigate the importance of categorical variables (see Fig. 1).

The data are presented in different scales and units as e.g. millimeters, percentages and degrees centigrade. Although it does not have any impact on the predictive performance of trees and forests, reduction to similar scale to allow fair comparison of parameters was applied through normalization by scaling all numerical variables in the interval [-1.0, 1.0]. This resulted in twenty seven features including the introduced temperature variation variable determined as the difference between maximal and minimal temperatures. Tenfold cross-validation was employed on training dataset for tuning the parameters of random forest:

m - number of predictors to consider on splitting each node and, *ntree* - the number of trees. Two strategies were applied:

1. Based on multiples of $m = \sqrt{p}$, where *p* - is the number of variables or features.

2. Using the default $m = \sqrt{p}$, half of default and twice the default value.

Table 1 shows the values found for these parameters. The values used in the analysis are *m = 15* and *ntree = 900 trees*.

TABLE I
TUNING PARAMETERS OF RANDOM FOREST

m	ntree	Error	Dispersion
Based on multiples of number of variables to split at each node			
5	900	0.01737	0.00781
10	900	0.01608	0.00647
15	900	0.01598	0.00632
20	900	0.01621	0.006208
25	900	0.01632	0.006139
Based on default, half and twice the default value of *m*			
3	900	0.02189	0.010887
5	900	0.018075	0.008532
10	900	0.016367	0.007511

B. Trees and Forest of Trees

Two types of models were considered: decision trees and ensembles. Decision-trees are generated by means of recursive partitioning (divide and conquer) [12]. These tree-shaped structures allow for representing sets of decisions. The recursive partitioning package Rpart [12] was used to perform tree regression predictions on supplied data in the R-statistical analyses environment.

The ensemble modeling approach deals with combinations of multiple prediction models. This results in classifier models that typically improve performance of a single classifier [14]-[16]. Bagging [14] is one of the standard approaches of ensemble building. Random forests were firstly introduced by Breiman and are based on bagging which builds a large collection of de-correlated trees, followed by aggregation through averaging of their results. The approach is to decrease variance levels of bagging by reducing the correlation between the trees in the forest. This is achieved by randomly selecting

$m \leq p$ input variables as candidates for splitting before each node split, where p is the total number of variables in dataset.

The random forest package employed in R statistical analyses tool, fitted regression trees to malaria, in-door residual spray, environmental and administrative district data set; and then combine the predictions from all the trees to produce the desired results. The package provides two estimates of variable importance: *permutation accuracy* and *node impurity*. This research reports importance using permutation accuracy as it is the most often recommended method [16]. The idea basically is to see how poorly the model performs when each predictor variable is assigned random realistic values, and the remaining variables are kept unchanged. The worse the model performs without a given random variable; the more important it is considered to be for predicting the response variable. The calculation of estimates is done on each tree of the forest and the prediction error is determined on the portion of out-of-bag data. The same calculation is performed for each variable using a random permutation of its values. Finally, averages of differences in prediction errors over all trees are calculated for each variable. In random forest using regression trees, the result is the percentage increase of mean squared errors (reported on the scale 0-100%), with higher values indicating more important variables.

III. RESULTS

Results from applying the nine models generated from different time windows to predict the number of malaria cases and the level of malaria incidence are shown in Table 2.

TABLE II
MODEL PERFORMANCE STATISTICS

Model	Mean Squared Error	Correlation Coefficient
Nine Years: 1999 – 2007		
Tree	0.0328	0.8965
Forest	0.0509	0.8897
Eight Years: 2000 – 2007		
Tree	0.0332	0.891
Forest	0.0514	0.8943
Seven Years: 2001 – 2007		
Tree	0.0314	0.8894
Forest	0.0496	0.9010
Six Years: 2002 – 2007		
Tree	0.0375	0.8851
Forest	0.0457	0.8976
Five Years: 2003 – 2007		
Tree	0.0389	0.8877
Forest	0.0454	0.8866
Four Years: 2004 – 2007		
Tree	0.0339	0.8626
Forest	0.0442	0.8889
Three Years: 2005 – 2007		
Tree	0.0384	0.8911
Forest	0.0401	0.8998
Two Years: 2006 – 2007		
Tree	0.0379	0.3142
Forest	0.0171	0.7116
One Year: 2007		
Tree	0.0305	-0.00248
Forest	0.0213	0.4117

