ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1783

Integrating ARIMA and Spatiotemporal Bayesian Networks for High Resolution Malaria Prediction

A. H. M. Imrul Hasan [1] and **Peter Haddawy**[2]

Abstract. Since malaria is prevalent in less developed and more remote areas in which public health resources are often scarce, targeted intervention is essential in allocating resources for effective malaria control. To effectively support targeted intervention, predictive models must be not only accurate but they must also have high temporal and spatial resolution to help determine when and where to intervene. In this paper we take the first essential step towards a system to support targeted intervention in Thailand by developing a high resolution prediction model through the combination of Bayes nets and ARIMA. Bayes nets and ARIMA have complementary strengths, with the Bayes nets better able to represent the effect of environmental variables and ARIMA better able to capture the characteristics of the time series of malaria cases. Leveraging these complementary strengths, we develop an ensemble predictor from the two that has significantly better accuracy that either predictor alone. We build and test the models with data from Tha Song Yang district in northern Thailand, creating village-level models with weekly temporal resolution.

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 [23]. In Thailand, 31,121 and 15,446 confirmed cases were reported in 2014 and 2015, respectively [18]. Since malaria is prevalent in less developed and more remote areas in which public health resources are often scarce, targeted intervention is essential in allocating resources for effective malaria control. Since 2009 Thailand has implemented an E-Malaria Information System (EMIS) [13] to systematically gather case data and data on relevant covariates in order to support control policy decisions as well as to track their effectiveness. With the rise of resistant strains of malaria as well as the greatly increased incidence of dengue (another mosquito vector borne disease) in Thailand and neighbouring Malaysia, Thailand's Center of Excellence for Biomedical and Public Health Informatics, which houses EMIS, has expressed interest in exploring use of this and related data sources to support targeted intervention by producing appropriate predictive models. To effectively support targeted intervention, predictive models must be not only accurate but they must also have high temporal and spatial resolution to help determine when and where to intervene [16]. While much work has been done on malaria prediction models, high resolution prediction remains a challenge [21].

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. Among the numerous techniques that have been used to create predictive models [30], ARIMA is the most popular because of its ability to accurately model characteristics of the time series as well as capture some dependence on covariates. Despite a variety of modeling approaches (ARIMA, regression, neural nets, SIR models) having been explored, no work has yet explored the potential of Bayes nets as a modeling framework for malaria. Bayesian networks [19] provide a number of advantages for modeling of malaria, including the ability to explicitly represent uncertainty, handle missing data, and represent nonlinear relations. In addition, the model structure, which typically reflects the problem structure, can be used to generate explanations of the predictions.

In this paper we take the first essential step towards effective targeted intervention by developing a high resolution prediction model through the combination of Bayes nets and ARIMA. Bayes nets and ARIMA have complementary strengths, with the Bayes nets better able to represent the effect of environmental variables and ARIMA better able to capture the characteristics of the time series of malaria cases. We find that for one week prediction the Bayes net model performs best for high and mid-level incidence while ARIMA performs better for low-level incidence. For two week prediction, ARIMA performs best for all incidence levels. Leveraging these complementary strengths, we develop an ensemble predictor from the two that has significantly better accuracy that either predictor alone at every incidence level, using model trees to select the features and the weights to put on each model. We build and test the models with data from Tha Song Yang district in northern Thailand, creating village-level models with weekly temporal resolution. This is the first work to use Bayesian networks to model malaria and the first to create an ensemble forecasting model using Bayes nets and ARIMA.

2 RELATED WORK

A number of researchers have explored the combination of neural networks and ARIMA. One approach has been to use the neural network to classify the residuals [6, 14, 28] from the ARIMA model. A residual is the difference between an actual value and it's prediction. The logic behind this is that the residuals will contain non-linearity since ARIMA cannot capture the non-linear structure of time series. Adhikari [1] proposed an approach to combine neural nets with a number of forecasting models (Box-Jenkins ARIMA, FANN, EANN and SVM) in two steps. First a set of sample weights is obtained from the inverse relation between absolute forecast and error forecast of the respective model. Second a neural net model is made to predict the combining weights by going through the sample weights.

[1] Mahidol University, Thailand, email: ahmimrul.has@student.mahidol.ac.th
[2] Mahidol University, Thailand, email: peter.had@mahidol.ac.th

One common combining method for ensemble techniques is to assign weights to component forecasts where each weight is inversely proportional to the prediction error of the corresponding component model, which ensures that the model with higher error gets the lower vote for the combined contribution. A nonlinear framework [2] is also proposed by researchers where correlations between pairs of component forecasts are considered along with the optimal weights determined from pairs of train and test sets. Wichard [24] proposed hybrid ensembles that combine multiple models (Nearest Trajectory, Neural Network, Difference, Trend cycle and AR model) by using a weight which is proportional to symmetric mean absolute percent error (SMAPE) computed over a left-out part of the time series.

