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,

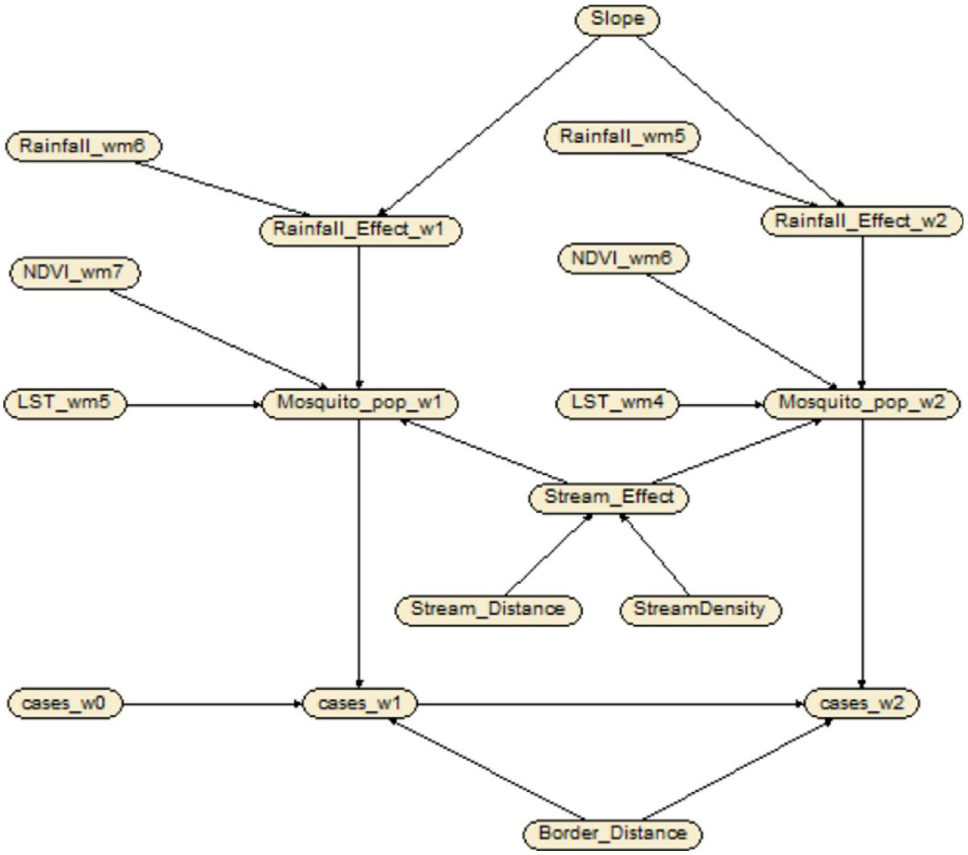


Fig. 1. Initial Structure of Single Village Bayesian Network Model (shown in Netica).

distance of Republic of Korea (ROK) locations from the DMZ, elevation, rainfall, land surface temperature, NDVI and EVI, southern oscillation index, and sea surface temperature anomaly. The distance of locations from the DMZ is included to represent the effect of movement of the vector across the border. They use fuzzy association rule mining to predict three levels of incidence rate: high, medium, and low for prediction horizons of 7–8 weeks. One classifier is trained and then used for all 64 regions. Data are divided into training, fine tuning, and test sets with performance evaluated in terms of positive predictive value, negative predictive value, sensitivity, and specificity. The overall fuzzy association rule mining results are found to be significantly better than those obtained by decision tree, random forest, support vector machine and Holt-Winters exponential smoothing methods.

While no previous work has used Bayes nets to build predictive models for malaria or other vector-borne diseases, relevant work includes application of Bayes nets to environmental modeling, modeling of non-infectious disease, and knowledge-based construction of spatial models. Most Bayes net environmental models to date have either focused on spatial aspects [12,13] or temporal aspects [14], with only the recent work of Wilkinson et al. [15] 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. [16] on modeling spatiotemporal patterns for non-contagious diseases that can cause outbreaks in a population such as may occur in bioterrorist attacks. Their focus is more on population modeling rather than modeling interaction between disease transmission, environmental factors and other disease covariates. Use of knowledge-based model construction linked to GIS data has been successfully applied by Laskey et al. [17] who applied it to generate spatial 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. The case data for our model consists of weekly microscopically confirmed malaria cases obtained from Thailand's national E-Malaria Information System (EMIS) [18]. The data covers each of the 66 villages for the years 2012 and 2013, providing a total of 6579 weekly village reports with 12,800 total cases over all reports (*Plasmodium falciparum, Plasmodium vivax*). The numbers of cases per village per week ranged from 0 to 82 with a mean of 2.1. According to 2013 government census data, the population of Tha Song Yang district was 62,373. The population of the villages ranged from 242 to 4350, with a mean of 945 and a median of 775. The actual populations are known to differ from the official numbers due to large numbers of unofficial migrants from neighboring Myanmar. The use of population data in our models was investigated but found to not be helpful due to its inaccuracy.

In addition to the case data, our model makes use of a number of environmental factors associated with malaria. Predictive models often 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