An analysis of the table shows that the model using trees only, obtains its best performance in time-window of one year when looking at its MSE value, i.e. predicting new malaria cases based on the data from previous year data only. This result however, is associated with a weak negative correlation, as shown by its correlation coefficient value. The random forest models obtain good performance as from three to one year time-windows. However, its best performance is observed for the model of two years data with a linear correlation of observed versus predicted that exceeds 70%. Hence, if we are interested in predicting the actual number of malaria cases the forest model should be chosen considering its high predictive performance. The observed differences in error are statistically significant for most of the models, except for the models obtained from the three and four years. Table 3 shows the p-values for the nine periods when comparing the single tree model to the corresponding forest model, as calculated by the paired t-test within [11], with significant criterion value of 5%, i.e. $\alpha = 0.05$.

TABLE III
COMPARISON OF MODEL PERFORMANCE

Period	P-value
Nine Years: 1999 – 2007	1.181E-05
Eight Years: 2000 – 2007	1.013E-05
Seven Years: 2001 – 2007	1.627E-05
Six Years: 2002 – 2007	0.0005609
Five Years: 2003 – 2007	0.002427
Four Years: 2004 – 2007	0.3666
Three Years: 2005 – 2007	0.09023
Two Years: 2006 – 2007	1.739E-09
One Year: 2007	0.02108

Fig. 1 illustrates the measured relative importance of each predictor variable for best forest model using the two year time window. The most important predictor is IRS. This rates these campaigns very highly for control and reduction of malaria incidence risk in the studied region. In the group of environmental variables, the most important predictors are minimal temperature and rainfall. The humidity variable comes last in the list. From a regional (administrative) perspective, the districts that exhibit high incidence of malaria

are Manhiça, Matola and Marracuene; followed by Magude and Matutuine. The results for the month attribute show that January is associated with highest malaria incidence, followed by July, June, September, August and October. In November however, malaria incidence is on the lowest level. The high ranking obtained for January agrees with general assumption that most malaria cases are observed in the rainy season.

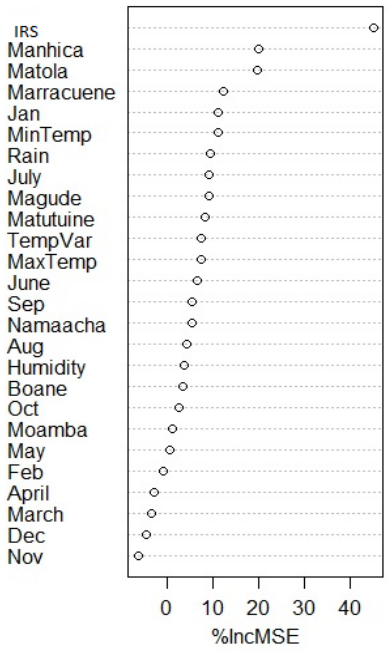

Fig. 1 Variable Importance of Two Years Model

IV. CONCLUSION AND FUTURE WORK

This paper studied the prediction of malaria cases in the Maputo province, Mozambique. It employs multiple regression using single trees and forests with 900 trees as implemented in [11]. Models were generated using number of malaria cases, names of administrative districts, month of the year, level of indoor residual spray applied, measurements of: temperature (maximum and minimal), relative humidity and rainfall as attributes, considered in different time frames. The subdivision of the basic data set into several sub-training sets allowed for the analysis and determination of the time frame that best predicts future malaria cases and incidence. The model allowed for recognizing the most important variables in the study. The most important factors according to the analysis were the implementation of indoor-residual spraying campaigns, followed by the districts of Manhiça, Matola and Marracuene. The three districts have characteristics of being crossed by waterways from regular levels and also having extensive areas with marshy ground. Thus, in rainy season the stagnant waters increase the mosquitoes breeding sites. The month of January, minimal temperature and rainfall also enter the list of the seven most important factors in the analysis of the predicted number of malaria cases in this study. Furthermore, a consistent application of indoor residual spray may lead to a stable decrease of malaria cases in most of the districts in the study, as the results indicate.