Relevant work on using Bayes nets for disease modeling includes that of Cooper et al. [5] on modeling spatiotemporal patterns for non-contagious diseases that can cause outbreaks in a population such as may occur in bioterrorist attacks. Spatiotemporal Bayes nets have been applied to a number of environmental modeling problems. Most Bayes net environmental models to date have either focused on spatial aspects [12, 7] or temporal aspects [11], with only the recent work of Wilkinson et al. [25] addressing the combined dimensions of spatial heterogeneity, spatial influence, and temporal evolution.

In this paper, we combine Bayesian network and ARIMA forecasting models by a simple linear function of weighted component models and a few selected features. The combining weights and features are chosen by a model tree algorithm.

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, particularly given the remoteness of some villages, 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 covers each of the 66 villages for the years 2012 and 2013, providing a total of 6,579 records with 12,800 total cases (plasmodium falciparum, plasmodium vivax). 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. 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 reservoir of the parasite and the population susceptibility [8]. Since seasons affect the environmental factors, models also often incorporate some representation of time or seasonality. The factors included 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,
- Distance to border: Euclidean distance from village center to the closest point on the border with Myanmar,
- 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 thought 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. The model includes temporal nodes such as NDVI at week zero (NDVI_w0), and non-temporal nodes for random variables whose states do not change with time, such as Border_Distance.

Time lags in the model include a one week lag in the effect of Rainfall on NDVI 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. Because of the inclusion of latent variables, the conditional probability tables (CPTs) for the Bayes net were learned using the expectation maximization (EM) algorithm.

The model is used for prediction by entering known values for the variables at week zero (w0), rainfall at week minus 2 (Rainfall_wm2), and Month for weeks zero and one, 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. Each node of malaria incidence is divided into 14 ranges. The predicted incidence is then computed as the expected value of the incidence random variable:

$$E(incidence\_w_1) = \sum_{i=1}^{14} \{P(range_i) * mean(range_i)\} \qquad (1)$$

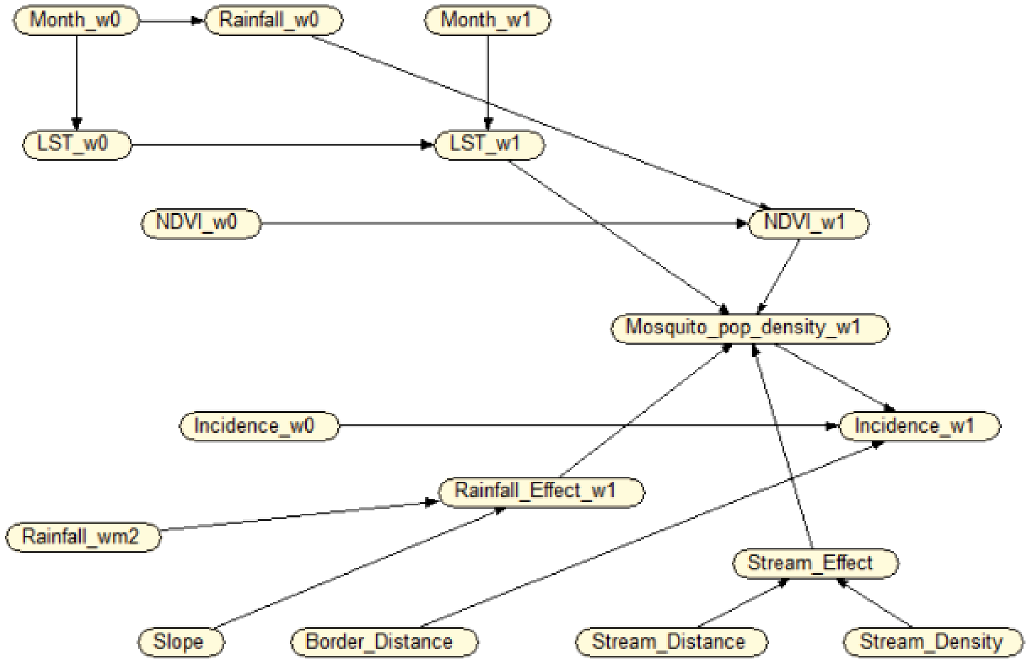

Figure 1. Bayesian network model showing two time slices.w0 = week 0, w1 = week 1, wm2 = week minus 2

where $i = 1, 2, ..., 14$ are the ranges, $P(range_i)$ is the probability of ith range and $mean(range_i)$ is the mean of the distribution of the data over the ith range. The prediction accuracy was evaluated by mean absolute error (MAE):

$$MAE = \frac{\sum_{case=1}^{N} Abs(Predicted_{case} - Actual_{case})}{N} \quad (2)$$

where N is the number of cases. The MAE of the Bayes net model for one week prediction over all 66 villages is 1.098 and the MAE for two week prediction is 1.417.