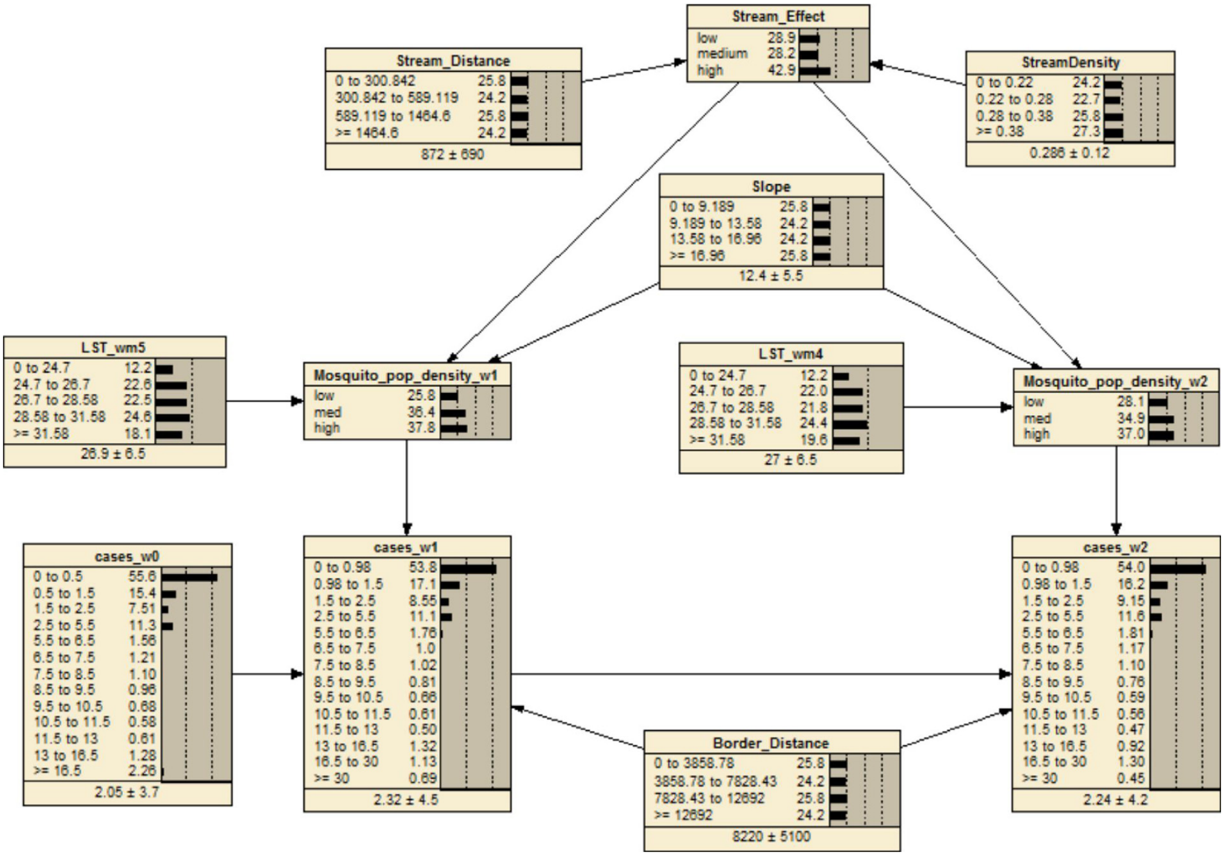


Fig. 2. Final Bayesian Network Prediction Model (shown in Netica).

reservoir of the parasite and the population susceptibility [3]. The factors considered for inclusion in our model 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: Euclidean distance from village center to closest point on the stream,
- Stream density: total stream length in 4 km buffer around each village, and
- Distance to border: Euclidean distance from village center to the closest point on the border with Myanmar.

NDVI, LST, and Rainfall 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 Distance to nearest stream, and Stream density are thought to positively impact malaria incidence. NDVI is generally correlated with malaria transmission, but the relation is complex, with studies in some regions showing a positive correlation with malaria incidence [3] and others a negative correlation [19]. LST has a nonlinear effect on malaria with malaria incidence low for low temperatures, increasing over some range, and then dropping off for high temperatures [20]. Rainfall also has a nonlinear effect, with malaria incidence increasing with rainfall until the point where the flushing effect is reached, at which point it decreases [21]. 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. Model design and development

The model was constructed using the first 70% of the data, with the other 30% reserved for testing. This approach to separating training from testing data was used so that the data remained contiguous, which is necessary for building ARIMA models for comparison with the Bayes net. The first step in model construction is to determine the appropriate time lags for the temporal covariates LST, Rainfall, and NDVI. This is done by computing cross-correlation between the time series for each covariate and the time series for the number of cases. In this process, pre-whitening [22] is first applied to the time series since temporal autocorrelation in the time series can cause spurious correlation. The process consists of fitting an ARIMA model to covariate time series (X), using this to filter the dependent variable time series (Y), and calculating the cross correlation between the residuals for X and the filtered Y. This results in cross-correlation graphs, as shown in Appendix A in Supplementary material. Since the graphs can indicate several potential time lags for a given covariate, each is tested using regression to determine the one with the most predictive power. The analysis resulted in identification of optimal time lags of 6 weeks for LST, 7 weeks for Rainfall, and 8 weeks for NDVI.

Using these lags we produced the initial Bayes net model shown in Fig. 1. Malaria is modeled using one Dynamic Bayes net (DBN)

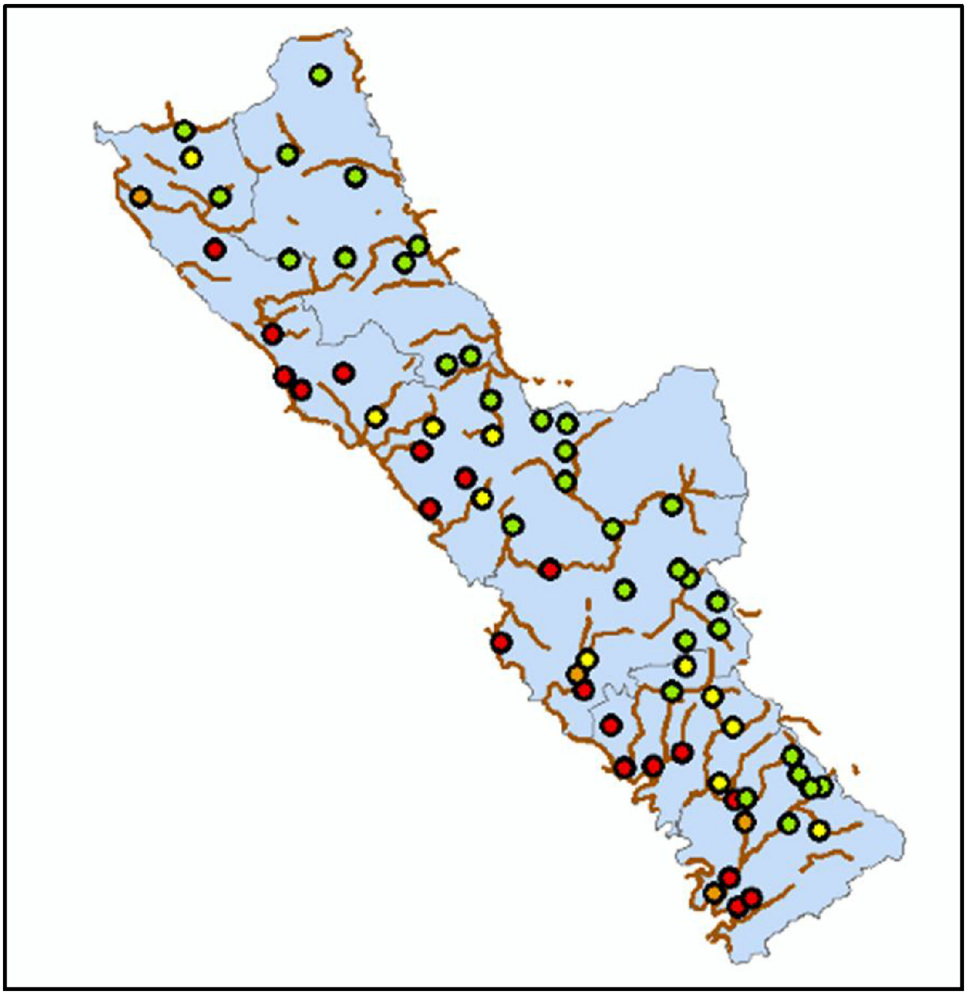


Fig. 3. GIS interface to Bayes net model showing predictions for villages in Tha Song Yang district of Tak Province.