The provision of prediction models of future malaria cases is expected to benefit health authorities, policy makers and communities in their main objective of strengthen prevention and control of this disease. Most of all, it will be of great advantage for local communities in understanding and increasing the awareness of how to prevent malaria epidemic becoming pervasive. In order to predict future malaria cases, the use of data from the previous two years appears to result in the most accurate models, according to the findings in this study.

Future studies will focus on the inclusion of the time interval from mosquito bites to the first appearance of malaria symptoms, including additional information considered relevant. Further work will also consider the proposition of a health decision making system so as to contribute to the management of all activities of the malaria program in an integrated and efficient way.

ACKNOWLEDGMENT

The authors would like to thank SIDA and UEM-project for Global Research in Mathematics, Statistics and Informatics at Eduardo Mondlane University for supporting this research. Thanks are also due to the Ministry of Health - Maputo Provincial Directorate of Health and the National Institute of Meteorology for providing data for this project.

REFERENCES

[1] Instituto Nacional de Estatística (INE)*: Publicações Periódicas de Indicadores de Saúde. Mozambique 2007.*

[2] A. Gomez-Elipe, A. Otero and M. van Herp, A. Aguirre-Jaime, "Forecasting malaria incidence based on quarterly cases reports and environmental factors in Karuzi, Burundi, 1997-2003," *Malaria Journal*, 2007, 6: 129.

[3] O. J. T. Briet, P. Vounatsou, D. M. Gunawardena, G. N. L. Galappaththy, P. H. Amersinghe, "Models for short term malaria prediction in Sri Lanka," *Malaria Journal*, 2008, 7: 76.

[4] D. S. Ross, M. J. Crawford, R. G. Donald *et al*, "Mining plasmodium genome database to define organellar function: what does the apicoplast do?," *Philo. Transactions of Royal Society Biological Sciences (2002)*, vol. 357, pp. 35-46.

[5] M. Llinás and H. A. del Portillo, "Mining the malaria transcriptome," *Trends in parasitology*, vol. 21, issue 8, 2005, pp. 350-352.

[6] X. Zhou, Y. Peng and B. Liu, "Text mining for traditional Chinese medical knowledge discovery: A survey," *Journal of Biomedical Informatics*, 2010.

[7] O. Olugbenga, O. Uzoamaka and O. Nwinyi, "A knowledge-based data mining system for diagnosing malaria related cases in healthcare management," *Egyptian Computer Science Journal*, vol. 34 (4), May 2010.

[8] N. Rachata, P. Charaoenkwan et al, "Automatic prediction system of dengue haemorrhagic-fever outbreak risk by using entropy and artificial neural network," *Proceedings of international symposium in information technology (ITSim 2008)*, 2008, pp. 210-214.

[9] K. B. Tan, H. L. Koh and Su L. Teh, "Modeling dengue fever subject to temperature change," *Proceedings from international conference in fuzzy systems and knowledge discovery (FSKD'09)*, 2009, pp. 61-65.
[10] N. A. Husin, N. Salim and A. R. Ahmad, "Modeling dengue outbreak prediction in Malaysia: A comparison of neural network and nonlinear regression model," *Proceedings of international symposium in information technology (ITSim 2008)*, 2008, pp. 1-4.
[11] R-Statistical tool for data analysis, url = http://CRAN.R-project.org/ - accessed September 16, 2012.
[12] T. A. Therneau and B. Atkinson, "Manual of Rpart – recursive partitioning," *R package version 3.1-53*, url = http://CRAN.R-project.org/package=rpart – accessed September 16, 2012.
[13] A. Liaw and M. Wierner, "Classification and regression by randomForest," *R News*, volume 2:3, pp. 18–22. url = http://CRAN.R-project.org/doc/Rnews/ - accessed September 16, 2012.
[14] L. Breiman, "Bagging predictors," *Machine Learning 24 (2)*, 1996, pp. 209-217.
[15] E. Bauer and R. Kohvi, "An empirical comparison of voting classification algorithms: bagging, boosting and variants," *Machine Learning 36 (1-2)*, 1998, pp. 105-139.
[16] L. Breiman, "Random forests," *Machine Learning 45*, 2001, pp. 5-32.