Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Spatiotemporal Bayesian networks for malaria prediction

Peter Haddawy[a,*], A.H.M. Imrul Hasan[a], Rangwan Kasantikul[a], Saranath Lawpoolsri[b], Patiwat Sa-angchai[b], Jaranit Kaewkungwal[b], Pratap Singhasivanon[b]

[a] *Faculty of ICT, Mahidol University, 999 Phuttamonthon 4 Rd, Salaya, Nakhonpathom 73170 Thailand*
[b] *Faculty of Tropical Medicine, Mahidol University, 420/6 Ratchawithi Rd, Bangkok 10400 Thailand*

ARTICLE INFO

Article history:
Received 25 November 2016
Received in revised form
12 September 2017
Accepted 4 December 2017

Keywords:
Bayesian networks
Probability logic
Spatiotemporal models
Malaria prediction

ABSTRACT

Targeted intervention and resource allocation are essential for effective malaria control, particularly in remote areas, with predictive models providing important information for decision making. While a diversity of modeling technique have been used to create predictive models of malaria, no work has made use of Bayesian networks. Bayes nets are attractive due to their ability to represent uncertainty, model time lagged and nonlinear relations, and provide explanations. This paper explores the use of Bayesian networks to model malaria, demonstrating the approach by creating village level models with weekly temporal resolution for Tha Song Yang district in northern Thailand. The networks are learned using data on cases and environmental covariates. Three types of networks are explored: networks for numeric prediction, networks for outbreak prediction, and networks that incorporate spatial autocorrelation. Evaluation of the numeric prediction network shows that the Bayes net has prediction accuracy in terms of mean absolute error of about 1.4 cases for 1 week prediction and 1.7 cases for 6 week prediction. The network for outbreak prediction has an ROC AUC above 0.9 for all prediction horizons. Comparison of prediction accuracy of both Bayes nets against several traditional modeling approaches shows the Bayes nets to outperform the other models for longer time horizon prediction of high incidence transmission. To model spread of malaria over space, we elaborate the models with links between the village networks. This results in some very large models which would be far too laborious to build by hand. So we represent the models as collections of probability logic rules and automatically generate the networks. Evaluation of the models shows that the autocorrelation links significantly improve prediction accuracy for some villages in regions of high incidence. We conclude that spatiotemporal Bayesian networks are a highly promising modeling alternative for prediction of malaria and other vector-borne diseases.

1. Introduction

Malaria remains a global public health problem with an estimated 214 million cases of malaria globally in 2015 and 438,000 malaria deaths [1]. Since malaria is prevalent in less developed and more remote areas in which public health resources are often scarce, prediction and targeted intervention are essential elements in effective malaria control. Modeling of malaria is challenging because disease transmission can exhibit spatial and temporal heterogeneity, spatial autocorrelation, and seasonal variation. In addition, some covariates such as temperature affect incidence rates in a nonlinear fashion.

Numerous techniques have been used to create predictive models [2] including regression [3], Autoregressive Integrated Moving Average (ARIMA) [4], Susceptible-Infected-Recovered (SIR) models [5], and Neural Networks [6]. No work has yet explored the potential of Bayesian networks as a malaria modeling tool. A Bayesian network is a graphical representation of probability distribution in which nodes represent random variables and links represent direct probabilistic influence among the variables. The relation between a node and its parents is quantified by a conditional probability table (CPT), specifying the probability of the node conditioned on all combinations of the values of the parents. The structure of the network encodes information about probabilistic independence such that the CPTs along with the independence relations provide a full specification of the joint probability distribution over the random variables represented by the nodes. By decomposing a joint probability distribution into a collection of smaller local distributions (the CPTs), a Bayesian network provides a highly compact representation of the complete joint distribution, making it possible to represent and compute with probability distributions over hundreds and thousands of variables. Bayesian networks provide a

[*] Corresponding author.
E-mail addresses: peter.had@mahidol.ac.th, haddawy@gmail.com (P. Haddawy).

https://doi.org/10.1016/j.artmed.2017.12.002
0933-3657/© 2017 Elsevier B.V. All rights reserved.

number of advantages for modeling of malaria, including the ability to represent uncertainty and handle missing data, the ability to represent nonlinear relations, and the availability of efficient algorithms for diagnostic and predictive reasoning as well as sensitivity analysis. In addition, the model structure, which typically reflects the problem structure, can be used to provide explanations.

In this paper we explore the use of Bayes nets to model malaria, demonstrating the approach with village-level weekly prediction models for Tha Song Yang district in northern Thailand. We first create a dynamic Bayes net that models malaria in each village. The network is learned from two years of case data as well as environmental covariates. The network models incidence over time and captures time lagged and nonlinear effects. Evaluation on test data shows that the Bayes net has prediction accuracy in terms of mean absolute error of about 1.4 cases for 1 week prediction and 1.7 cases for 6 week prediction. Comparison of the Bayes net prediction accuracy with several traditional modeling approaches shows the Bayes net to outperform the other models on the most important cases: longer time horizon prediction of high incidence transmission. We produce a binary version of this network for predicting outbreaks and show that it has an ROC AUC prediction accuracy on high incidence villages of above 0.9 for all time horizons. We then elaborate the model with links between the village models to capture spatial autocorrelation of malaria incidence. This results in some very large models which would be far too laborious and error prone to build by hand. So we represent the models as collections of probability logic rules and automatically generate the networks. Evaluation of the models shows that the autocorrelation links significantly improve prediction accuracy for some villages in regions of high incidence.