5 ARIMA PREDICTION MODEL

Auto Regressive Integrated Moving-Average (ARIMA), also known as the Box-Jenkins approach [4], is the most popular stochastic time series forecasting model of the past few decades. It is a modeling approach that can be used to calculate the probability of a future value lying between specific limits. It has three parameters: autoregression (AR), integration (I), and moving average (MA). The ARIMA$(p, 0, 0)$ or autoregressive model is represented as

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t \quad (3)$$

where θ_0 is the intercept, p is the number of auto regressive terms, Y_t is predicted result, Y_{t-p} is the observation of time $t - p$, $\phi_1, \phi_2, ..., \phi_p$ is a set of parameters that are calculated by linear regression, and e_t is the regression error. The ARIMA$(0, 0, q)$ or moving average only depends on q past random terms and the current random term e_t and is expressed as

$$Y_t = \mu - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ...... - \theta_q e_{t-q} + e_t \quad (4)$$

where q is the number of the moving averages, $\theta_1, \theta_2, ..., \theta_q$ is a set of parameters, and μ is the mean of the series. The ARIMA$(p, 0, q)$ model is

$$\begin{aligned} Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \\ \mu - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q} + e_t \quad (5) \end{aligned}$$

The ease of use of the Box-Jenkins methodology [4] for optimal model fitting, as well as the flexibility of the representation have made ARIMA a highly popular modeling approach. A variant of ARIMA capable of modeling seasonal data called SARIMA [9] includes additional terms in the ARIMA model and is written as $(p, d, q)(P, D, Q)_m$. The upper-case notation is for seasonal parts of the model. The term m is the number of units per season, P is the number of seasonal autoregressive (SAR) terms, D is the number of seasonal differences and Q is the number of seasonal moving average (SMA) terms. ARIMA models are used frequently by researchers for disease surveillance and prediction (malaria [8, 22, 29], dengue [20], HFMD [17]). The version of ARIMA that includes external predictors is known as ARIMAX, which is denoted by ARIMA(p, q, d)X (where X is the external independent variables). The ARIMA model with extra variables often performs better [15, 3] than simple univariate ARIMA when the dependent variable is explainable by other external factors.

In this study, the ARIMA models were developed in the R Software package v3.2.3. A best fit was obtained from the combination of all 66 time series (66 villages) by applying the auto.arima() function available in the R package called ”forecast” [10]. We took 70% incidence data for training from each village and concatenated them by inserting Null values in between to keep the seasonality intact. The remaining 30% of the data was used for model testing. We trained a seasonal ARIMA((1,1,0)(1,0,0)) model for this study. For multivariate ARIMA, we chose explanatory variables from the environmental factors (Month_w0, Rainfall_w0, LST_w0, NDVI_w0, Rainfall_wm2) that we used for the Bayes net (section 4). An iterative greedy method was used to select the external variables that provided significant improvement in prediction accuracy based on MAE. The only variable so selected as external covariate was Month_w0. We computed two-period-ahead forecasts by following the rolling window method. Over all 66 villages, the MAE of ARIMA for one week prediction is 1.102 and for two week prediction is 1.217, while the MAE of ARIMAX for one week prediction is 1.074 for two week prediction is 1.251.

6 COMBINING BAYESIAN NETWORK AND ARIMA PREDICTION MODELS

We analyzed the prediction accuracy of the ARIMA and Bayes net models by testing them on three different subsets of villages divided according to average incidence rate: (1) 13 villages with high incidence{Min: 0, Max: 82, Avg.: 7.43}, (2) 13 villages with medium incidences{Min: 0, Max: 16, Avg.: 1.91} (3) 14 villages with low incidence{Min: 0, Max: 3, Avg.: 0.099} and all 66 villages containing the entire spectrum of incidence. The results are shown in table 1 (columns: BN, ARIMA, ARIMAX). It can be seen from the table that the Bayes net and the ARIMA models have complementary strengths. For one week prediction, the Bayes net model has the best performance for high- and mid-level incidence villages and the ARIMA models have the best performance for low-incidence villages. For two week prediction, the ARIMA models perform best for all classes of villages.

