Exploring Complexity in Health: An Interdisciplinary Systems Approach
A. Hoerbst et al. (Eds.)

doi:10.3233/978-1-61499-678-1-773

Spatiotemporal Bayesian Networks for Malaria Prediction: Case Study of Northern Thailand

Peter HADDAWY[1], Rangwan KASANTIKUL, A.H.M. Imrul HASAN, Chunyanuch RATTANABUMRUNG, Pichamon RUNGRUN, Natwipa SUKSOPEE, Saran TANTIWARANPANT and Natcha NIRUNTASUK
Faculty of Information and Communication Technology, Mahidol University

Abstract. 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 of inferences. This paper explores the use of Bayesian networks to model malaria, demonstrating the approach by creating a village level model with weekly temporal resolution for Tha Song Yang district in northern Thailand. The network is learned using data on cases and environmental covariates. The network models incidence over time as well as evolution of the environmental variables, and captures time lagged and nonlinear effects. Out of sample evaluation shows the model to have high accuracy for one and two week predictions.

Keywords. Bayesian networks, spatiotemporal models, malaria prediction

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. Predictive models commonly make use of environmental factors such as rainfall, temperature, and vegetation as determinants of mosquito vector density and infectivity, as well as malaria incidence in the preceding time period (typically week or month) as an estimator of the human reservoir of the parasite and the population susceptibility [2]. Since seasons affect the environmental factors, models also often incorporate some representation of seasonality.

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 [3] including regression [2], ARIMA [4], SIR based models [5], and Neural Networks [6]. No work has yet explored the potential of Bayesian networks as a malaria modeling framework.

[1] Corresponding Author: haddawy@gmail.com

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. 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 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 of the predictions.

In this paper we explore the use of Bayes nets to model malaria, demonstrating the approach with a village-level weekly prediction model for Tha Song Yang district in northern Thailand. The network is learned from case data as well as environmental covariates. The network models incidence over time and evolution of the environmental variables, and captures time lagged and nonlinear effects. Out of sample evaluation shows the model to have high accuracy for one and two week predictions.

2. Related Work

While no previous work has used Bayes nets to build predictive models for malaria or other infectious diseases, relevant work includes application of Bayes nets to environmental modeling, modeling of non-infectious disease, and making explicit uncertainty in geospatial information. Most Bayes net environmental models to date have either focused on spatial aspects [7, 8] or temporal aspects [9], with only the recent work of Wilkinson et al [10] addressing the combined dimensions of spatial heterogeneity, spatial influence, and temporal evolution. Relevant work on using Bayes nets for disease modeling includes that of Cooper et al. [11] on modeling spatiotemporal patterns for non-contagious diseases that can cause outbreaks in a population such as may occur in bioterrorist attacks. Laskey et al [12] show how to use Bayes nets to reason about cross-country mobility. They create a separate Bayes net for each map pixel, tailored to the features in the pixel, but with no temporal aspect. They link the networks to a GIS and provide a bivalent visualization of the predictions and the degrees of confidence in them.

3. Geographic Region and Data

We demonstrate our approach with the problem of weekly village-level malaria prediction in Tha Song Yan district of Tak province of Thailand. Tha Song Yang is a hilly area with 66 villages in which malaria is endemic. It is located along the border with Myanmar and this proximity to the border results in imported cases. Policy makers

were interested in having a predictive model that can assist in timely targeted intervention, as well as in understanding the factors that most influence the malaria incidence.

The case data for our model consists of weekly clinically confirmed malaria cases obtained from Thailand's national E-Malaria Information System (EMIS) [13]. The data covered each of the 66 villages for the years 2012 and 2013, providing a total of 6,579 records with 12,800 total cases (PF, PV). The numbers of cases per village per week ranged from 0 to 82 with a mean of 2.1.