Table 1
Prediction accuracy (MAE) of Bayesian Network model for 1–6 week prediction. Highlighted cells indicate where this model performs best among those evaluated.

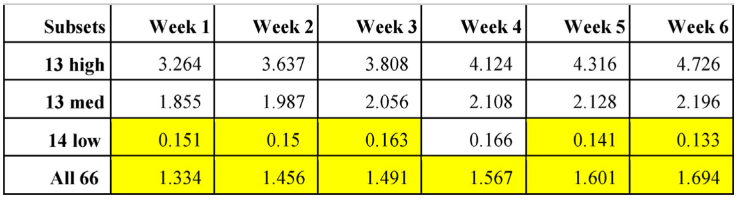
Subsets	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
13 high	3.387	3.404	3.454	3.504	3.43	3.457
13 med	1.886	2.073	2.129	2.319	2.299	2.411
14 low	0.305	0.368	0.422	0.463	0.519	0.538
All 66	1.415	1.501	1.557	1.644	1.657	1.729

per village. Fig. 1 shows the structure of the DBN prediction model for two time slices: 1 week and 2 week prediction. 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; in this case one week. Temporal nodes represent the state of a random variable at a point in time, such as NDVI at week minus 7 (NDVI_wm7), 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. Links within a time slice represent probabilistic relations among variables at a given instant and links between time slices represent lagged effects.

Our malaria model includes three latent variables: Rainfall_Effect_wi represents the interaction of rainfall and slope; Stream_Effect summarizes the effect of stream distance and stream density; and Mosquito_pop_density_wi 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 Cases_w1 which would otherwise be too large to learn from the available data.

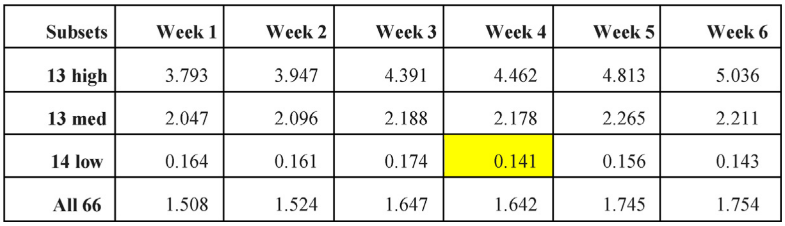
Using backward elimination of covariates on this model showed that performance was improved by removing Rainfall and NDVI. This resulted in the simplified model shown in Fig. 2. In addition to the structure, the figure shows the states of all the nodes and their marginal prior probabilities. The latent variables Stream_Effect and Mosquito_pop_density_wi are discrete. The remaining variables are continuous and needed to be discretized. Initial discretizations were produced by using unsupervised binning in the Weka package [23] with approximately equal data counts in each bin. The discretizations were then fine-tuned experimentally.

Table 2
Prediction accuracy (MAE) of ARIMA model for 1–6 week prediction. Highlighted cells indicate where this model performs best among those evaluated.



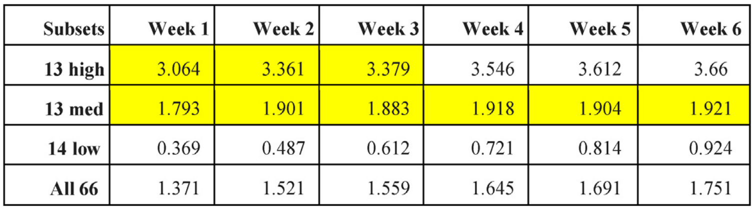
Subsets	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
13 high	3.264	3.637	3.808	4.124	4.316	4.726
13 med	1.855	1.987	2.056	2.108	2.128	2.196
14 low	0.151	0.15	0.163	0.166	0.141	0.133
All 66	1.334	1.456	1.491	1.567	1.601	1.694

Table 3
Prediction accuracy (MAE) of ARIMAX model for 1–6 week prediction. Highlighted cells indicate where this model performs best among those evaluated.



Subsets	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
13 high	3.793	3.947	4.391	4.462	4.813	5.036
13 med	2.047	2.096	2.188	2.178	2.265	2.211
14 low	0.164	0.161	0.174	0.141	0.156	0.143
All 66	1.508	1.524	1.647	1.642	1.745	1.754

Table 4
Prediction accuracy (MAE) of Linear Regression model for 1–6 week prediction. Highlighted cells indicate where this model performs best among those evaluated.



Subsets	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
13 high	3.064	3.361	3.379	3.546	3.612	3.66
13 med	1.793	1.901	1.883	1.918	1.904	1.921
14 low	0.369	0.487	0.612	0.721	0.814	0.924
All 66	1.371	1.521	1.559	1.645	1.691	1.751

The Bayes net model is used for prediction by entering the known value for Cases at week zero (cases_w0), LST at weeks minus 5 and minus 4 (LST_wm5, LST_wm4), and Stream Distance (Stream_Distance), Stream Density (StreamDensity), Slope (Slope), and Border Distance (Border_Distance), and computing the posterior probability of cases at week 1 and week 2 (cases_w1, cases_w2). To predict cases for later weeks, additional time slices are included with similar repeated structure. Predictions for each village are produced by instantiating the Bayes net model with the parameter values for that village. The predicted number of cases is then the expected value of the Cases random variable (e.g. cases_w1). When computing the expected value, the mean of each bin is used based on the distribution of data values over the range of the bin.

As shown in Fig. 3, predictions can be displayed in color on a map using a modification of the Bayesian network Classification tool [24], which is implemented as an extension to ArcGIS. The tool also provides detailed information concerning predictions and permits the user to enter data into the model through the GIS interface.

5. Evaluation

5.1. Accuracy of the Bayes net model and comparison with traditional modeling approaches

The accuracy of the Bayes net model was evaluated on the testing data for one to six week prediction horizons using mean absolute error (MAE) as a metric. We analyzed the prediction accuracy by testing on three different subsets of villages divided according to average weekly incidence:

- 13 villages with high incidence {Min: 0, Max: 82, Ave: 7.43},
- 13 villages with medium incidence {Min: 0, Max: 16, Ave: 1.91}
- 14 villages with low incidence {Min: 0, Max: 3, Ave: 0.099}, and
- all 66 villages containing the entire spectrum of incidence.

Table 5
Prediction accuracy (MAE) of Poisson regression model for 1–6 week prediction.

Subsets	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
13 high	5.49	5.395	5.204	5.098	4.741	4.305
13 med	1.952	1.937	1.896	1.842	1.779	1.692
14 low	1.092	1.091	1.091	1.083	1.093	1.101
All 66	1.975	1.956	1.919	1.902	1.831	1.731

The results are shown in Table 1. Relative to the average incidence, the Bayes net performs best in the high incidence villages, less well in the medium incidence villages, and worst in the low incidence villages. We see this positively since accurate prediction of periods of high incidence is of most importance for targeted intervention and resource allocation.