2. Related work

Work on malaria prediction has used numerous techniques including various types of regression [3], ARIMA models [4], SIR based models [5], and AI techniques such as neural networks [6]. Models are most commonly built with weekly or monthly temporal resolution and spatial resolutions range from village to district to province, with district being the most common. Here we discuss a few of the most relevant examples of work on models for malaria prediction. Zinser et al. [2] provide a nice comprehensive survey of work on malaria prediction.

Kiang et al. [6] produce predictive models for malaria in nineteen provinces of Thailand, including Tak province in which Tha Song Yan is located. They use neural networks with data on total number of monthly provincial malaria cases for the years 1994 through 2001, as well as data on air temperature, rainfall, relative humidity, and NDVI. The predictor variables in their model include the meteorological variables of the current month, the rainfall of the previous month, and time, but not previous cases. The data is divided into five years for training and one year for testing. Malaria cases are divided into 20 bands with the classification considered correct if the prediction falls into the correct band or one of the two adjacent bands. Using this measure, prediction accuracy for Tak province on the test data is found to be 67%. The authors mention that proximity to the border of Tak and some of the other provinces complicates malaria prediction because of imported cases due to migration.

Kulkarni et al. [7] use occurrence records for malaria vectors in north eastern Tanzania and select among 11 temperature and 8 precipitation bioclimactic variables as well as land cover classification to produce binary habitat suitability maps for each of the vector species for 24 villages. Land in a buffer region around each village is classified as suitable or unsuitable. The niche models are produced using maximum entropy. Altitude with and without the percent suitable habitat around each of the 24 villages is then used to predict malaria prevalence in children aged 2–9 years. Linear regression is used for the altitude only models and conditional autoregressive modeling (CAR) is used for the models with altitude and habitat information. Evaluation on 25% of the data reserved for testing shows the model including the habitat variable to significantly outperform the one with only altitude.

Zinser et al. [8] produce Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) models to predict malaria in six catchment areas in Uganda, with each catchment having a population of approximately 60,000. One-week predictions are produced with horizons of 1–52 weeks. Clinical data used includes confirmed cases, numbers of individuals tested, and numbers of individuals treated with various antimalarial drugs. Environmental variables include temperature, rainfall, and enhanced vegetation index (EVI). The predictors used in the final models vary by catchment. About half the predictor series are lagged, with the lags determined using pre-whitening ranging from 1 to 52 weeks. Data is divided into training and testing sets and accuracy evaluated using symmetric mean absolute percentage error (SMAPE). While high frequency variation in cases is best predicted for the short-term horizons (1–4 weeks), the peaks are predicted 1–4 weeks after they occur. The SMAPE is best when the observed counts are low or zero.

Haghdoost et al. [9] produce a Poisson regression model to predict malaria in Kahnooj district of Iran. The dataset consists of confirmed P. vivax and P. falciparum malaria cases for the years 1994 through 2001. Meteorological variables used are mean daily temperature, relative humidity, and rainfall. The predictive model uses the meteorological variables as well as number of previous cases to predict pf and pv cases. They use a 10-day (dekad) temporal resolution. Various values of time lag are selected based on Pearson correlation and the best fitting one then chosen, resulting in a model with a time lag of three dekads (one month) between all explanatory variables and the predicted variable. The data is divided into six years of training data and two years of test data with performance evaluated in terms of mean absolute percent error (MAPE).

Teklehaimanot et al. [10] use ten years of data on weekly confirmed PF malaria cases in ten districts of Ethiopia as well as temperature and rainfall to produce weekly predictions in each of the districts. They use Poisson regression with lags of 4–12 weeks for rainfall and 4–10 weeks for minimum and maximum temperatures, as well as an autoregressive term based on the number of cases 4, 5, and 6 weeks before. Due to the time lags used, the prediction horizon is set to 4 weeks. Accuracy of predictions are evaluated on one year of held out data using percentage of correct predictions above a given threshold as measure as well as potentially prevented cases by comparing to alerts generated by a detection system based on using actual cases. The predictions estimate the overall patterns well but underestimate the heights of the largest peaks and some predictions lag behind the actual values.

Gomez-Elipe et al. [11] develop a model to predict malaria in a province of Burundi highlands using data on monthly notifications of malaria cases (based on symptoms), as well as data on rainfall, mean maximum temperature, and NDVI. They use an ARIMAX model to produce monthly predictions with all variables lagged by one month, resulting in a one month prediction horizon. Time lags are determined by cross-correlation after pre-whitening of the case time series. Data is separated into training and testing sets but the distribution of cases in the two sets differs greatly, with the test set containing no periods of high incidence as in the training set. Model accuracy is reported in terms of the R^2 value (82%) for their linear model. Graphs of actual and predicted malaria rates show that the predictions seem to track the actual rate, lagged by one month.

Buczak et al. [11] develop a model to predict malaria in 64 regions of South Korea using weekly case data for the provinces as well as data on Democratic People's Republic of Korea (DPRK) cases, DPRK mosquito net distribution, DPRK malaria control financing,

Please cite this article in press as: Haddawy P, et al. Spatiotemporal Bayesian networks for malaria prediction. Artif Intell Med (2017), https://doi.org/10.1016/j.artmed.2017.12.002