In addition to the case data, our model makes use of a number of environmental factors associated with malaria. The factors and the source for each are: Normalized Difference Vegetation Index (NDVI) - monthly satellite data from MOD11A3; Land Surface Temperature (LST) - monthly satellite data at 5 km resolution from MOD11C3; Rainfall - daily satellite data at 10 km resolution from JAXA Global Rainfall Watch; Slope - average in 1 km buffer around each village, computed from elevation data; Distance to nearest stream - distance from village center to closest point on the stream; Stream density - total stream length in 4 km buffer; Distance to border - distance from village center to the closest point on the border with Myanmar; and Month - month of the year. NDVI, LST, Rainfall, and Month are temporal variables whose values are indexed by week, while Slope, Stream density, Distance to nearest stream, and Distance to border are non-temporal variables whose values are constant over time. The variables NDVI, Distance to nearest stream, and Stream density are known to positively impact malaria incidence. LST has a nonlinear effect on malaria with malaria incidence low for low temperatures, increasing over some region, and then dropping off for high temperatures. Rainfall is known to have a positive effect on malaria incidence except for very heavy rainfall which can wash away the larvae. Slope is included because it interacts with rainfall, with rain draining off more quickly the higher the slope. Distance to border is a proxy for the number of imported cases and is thought to have a positive effect on incidence. Some values for the variables obtained from satellite data were missing due to cloud cover during some time periods. Missing values were filled in using temporal and spatial interpolation as appropriate.

4. Bayesian Network Prediction Model

Malaria may be modeled using one Dynamic Bayes net (DBN) per village. Figure 1 shows the structure of the DBN prediction model for two time slices: week 0 and week 1. A DBN is a probabilistic representation of the state of a system over time. Time is modeled discretely with a fixed interval between time slices. Temporal nodes represent the state of a random variable at a point in time, such as NDVI at week zero (NDIV_w0), and non-temporal nodes represent random variables whose state does not change, such as Border Distance. Temporal nodes are organized into time slices, representing the state of the system at a point in time. A DBN contains two types of links. Links within a time slice represent probabilistic relations among variables at a given instant and links between time slices represent temporal correlation and lagged effects. The link from NDVI_w0 to NDVI_w1 indicates that NDVI values tend to persist over time. Time lags in the model include a one week lag in the effect of Rainfall on NDIV and a three week lag in the effect of Rainfall on Mosquito Population Density.

Our malaria model includes three latent variables: Rainfall_Effect_w1, which represents the interaction of rainfall and slope; Stream_Effect, which summarizes the effect of stream distance and stream density; and Mosquito_pop_density_w1, which

represents the effect of various environmental factors on the vector density. Inclusion of these variables increases the explanatory power of the network and, importantly, reduces the size of some of the conditional probability tables. For example, inclusion of Mosquito_pop_density_w1 reduces the size of the CPT for the node Incidence_w1 which would otherwise be too large to learn from the available data.

The model is used for prediction by entering known values for variables at week zero (w0), rainfall at week minus 2 (Rainfall_wm2), and Month for weeks one and two and computing the posterior probability of incidence at week 1 (Incidence_w1). To predict incidence for week two, an additional time slice is included with similar repeated structure. The predicted incidence is then the expected value of the incidence random variable. As shown in Figure 2, predictions are displayed in color on a map using a modification of the Bayesian network Classification tool [14], which is implemented as an extension to ArcGIS.

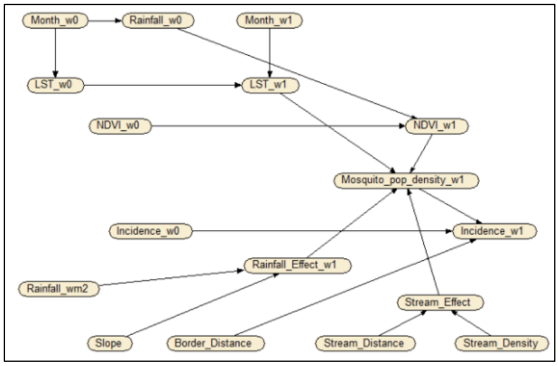

Figure 1. Bayesian network prediction model – two time slices

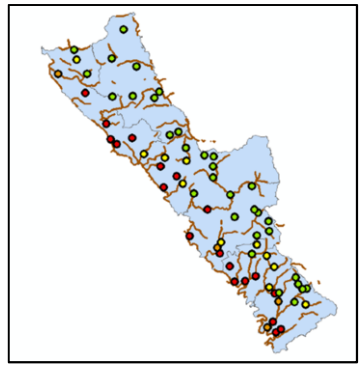

Figure 2. Village-level malaria predictions

5. Evaluation

We evaluated the out of sample prediction accuracy using k-fold cross validation with k = 11. The folds were stratified by time so that all villages were represented in the training and testing data. The average of the Mean Absolute Error (MAE) for the Bayes net model over all 66 villages is 1.44 (SD = 1.73) for one week prediction and 1.61 (SD = 1.88) for two week prediction. Sensitivity analysis of the model shows the previous week incidence to be the most influential factor, followed by Distance to Border. The influence of Mosquito population density increases for moderate values of previous week incidence and then falls off for high values.