For comparison purposes, we created prediction models using four traditional approaches: ARIMA, ARIMAX, Linear Regression, and Poisson Regression. The R package was used for all models. All variables appropriately lagged were considered for the Poisson and linear regression models. The Poisson Regression model selected the predictors Stream_Distance, Border_Distance, Slope, Stream_Density, Cases_w0, and LST (lagged 6 weeks) for all prediction horizons. The linear regression model left out slope and included different sets of the predictors for different prediction horizons. The training data for the ARIMA and ARIMAX models was prepared by concatenating the data for the 66 villages and inserting null values between each adjacent pair of village time series in order to not produce spurious patterns. The number of null values (91) was selected to keep the seasonality intact. The best fitting ARIMA model as found to be ARIMA((0,1,0)(1,0,0)). For the ARI-

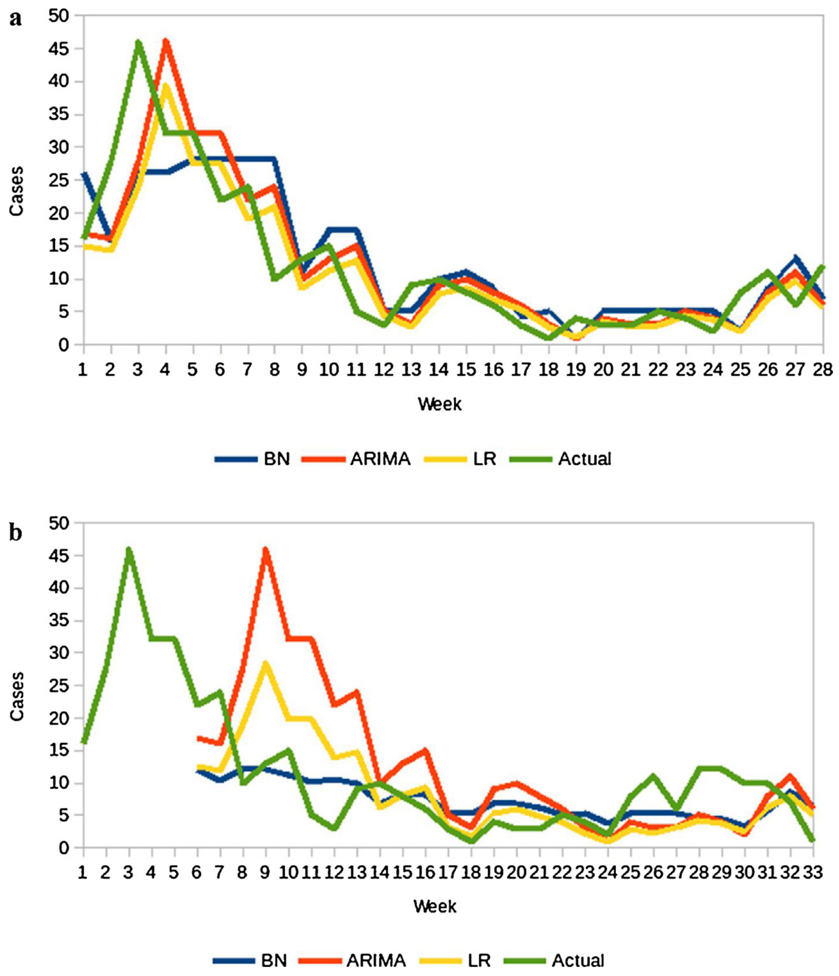


Fig. 4. (a) One week ahead predictions for a high incidence village. (b) Six week ahead predictions for a high incidence village.

Table 6
Performance of Bayes net relative to ARIMA model (% improvement).

Subset	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
13 high	−3.8%	6.4%	9.3%	15.0%	20.5%*	26.9%*
13 med	−1.7%	−4.3%	−3.6%	−10.0%	−8.0%	−9.8%
14 low	−102.0%*	−145.3%*	−158.9%*	−178.9%*	−268.1%*	−304.5%*
All 66	−6.1%	−3.1%	−4.4%	−4.9%	−4.6%	−2.1%

* Indicates difference statistically significant (two-tailed T test $p < 0.05$).

Table 7
Performance of Bayes net relative to Linear Regression model (% improvement).

Subset	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
13 high	−10.5%	−1.3%	−2.2%	1.2%	5.0%	5.6%
13 med	−5.2%	−9.0%	−13.1%	−20.9%*	−20.8%*	−25.5%*
14 low	17.3% *	24.4%*	31.1%*	35.8%*	36.2%*	41.8%*
All 66	−3.2%	1.3%	0.13%	0.06%	2.0%	1.3%

* Indicates difference statistically significant (two-tailed T test $p < 0.05$).

Table 8
Sensitivity of Cases in weeks 1, 4, 6 on previous cases and covariates in terms of variance reduction.

Variable	Cases Week 1	Cases Week 4	Cases Week 6
Cases_w0	13.93	2.492	0.7276
Border_Distance	0.2172	1.401	1.981
Stream_Effect	0.0002575	0.001498	0.001658
LST (lagged 6 weeks)	0.0001495	0.0001589	0.002991
Slope	0.0001251	0.002776	0.006755
Stream_Density	4.653e-05	0.0003213	0.0001194
Stream_Distance	1.32e-05	0.0002400	0.0003457

MAX model only the temporal variables (NDVI, Rainfall, LST) were considered since ARIMAX does not support static variables. Forward selection was used and only the variable LST identified as a significant predictor.

Evaluation of the prediction accuracy of the models is shown in Tables 2–5. The cells for which each model is the top performer are highlighted. The Poisson regression model does not outperform the other models for any prediction scenario and the ARIMAX model performs best for only one scenario. The other three models have clear categories for which they perform best. The Bayes net model performs best for 4–6 week predictions for the high incidence villages. The linear regression model performs best for 1–3 week predictions for the high incidence villages and performs best for all prediction horizons for the medium incidence villages. The ARIMA model performs best for all prediction horizons for the low incidence villages, except for 4 week prediction, where ARIMAX performs better. We can see that the Bayes net, linear regression, and the ARIMA models are complementary. The Bayes net model performs best for the most important scenarios: providing sufficient lead time to target intervention for high incidence transmission.

