**Bioinformation** by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Disease Management Model**

# Management of filariasis using prediction rules derived from data mining

**Duvvuri Venkata Rama Satya Kumar, Kumarawsamy Sriram, Kadiri Madhusudhan Rao and Upadhyayula Suryanarayana Murty***
Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology, Uppal Road, Hyderabad - 500 007, Andhra Pradesh, India;
Upadhyayula Suryanarayana Murty* - Email: murtyusn@gmail.com; * Corresponding author

received March 21, 2005; revised March 29, 2005; accepted April 04, 2005; published online April 06, 2005

## Abstract:

The present paper demonstrates the application of CART (classification and regression trees) to control a mosquito vector (*Culex quinquefasciatus*) for bancroftian filariasis in India. The database on filariasis and a commercially available software CART (Salford systems Inc. USA) were used in this study. Baseline entomological data related to bancroftian filariasis was utilized for deriving prediction rules. The data was categorized into three different aspects, namely (1) mosquito abundance, (2) meteorological and (3) socio-economic details. This data was taken from a database developed for a project entitled "Database management system for the control of bancroftian filariasis" sponsored by Ministry of Communication and Information Technology (MC&IT), Government of India, New Delhi. Predictor variables (maximum temperature, minimum temperature, rain fall, relative humidity, wind speed, house type) were ranked by CART according to their influence on the target variable (month). The approach is useful for forecasting vector (mosquito) densities in forthcoming seasons.

**Key words:** Disease management; vector-borne disease; bancroftian filariasis; data mining; classification; regression trees

### Background:

Public health management requires an understanding of disease transmission, vector control and disease morbidity. Bancroftian filariasis is a mosquito borne disease, infecting nearly 60 million people in South East Asian countries. The annual economic loss due to filariasis in India alone is U.S$1.5 billion. **[1-3]** The tropical and sub-tropical climate facilitates the proliferation of the mosquito vector (*Culex quinquefasciatus*) for filariasis. **[4, 5]** The mosquito borne disease is a threat to human population despite the practice of several control strategies. **[6]** Proper planning and implementation of control measures require adequate exploitation of the available data for disease management. Therefore, it is of interest to develop prediction methods to augment existing mosquito control strategies. Murty *et al.* used rule-based systems for rapid and accurate identification of malaria causing 54 Indian *Anopheline* mosquito species. **[7]** Thus, the use of prediction models in disease management has been realized. **[8-10]** These tools help epidemiologists to predict the future courses of vector borne diseases. Here, we derive decision rules for vector surveillance using CART (classification and regression tree).

### Methodology:

### Dataset:

A mosquito abundance dataset consisting of 5790 subjects or records with 15 attributes each reflecting the meteorological and socio-economic conditions influencing mosquito survival was used. The mosquito density was expressed in PMHD (per man hour density), which is the total catch of female *Culex quinquefasciatus*, per hour spent for mosquito collection. **[4]**

### Dataset Processing:

The raw data was stored in Excel 2000 (Microsoft Corporation). The data consists of several fields describing each attribute. The attributes include (1) collection date, (2) door number, (3) village name, (4) taluk name (5) district name, (6) unit name, (7) family background, (8) number of children, (9) knowledge of filariasis, (10) house type (11) maximum temperature, (12) minimum temperature, (13) total rainfall, (14) relative humidity, (15) wind speed and (16) mosquito density (PMHD). Seven of the sixteen attributes were further used for developing association rules. These include (1) maximum temperature, (2) minimum temperature, (3) total rainfall, (4) relative humidity, (5) wind speed, (6) house type and (7) mosquito density. These attributes form the independent (predictor) variables. The dependent (predictive) variable is month describing different seasons and climatic conditions of the region. All variables except house type and month are continuous. The four house types include, (1) hut, (2) RCC (reinforced concrete cemented), (3) thatched and (4) tiled.