6. Conclusions & Future Research

We have shown how Bayesian networks may be used for accurate malaria prediction at high temporal and spatial resolution. The networks are able to integrate environmental and case information and make use of temporal and non-temporal covariates. The current model does not incorporate spatial autocorrelation among villages, which we know to exist in the modeled region. Including this information increases model complexity,

particularly for predictions beyond one week, making it impossible to build such models by hand. We are currently working on using automated Bayes net construction techniques [15] to build such models from libraries of model fragments.

Acknowledgements

This paper is based upon work supported by the US Army International Technology Center Pacific (ITC-PAC) under contract FA5209-15-P-0183. We thank Saranath Lawpoolsri for providing the malaria and environmental data.

References

[1] WHO, World Malaria Report 2015, World Health Organization, ISBN 978 92 4 156515, 2015.
[2] A. Gomez-Elipe, A. Otero, M. van Herp, A. Aguirre-Jamie, Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1977-2003, Malaria Journal, 6 (2007), 129.
[3] K. Zinszer, A.D. Verma, K. Charland, T.F. Brewer, J.S. Brownstein, Z. Sun, D.L. Buckeridge, A scoping review of malaria forecasting: past works and future directions, BMJ Open 2012;2:e001992.
[4] K. Wagdi, P. Singhasivanon, T. Silawan, S. Lawpoolsri, N. White, J. Kaewkungwal, Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan, Malaria Journal, 9 (2010), 251
[5] K. Laneri, A. Bhadra, E.L. Ionides, et al. Forcing versus feedback: epidemic malaria and monsoon rains in northwest India. PLoS Computational Biology, 6 (2010), 1-13.
[6] R. Kiang, F. Adimi, V. Soika, et al. Meteoroligical, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand. Geospatial Health 1 (2006), 71-84.
[7] Johnson, S. Mengersen, K., de Waal, A., Marnewick, K., Cillers, D., Houser, A.M., Boast, L. Modelling cheetah relocation success in southern Africa using an iterative Bayesian network development cycle. Ecological Modeling, 221 (2010), 641-651.
[8] Dlamini, W.M. A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland, Environmental Modelling & Software, 25:199-208, 2010.
[9] Johnson, S., Fielding, F., Hamilton, G., Mengersen, K. An integrated Bayesian network approach to Lyngbya majuscula bloom initiation, 69:27-37, 2010.
[10] Wilkinson, L., Chee, Y.E., Nicholson, A.E., Quintana-Ascencio, P. An object-oriented spatial and temporal Bayesian network for managing willows in an American heritage river catchment. UAI Workshop on Models for Spatial, Temporal, and Networked Data, July 2013.
[11] Cooper, G.F., Dash, D.H., Levander, J.D., Wong, W., Hogan, W.R., and Wagner, M.M. Bayesian Biosurveillance of Disease Outbreaks, Proceedings of the 20th conference on Uncertainty in Artificial Intelligence (UAI '04). AUAI Press, Arlington, Virginia, United States, 94-103, 2004.
[12] Laskey, K.B., Wright, E.J., da Costa P.C.G. Envisioning uncertainty in geospatial information, International Journal of Approximate Reasoning, 51:209-223, 2010.
[13] A. Khamsiriwatchara, P. Sudathip, S. Sawang, et al, Artemisinin resistance containment project in Thailand. (I): Implementation of electronic-based malaria information system for early case detection and individual case management in provinces along the Thai-Cambodian border, Malaria Journal, 11 (2012), 300.
[14] ArcGIS Bayesian Classification Tool Addin, User's Guide, CSER, University of Queensland, http://ww2.gpem.uq.edu.au/CRSSIS/tools/bngis/BNClassificationAddinManual.pdf, accessed 6 March 2016.
[15] L. Ngo and P. Haddawy. Answering Queries from Context-Sensitive Probabilistic Knowledge Bases. Theoretical Computer Science, 171(1997), 147-177.