Tables 6 and 7 compare the performance of the Bayes net model with the ARIMA model and Linear Regression model, respectively. It is particularly instructive to compare the linear regression and the Bayes net models for the high incidence villages. Here the Linear Regression model performs best for 1 − 3 week prediction and the

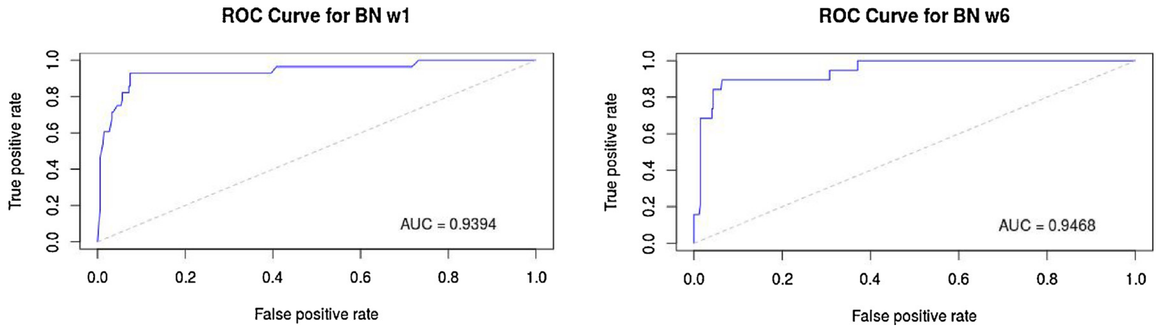


Fig. 5. ROC curves for Bayes net outbreak prediction for weeks 1 and 6.

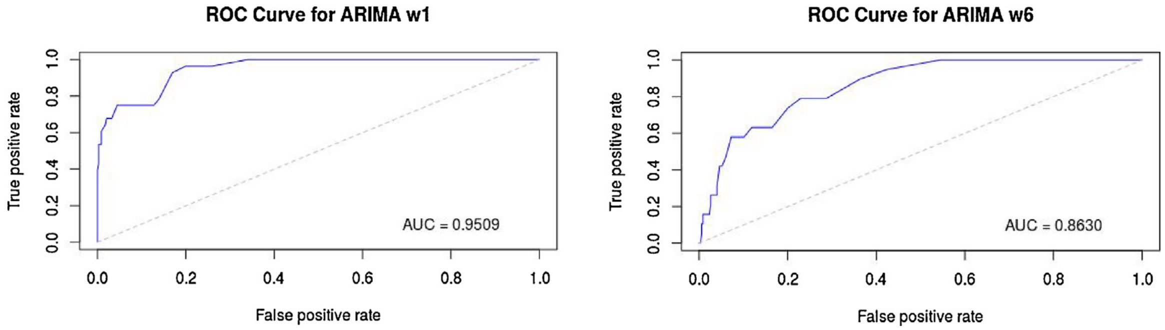


Fig. 6. ROC curves for ARIMA outbreak prediction for weeks 1 and 6.

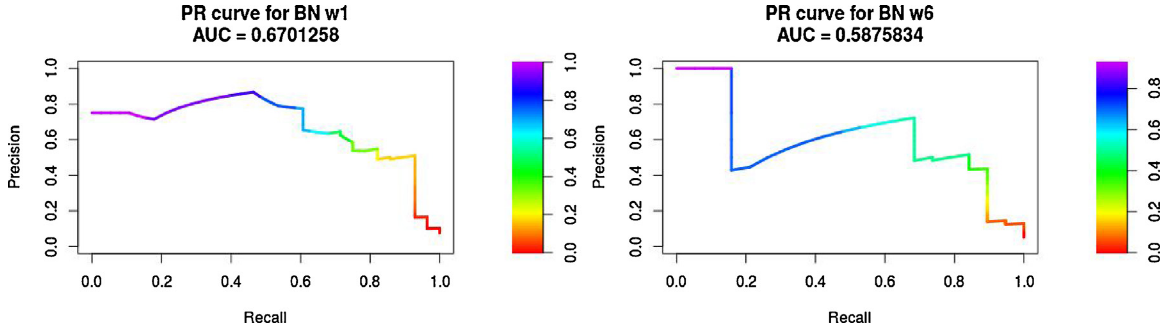


Fig. 7. Precision recall curves for Bayes net outbreak prediction for weeks 1 and 6.

Bayes net performs best for 4–6 week prediction, with the magnitude of the improvement of the Bayes net over the Linear Regression model generally increasing with increasing prediction horizon. A possible explanation for this is the increasing importance of the environmental covariates in the prediction as the horizon increases. In the case of the Linear Regression model, a separate regression model was fitted for each time horizon and we can observe a systematic change in the regression equations as the time horizon increases, with the magnitude of the coefficient on week zero cases decreasing as the time horizon increases and the magnitude of the coefficient on LST increasing. The coefficient on cases for week zero is 0.83 for one week prediction and 0.61 for six week prediction, while the coefficient on LST is 0.054 for one week prediction and 0.21 for six week prediction. Table 8 shows a sensitivity analysis for the Bayesian network prediction of cases in weeks 1, 4, and 6 on cases in week zero as well as the various covariates. We can see that the longer the prediction horizon, the more influence the covariates have relative to week zero cases.

Table 9
Precision recall AUC values of ARIMA, Linear Regression, and binary Bayes net models in predicting outbreaks in the 13 high incidence villages for 1–6 week prediction.

Model	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Bayes Net	0.670	0.667	0.675	0.591	0.566	0.588
ARIMA	0.760	0.714	0.617	0.512	0.358	0.250
Linear Regression	0.763	0.716	0.619	0.527	0.364	0.260

We can gain additional insight into the reason for the superior performance of the Bayes net for longer prediction horizons by examining some prediction graphs. Fig. 4a shows the actual number of cases for one of the high incidence villages and the one week ahead predictions of the Bayes net, ARIMA, and Linear Regression models. The ARIMA and Linear Regression models do a better job of prediction than the Bayes net, but their curves are very close to those of the actual number of cases lagged by one week. This strong reliance of the ARIMA and Linear Regression models on the week zero cases is seen again in Fig. 4b, which shows the curves for six week ahead prediction. The ARIMA and Linear Regression curves again closely mirror the actual number of cases, but now lagged by six weeks, resulting in relatively poor prediction accuracy. Except for the lag, the ARIMA model prediction curve is essentially unchanged from one week to six week prediction. The Linear Regression prediction curve retains the same shape but the peaks are attenuated down in the six week prediction as would be expected from the decreased value of the coefficient on week zero cases mentioned above. In contrast, the Bayes net also tends to follow the week zero cases for one week prediction but shows more of an averaging behavior for six week prediction. This is a consequence of the relatively stronger sensitivity to the environmental covariates for the longer term predictions.

5.2. Outbreak prediction

An important measure of accuracy is the ability to predict outbreaks, i.e. whether the number of cases will exceed a given threshold. For the thirteen high incidence villages, we define an outbreak as any week in which the number of cases exceeds the mean for the villages plus one standard deviation. The mean for the villages is 7.49 and the standard deviation 9.40, resulting in 96 outbreaks out of 845 instances in the training set and 30 outbreaks out of 364 instances in the test set. To predict outbreaks, a variant of the Bayes net in Fig. 2 was produced in which the Cases nodes were replaced with binary nodes which were positive when the number of cases exceeded the defined threshold.