**Table 1: Classification of predictive variable based on predictor variables**

| | | | Predictor (independent variables) | | | | | Predictive (dependent variable) |
|---|---|---|---|---|---|---|---|---|
| **S. No** | **WS (Km/hr)** | **Max. Temp (x) (${}^0$C)** | **Min. Temp (${}^0$C)** | **RH (%)** | **TRF (mm)** | **HT** | **P.M.H.D** | **M** |
| 1 | 1.5< to <=6.5 | 32.4< to <40.15 | <=21.85 | > 54 | <=261 | Any | >17.75 to 18.03 | February |
| 2 | NC | NC | >21.85 | NC | <=9.45 | Any | >20.75 | March |
| 3 | <= 8.5 | 36.95 < to <=40.15 | >21.8 | NC | <= 54 | Any | <= 2.4295 | April |
| 4 | <=4.5 | <=34.9 | >21.85 | >54 | >19.7 to <=142 | Any | >20.75 | April |
| 5 | <=8.5 | <=40.15 | NC | NC | NC | Any | <=2.42 | May |
| 6 | NC | 35.6< to <=38.8 | <=25.1 | NC | NC | Any | >11.7 to <=13.7 | June |
| 7 | <=6.5 | <=34.9 | NC | NC | >26.6 to <=261 | Any | >17.7 to <=18.03 | August |
| 8 | <=8.5 | 33.4< to <=34.2 | >21.85 | NC | NC | Thatched, Tiled, RCC | >11.7 to <=13.75 | September |
| 9 | <=8.5 | 33.4< to <=34.2 | >21.85 | > 142.4 | NC | Hut | >13.75 | September |
| 10 | NC | 33.4< to <=35.1 | <=25.1 | NC | NC | Hut | >11.7 to <=13.75 | October |
| 11 | NC | NC | >21.85 | >142.2 | NC | RCC | >29.2 to <=51 | October |
| 12 | NC | 33.4< to <=35.1 | <=25.1 | NC | NC | Thatched, Tiled | >39 to <=44.9 | October |
| 13 | NC | NC | >21.85 | >142.2 | NC | Hut | >63 to <=84 | November |
| 14 | NC | NC | >21.85 | >142.2 | NC | Thatched, Tiled, RCC | >51 to <=64 | November |
| 15 | NC | NC | >21.85 | >142.2 | NC | Hut | >64 to <=84 | December |
| 16 | NC | NC | >21.85 | >142.2 | NC | Thatched, Tiled, RCC | <=64 | December |
| 17 | 1.5< to <=6.5 | 32.4< to <=36.05 | >21.85 | NC | <=261 | Any | >17.75 to <=18.03 | January |

**Table 1:** Classification of predictive variable based on predictor variables
WS = wind speed; Max. Temp = maximum temperature; Min. Temp = minimum temperature; TRF = total rainfall; HT = house type; M = month; NC = not considered for classification by CART and P.M.H.D = per man hour density

**Data formats:**
The raw data was stored in EXCEL and the analysis was performed using a commercial software CART (Salford systems Inc. USA). Hence, the raw data was converted to a CART compatible CSV (comma delimited) format.

**Data mining tool**
CART version 5.0 from Salford Systems, California, USA, was used for the current analysis. **[11]** CART is a robust and powerful tree based tool for data classification. **[12]** The tool is suited for the analysis of categorical (classification) and continuous (regression) datasets. The tool uses binary recursive partitioning, in which the parent nodes are exactly split into two child nodes in a recursive manner until the tree is terminated. This depends on the rules used for splitting each node in a tree until the tree is complete. In this process, each terminal node is assigned to a class outcome. CART contains sound statistical tool that enables the development of fast and accurate models. The steps used in the analyses are summarized as follow. The CSV formatted data is loaded to CART using the user interface. The loaded data is used to select and define independent variables (predictor) and predictive (dependent) variables. In this analysis, we defined month as predictive and the other seven variables as predictors. The GINI splitting function is used to maximize the average purity of two child nodes. **[12]** CART contains two tree types, namely (1) classification and (2) regression. The predictive variable (month) is categorical in this analysis. Hence, we used classification type tree model for this analysis.