For such a binary classification problem, ROC analysis is a commonly used evaluation technique. Fig. 5 shows the ROC curves for 1 week and 6 week predictions for the 13 villages. The area under the curve (AUC) values for 1–6 week predictions are consistently high, ranging from a low of 0.939 to a high of 0.968. Fig. 6 shows the ROC curve for 1 and 6 week predictions using the previous ARIMA model and thresholding at mean plus one standard deviation. For one week prediction the ARIMA model has an AUC of 0.95 which is slightly higher than the Bayes net, but for 6 week prediction the AUC drops to 0.86, significantly lower than the Bayes net.

Since the number of negative instances for this problem is far larger than the number of positive instances, precision recall (PR) curves can provide a more informative measure of performance [25]. For the binary Bayes net, the AUC for the PR curves ranged from a high of 0.675 (for 3 week prediction) to a low of 0.566 (for 5 week prediction). Fig. 7 shows the PR curves for 1 week and 6 week predictions. With the low frequency of outbreaks, the random baseline PR AUC is 0.082, making the performance of the Bayes net excellent. Fig. 8 shows the PR curves for 1 and 6 week predictions for the ARIMA model with thresholding and Table 9 provides the PR AUC values for 1–6 week outbreak predictions for the binary Bayes net and the ARIMA and Linear Regression models with thresholding. Similar to the results on prediction of number of cases, the performance of the Bayes net relative to the other two models improves with increasing prediction horizon. The ARIMA and linear regression models outperform the Bayes net for 1 and 2 week predictions, but the Bayes net outperforms the other models beginning at 3 weeks until the performance of the Bayes net is more than twice that of the other models at 6 weeks.

6. Automated model construction for modeling spatial autocorrelation

The DBN in Fig. 2 models malaria cases over time in each individual village. But malaria from one village can spread to another, resulting in significant spatial autocorrelation of incidence. An analysis of cross correlation among all pairs of the 66 villages shows highest correlation at 1 week time lag with the correlation dropping off markedly as villages are further than 3 km apart. This dependence of spatial autocorrelation on distance is in line with previous findings in the literature [26,27].

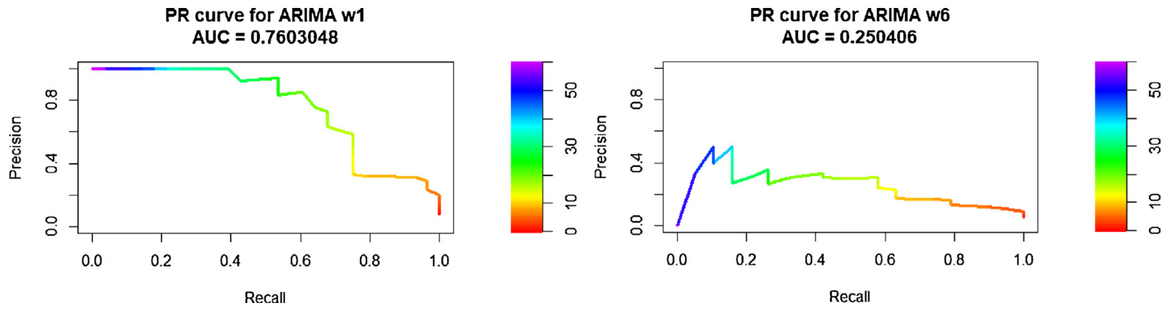
To represent autocorrelation of malaria incidence we augment the DBN in Fig. 2 with links between villages less than 3 km apart. Fig. 9 shows a schematic of the resulting spatiotemporal model for the case of a group of three villages close together. To simplify the figure, the nodes representing the environmental variables have been omitted. In this model the number of cases in a given week is not only a function of the cases in that village in the previous week but also the sum of the cases in the neighbors. The sum of the cases of the neighbors for a given village is represented by the node Neighbor_sum. Notice that the Neighbor_Sum node for week 1 is a function of the predicted values of the cases in the neighboring villages for week 1. Various other network topologies are possible depending upon the distances between villages. For example, one village might sit between two other villages which are close neighbors but the neighbors may not be close to each other. While the previous DBN models could be created by hand, the size and complexity of the spatiotemporal models that include autocorrelation now prohibits this. So we present a technique to automatically generate them.

The network construction algorithm takes a context-sensitive probabilistic knowledge base [28] consisting of a set of context-sensitive probability logic sentences or rules, a specification of the spatial and temporal scope, a set of evidence, and an optional query and generates a Bayes net for that particular modeling problem. Each rule is a schematic representation of the relation between a node and its parents and contains an optional context constraint, specifying under what conditions the probabilistic parent/child relationship holds. Since rules are schematic with all variables implicitly universally quantified, a set of rules specifies a potentially infinite set of Bayesian networks, with each generated Bayes net representing a set of ground instances of the rules. The file for the Bayes net model with spatial autocorrelation is shown in Appendix B in Supplementary material, three of which are

1) FOR (t:TIME)(x:LOCATION) WHERE prop(x,Total_Cases)<15 STATE Neighbor_Sum AT x,t IS (0.0,0.5,Infinity);
2) FOR (t1:TIME)(t2:TIME)(x:LOCATION) WHERE prop(x,TOTAL-CASES)<15 AND t1>=1 AND t2=t1+1 PARENT Cases AT x,t2 IS Cases AT x,t1 AND Border_Distance AT x,t2 AND Mosquito_Pop_Density AT x,t2 AND Neighbor_Sum AT x,t1 CPT INCD207;
3) FOR (t:TIME) (x1:LOCATION) (x2:LOCATION) WHERE t >= 1 AND dist(x1,x2) <= 3000 AND x1! = x2 PARENT NeighborSum AT x1,t IS Cases AT x2,t CPT AUTO_SUM;

The first rule is one of four that specifies conditional discretization for the Neighbor_Sum node. The discretization is dependent on the total cases of the village of which the node is a parent. The adjustment of discretization based on incidence in this way improves prediction accuracy because villages of high incidence are not as subject to influence from their neighboring villages as are villages of low incidence and thus different thresholds are appropriate. While one could in principle use a single fine discretization, doing so would require significantly more data to learn the CPT. The second rule says that Cases at location x and time t2 has parents Border_Distance at x and t2, Mosquito_Pop_Density at x and t2, and Neighbor_Sum at x and t1 (the previous week) with the CPT contained in the variable INCD207. This rule applies to villages with total cases less than 15. The third rule specifies the network fragment for the NeighborSum node. The context part says to create a link to a neighbor village if that village is within 3 km distance and is not the same as the current village. The CPT represents the sum of the incidences of the parents, designated by AUTO_SUM, which is a function in our modeling language that can apply to a variable number of parents.

The network generation engine can be run in two ways. It can generate a network to cover all times and all locations for a given modeling problem or it can generate just that subnetwork to answer a particular query (e.g. malaria incidence in a particular village at a particular time) given particular evidence. For very large problems the query network can be considerably smaller than the


PR curve for ARIMA w1
AUC = 0.7603048

PR curve for ARIMA w6
AUC = 0.250406

Fig. 8. Precision recall curves for ARIMA outbreak prediction for weeks 1 and 6.