**Results:**
The CART analysis generated a decision tree with 133 terminal nodes based on the selection criteria. Every terminal node represents a decision rule. Out of the 133 terminal nodes, 17 decision rules were in agreement with meteorological and socio-economic parameters. The decision rules (IF – THEN) used in this analysis are given in Table 1. Data in Table 1 shows the distribution of *Culex quinquefasciatus* density (≤ 2.42 to 84) in PMHD unit over different months of a year. A very low PMHD of ≤ 2.42 is reported for rules #3 and #5 in Table 1. These values correspond to the summer months April and May. This observation corresponds to high maximum temperature (≤ 40.15 $^{o}$C in April and >40.15 $^{o}$C in May) during these months. Thus, high temperature is an influencing parameter for low PMHD in April and May. However, it is also found that the PMHD is >20.75 in April when relative humidity (>54 %) and rainfall is high (≤ 142 mm). Interestingly, PMHD is significantly high during the monsoon and post monsoon months (June, August, September, October, November, December, January and February).

**Discussion:**
The disease transmission dynamics is modeled using the parameters such as vector (pathogen transmitting agent) surveillance, parasitic load in the human community and sudden environmental changes. **[6]** We used data mining tools in CART to find relationships between vector data and the predictive variable. These relations are generally hidden in a large dataset. The <IF-THEN> rules in the CART system is used for the prediction of filarial transmission vectors in an effective way.

The PMHD recorded during the summer months for rules #3 and #5 show that there is no risk of filariasis when the role of other influencing parameters is negligible. In Table 1, for rule #4, the PMHD is high due to high relative humidity and total rainfall. This results in an increased risk of disease transmission under these conditions in April. During the months of October, November and December, a high PMHD (>29.2 to <=84) is recorded for different house types (rules #11, #13, #14, #15 and #16). These rules suggest that the relative humidity is a critical variable on vector density. For rules #1, #2, #7 and #17, the PMHD is elevated due to high total rainfall. Table 1 shows that the four predictors, namely, (1) total rainfall, (2) maximum temperature, (3) minimum temperature, (4) relative humidity and (5) wind speed influenced the target variable in descending order. This is helpful in ranking the predictive variables. Thus, decision trees play an important role in the management of vector borne diseases.

**Conclusion:**
The principal vector for bancroftian filariasis is the mosquito *Culex quinquefasciatus*. Surveillance of the filariasis vector is an important issue in disease management. Here, we show that decision rules help to predict and forecast mosquito density during different months of a year in the region. Thus, prediction of vector density is important towards the effective control of vector borne diseases.

**Acknowledgement:**
Authors are grateful to the Director, IICT, Hyderabad for his continuous support and help. Thanks are due to Ministry of Communication and Information Technology, Government of India, New Delhi for their financial support.

**References:**

**[01]** E. A. Ottesen, *et al., Bull.Wld.Hlth.Org.,* 75:491 (1997) [PMID: 9509621]
**[02]** K. D. Ramaiah, *et al., Parasitol. Today,* 16:251 (2000) [PMID: 10827432]
**[03]** K. S. Snehalatha, *et al*., *Acta Trop*., 88:3 (2003) [PMID: 12943970]
**[04]** D. V. R. S. Kumar, ***et al.,*** *Southeast Asian J Trop Med Public Health*, 35:587 (2004) [PMID: 15689071]
**[05]** U. Suryanarayana Murty, *et al., South-East Asian J Trop Med. Pub. Hlt.,* 33:702 (2002) [PMID: 12757213]
**[06]** D. J. Gubler, *Emerging Infectious Diseases,* 3:1 (1998) [PMID: 9716967]
**[07]** U. Suryanarayana Murty, *et al*., *Comput Appl Biosci.,* 12: 491 (1996) [PMID: 9021267]
**[08]** C. V. Broome & J Loonsk, *MMWR Morb Mortal Wkly Rep*., 24: 199 (2004) [PMID: 15717392]
**[09]** D. C Sharma, *The Lancet Infectious Diseases,* 12: 66 (2002) [PMID: 11901646]

**[10]** P. H. Dessein, *et al., The Journal of Rheumatology*, 32: 435 (2005) [PMID: 15742434]

**[11]** http://www.salford-systems.com

**[12]** M. Garzotto, *et al., J Clin Oncol.,* 21 (2005) [PMID: 15781880]

**Edited by P. Kangueane**

**Citation: Kumar** *et al.,* Bioinformation 1(1): 8-11, (2005)