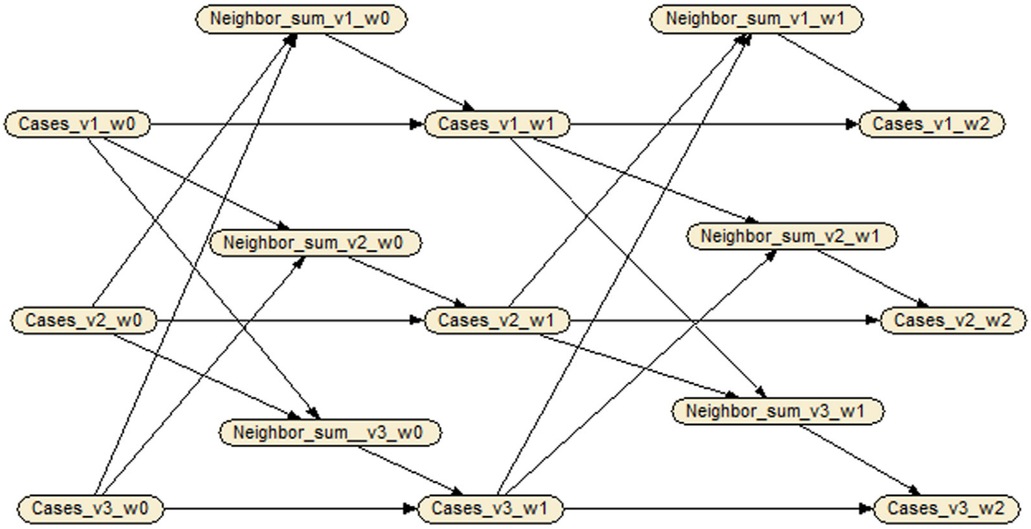


Fig. 9. Bayes net model of spatial autocorrelation among a group of three neighboring villages.

Table 10
Absolute and percent improvement in MAE for the six villages where modeling of spatial autocorrelation with neighboring villages improves prediction accuracy for all prediction time horizons.

Village No.	Total Incidence	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
109	2672	2.15* (17.8%)	1.92* (15.6%)	1.54 (11.8%)	1.31 (8.92%)	0.497 (3.34%)	0.94 (5.78%)
201	412	0.0278 (1.18%)	0.204 (7.75%)	0.226 (8.52%)	0.116 (4.24%)	0.226 (7.66%)	0.158 (5.38%)
208	274	0.0159 (0.88%)	0.0671 (3.34%)	0.237* (11.3%)	0.211 (9.77%)	0.236 (10.6%)	0.236 (10.4%)
205	83	0.123* (12.9%)	0.241* (22.5%)	0.271* (22.2%)	0.223 (18.2%)	0.165 (12.7%)	0.117 (9.05%)
410	79	0.0538 (5.03%)	0.0467 (4.06%)	0.0859 (6.84%)	0.126 (9.15%)	0.165 (11.0%)	0.132 (8.84%)
107	51	0.0326 (4.02%)	0.0542* (6.17%)	0.124 (12.5%)	0.165 (16.2%)	0.234* (21.6%)	0.233 (20.9%)

* Indicates difference statistically significant (two-tailed T test $p < 0.05$).

entire problem network, resulting in considerable savings in inference time. The generated Bayes net is then provided as an input file to a standard Bayes net inference package (we are currently using Hugin) where the network is compiled and evidence is propagated.

To examine the effect of including the links for spatial autocorrelation we identified the subset of 17 villages that have at least one neighbor with Kendall cross-correlation of incidence above 0.2. The number of neighbors of each village ranged from 1 to 3. For six of the villages addition of the autocorrelation links improved predictions for all time horizons; for six villages the links did not help for any prediction horizon; and for five of the villages addition of the links improved some predictions. Table 10 shows the six villages for which addition of the links improved all predictions. The highest percentage improvement is 22.5% for 2 week prediction for village 205. All villages and their neighbors are characterized by relatively high total incidence rates. All villages except 107 have total malaria incidence above the median of 70.5. The total incidence for village 107 is 51. In addition, with the exception of village 107, the neighboring villages of all villages in the table have total incidence at or above the median. The single neighbor of 107 has a total incidence of 44. These results are in line with previously published studies of the effect of autocorrelation on prediction of malaria transmission. In a study of malaria transmission using data from 101 surveys in Mali, Kleinschmidt et al. [29] find significant evidence of spatial autocorrelation and find that inclusion of spatial autocorrelation via kriging improves prediction accuracy of their logistic regression models for five of the surveys. One possible explanation they provide is that the logistic regression models already partially accounted for the spatial autocorrelation through the inclusion of environmental variables. Similar to our results, they further find that inclusion of spatial autocorrelation significantly improves prediction accuracy particularly in areas of high risk.

Our approach to modeling influence among villages is similar to a number of other modeling approaches, including coupled hidden Markov models (CHMM) [30], influence models [31,32], and two-level influence models [33]. A CHMM models interactions of multiple HMMs by directly linking the current state of one model with the previous states of all the other models. The problem with this is that the transition matrix grows exponentially with the number of interacting models. The influence model reduces the complexity of the modeled interaction by representing the influence as a convex combination of pairwise conditional probabilities. The influence model has been used in numerous applications to model influence from social interaction. Pan et al. [32] also introduce a dynamic version in which the weights used to sum influences can change over time. They show how the model can be used to uncover changing dynamics of relations among flu incidence in states of the US over time. Our approach to modeling interaction differs from theirs in that we first sum the incidences of villages and then model the influence of the sum on the village of interest. Zhang et al. [33] introduce the two-level influence model, a dynamic Bayesian network with a two-level structure in which the lower level networks represent the states of individuals over time and the higher level network represents the state of the group. In this model, the state of an individual at a time is a function of its state at the previous time as well as the state of the group at the previous time. They demonstrate the effectiveness of their model in reasoning about multi-player games and multi-party meetings. Since the NeighborSum node in our models can be thought of as the state of a group of individuals, our models can be considered a variation of the two-level influence model.

7. Conclusions & future research

We have shown how Bayesian networks may be used for accurate malaria prediction. Our networks are able to integrate environmental and case information and make use of temporal and non-temporal covariates. 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. Accuracy of predicting outbreaks for high incidence villages ranges from 0.566 to 0.67 PR AUC against a random baseline accuracy of 0.082. 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. This is the case for numeric case prediction as well as binary outbreak prediction. This seems due to the superior ability of the Bayes net to incorporate the effect of environmental covariates, which have increasing importance with longer prediction horizons.

We model the autocorrelation of malaria incidence between villages by adding links between the individual village networks for neighbors that are in close proximity. Since this results in networks that are too large to construct by hand, we represent the models with collections of probabilistic rules and dynamically construct the models. Evaluation of the models shows that the autocorrelation links significantly improve prediction accuracy for some villages in regions of high incidence. Physical proximity is, of course, not the only factor affecting spatial autcorrelation. Malaria may not spread between villages that are close due to separation by geographic features like hills or rivers and it may spread between villages that are more distant because they are linked by roads. It would be worthwhile to investigate whether prediction accuracy could be improved by a more refined model of spatial autocorrelation taking into account such geographic features. The models of cross country of Laskey et al. [17] may be useful for this.

Acknowledgements

This paper is based upon work supported by the US Army International Technology Center Pacific (ITC-PAC) under contract FA5209-15-P-0183. This work was partially supported through a fellowship from the Hanse-Wissenschaftskolleg Institute for Advanced Study, Delmenhorst, Germany to Haddawy.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.artmed.2017.12.002.

References

[1] WHO. World malaria report 2015. World Health Organization; 2015. ISBN 978 92 4 156515.
[2] Zinszer K, Verma AD, Charland K, Brewer TF, Brownstein JS, Sun Z, et al. A scoping review of malaria forecasting: past works and future directions. BMJ Open 2012;2:e001992.
[3] Gomez-Elipe A, Otero A, Van Herp M, Aguirre-Jaime A. Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1997–2003. Malar J 2007;6(September (1)):1.
[4] Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White NJ, Kaewkungwal J. 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. Malar J 2010;9(September (1)):1.
[5] Laneri K, Bhadra A, Ionides EL, Bouma M, Dhiman RC, Yadav RS, et al. Forcing versus feedback: epidemic malaria and monsoon rains in northwest India. PLoS Comput Biol 2010;6(September (9)):e1000898.
[6] Kiang R, Adimi F, Soika V, Nigro J, Singhasivanon P, Sirichaisinthop J, et al. Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand. Geospat Health 2006;1(November (1)):71–84.
[7] Kulkarni MA, Desrochers RE, Kerr JT. High resolution niche models of malaria vectors in northern Tanzania: a new capacity to predict malaria risk? PLoS One 2010;5(February (2)):e9396.
[8] Zinszer K, Kigozi R, Charland K, Dorsey G, Brewer TF, Brownstein JS, et al. Forecasting malaria in a highly endemic country using environmental and clinical predictors. Malar J 2015;14(June (1)):1.
[9] Haghdoost AA, Alexander N, Cox J. Modelling of malaria temporal variations in Iran. Trop Med Int Health 2008;13(December (12)):1501–8.
[10] Teklehaimanot HD, Schwartz J, Teklehaimanot A, Lipsitch M. Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia II. Weather-based prediction systems perform comparably to early detection systems in identifying times for interventions. Malar J 2004;3(November (1)):1.
[11] Buczak AL, Baugher B, Guven E, Ramac-Thomas LC, Elbert Y, Babin SM, et al. Fuzzy association rule mining and classification for the prediction of malaria in South Korea. BMC Med Inform Decis Mak 2015;15(June (1)):47.
[12] Johnson S, Mengersen K, de Waal A, Marnewick K, Cilliers D, Houser AM, et al. Modelling cheetah relocation success in southern africa using an iterative bayesian network development cycle. Ecol Modell 2010;221(4):641–51.
[13] Dlamini WM. A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. Environ Modell Softw 2010;25(February (2)):199–208.
[14] Johnson S, Fielding F, Hamilton G, Mengersen K. An integrated Bayesian network approach to Lyngbya majuscula bloom initiation. Mar Environ Res 2010;69(February (1)):27–37.
[15] Wilkinson LA, Chee YE, Nicholson A, 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 2013 Jul 15 2017:77–86.
[16] Cooper GF, Dash DH, Levander JD, Wong WK, Hogan WR, Wagner MM. Bayesian biosurveillance of disease outbreaks. In: Proceedings of the 20th conference on Uncertainty in Artificial Intelligence 2004 Jul 7. 2017. p. 94–103.
[17] Laskey KB, Wright EJ, da Costa PC. Envisioning uncertainty in geospatial information. Int J Approximate Reason 2010;51(January (2)):209–23.

[18] Khamsiriwatchara A, Sudathip P, Sawang S, Vijakadge S, Potithavoranan T, Sangvichean A, 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. Malar J 2012;11(July (1)):1.

[19] Haque U, Hashizume M, Glass GE, Dewan AM, Overgaard HJ, Yamamoto T. The role of climate variability in the spread of malaria in Bangladeshi highlands. PLoS One 2010;5(December (12)):e14341.

[20] Mordecai EA, Paaijmans KP, Johnson LR, Balzer C, Ben-Horin T, Moor E, et al. Optimal temperature for malaria transmission is dramatically lower than previously predicted. Ecol Lett 2013;16(January (1)):22–30.

[21] Koenraadt CJ, Harrington LC. Flushing effect of rain on container-inhabiting mosquitoes Aedes aegypti and Culex pipiens (Diptera: Culicidae). J Med Entomol 2008;45(January (1)):28–35.

[22] Chatfield C. The analysis of time series: an introduction. 6th ed. London: Chapman & Hall; 2004.

[23] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newslett 2009;11(November (1)):10–8.

[24] 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 September 2017].

[25] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(March (3)):e0118432.

[26] Yeshiwondim AK, Gopal S, Hailemariam AT, Dengela DO, Patel HP. Spatial analysis of malaria incidence at the village level in areas with unstable transmission in Ethiopia. Int J Health Geogr 2009;8(January (1)):5.

[27] Gemperli A, Sogoba N, Fondjo E, Mabaso M, Bagayoko M, Briët OJ, et al. Mapping malaria transmission in west and central africa. Trop Med Int Health 2006;11(July (7)):1032–46.

[28] Ngo L, Haddawy P. Answering queries from context-sensitive probabilistic knowledge bases. Theor Comput Sci 1997;171(January (1)):147–77.

[29] Kleinschmidt I, Bagayoko M, Clarke GP, Craig M, Le Sueur D. A spatial statistical approach to malaria mapping. Int J Epidemiol 2000;29(April (2)):355–61.

[30] Oliver N, Rosario B, Pentland A. Graphical models for recognizing human interactions. Advances in Neural Information Processing Systems 1999:924–30.

[31] Basu S, Choudhury T, Clarkson B, Pentland A. Learning human interactions with the influence model, vol. 539. Cambridge, MA: MIT Media Lab; 2001. Tech. Rep.

[32] Pan W, Dong W, Cebrian M, Kim T, Pentland AS. Modeling dynamical influence in human interaction. IEEE Signal Process Mag 2012;29(May):77–86.

[33] Zhang D, Gatica-Perez D, Bengio S, Roy D. Learning influence among interacting Markov chains. Advances in Neural Information Processing Systems 2006:1577–84.