We combined the models using stacked generalization [27], where outputs are collected from $level_0$ models (trained on $level_0$ data) and treated as data for another learning problem at $level_1$. The model applied in this step is referred to as the $level_1$ model. For the $level_1$ generalization we used model tree induction [26] on the Bayes net, ARIMA, and ARIMAX predictions along with a number of attributes that characterize the incidence rate in the current and previous weeks as well as overall:

- *incidence_W*$_0$*:* Incidence of the current week (W_0)
- *incidence_rate:* Sum of weekly incidence of a village divided by the maximum sum among all villages
- *incidence_WM1:* Incidence of previous week
- *incidence_Avg:* The average incidence of a village
- *BN_Prediction:* Predicted incidence by Bayesian Network
- *ARIMA_Prediction:* Predicted incidence by the ARIMA model
- *ARIMAX_Prediction:* Predicted incidence by the ARIMAX model
- *Last_Two_Avg:* Average of last two weeks incidence

We used the M5P [26] model tree algorithm from the WEKA (v3.6.10) data mining tool. The test set of $level_0$ was further divided into 70% and 30% then used as the train and test sets of $level_1$. For the combination of the Bayes net and ARIMA, the algorithm generated a tree with a single leaf node for both one- and two-week prediction and selected only incidence rate and current week incidence in addition to the BN and ARIMA predictions:

$$\begin{aligned} incidence\_W_1 = {} & (0.3424 \times incidence\_W_0) + \\ & (0.253 \times ARIMA\_Prediction\_W_1) + \\ & (3.8015 \times incidence\_rate) + \\ & (0.2451 \times BN\_Prediction\_W1) + 0.0128 \end{aligned} \tag{6}$$

$$\begin{aligned} incidence\_W_2 = {} & (0.3224 \times incidence\_W_0) + \\ & (0.2409 \times ARIMA\_Prediction\_W_2) + \\ & (6.6964 \times incidence\_rate) + \\ & (0.0824 \times BN\_Prediction\_W2) + 0.0448 \end{aligned} \tag{7}$$

For the combination of the Bayes net and ARIMAX the algorithm also generated a tree with only one node for one- and two-week prediction. The prediction models include the variables incidence rate, current incidence, and incidence average. For two-week prediction the combining function substitutes the value of last two week average for the BN prediction value.

$$\begin{aligned} incidence\_W_1 = {} & (0.3582 \times incidence\_W_0) + \\ & (0.2254 \times ARIMAX\_Prediction\_W_1) - \\ & (20.226 \times incidence\_rate) + \\ & (0.9385 \times incidence\_avg) + \\ & (0.2112 \times BN\_Prediction\_W1) - 0.0571 \end{aligned} \tag{8}$$

$$\begin{aligned} incidence\_W_2 = {} & (0.6355 \times incidence\_W_0) + \\ & (0.5065 \times ARIMAX\_Prediction\_W_2) - \\ & (26.1348 \times incidence\_rate) + \\ & (1.2706 \times incidence\_avg) - \\ & (0.5763 \times Last\_Two\_Avg) - 0.01 \end{aligned} \tag{9}$$

For one-week prediction, equations 6 and 8 each assign roughly the same weights to the BN and ARIMA/ARIMAX predictions. For two-week prediction, equation 7 assigns significantly lower weight to the BN prediction and equation 9 leaves it out altogether. In the four combination formulas, the relatively large magnitude coefficient on incidence rate is due to its relatively small range of values. The prediction accuracy of the (BN+ARIMA) and (BN+ARIMAX) models is shown in Table 1.

7 RESULTS AND DISCUSSION

We compared the accuracy for one- and two-week predictions of the two ensemble models with the BN and ARIMA models using data from three consecutive weeks from the months of July, September, and November.

Figure 2 shows one-week predictions of BN, ARIMA, ARIMAX, (BN+ARIMA) and (BN+ARIMAX) models versus actual incidence for the three three-week periods for a high incidence village. Figure 2(a) shows that the BN underestimates the peak in week 1 of the first month but does well with the remaining weeks of all three months. ARIMA overestimates weeks 1 and 3 of the first month while ARIMAX overestimates weeks 1 and 2 of the first month and week 1 of second month but both do well on the remaining low incidence weeks. This is consistent with the results in Table 1. The combinations (BN+ARIMA) and (BN+ARIMAX) correct for the major inaccuracies in the other models and perform equally well for this village with only slight differences in their predictions.

Figure 3 shows two-week predictions of BN, ARIMA, ARIMAX, (BN+ARIMA) and (BN+ARIMAX) models versus actual incidence for the same time periods for a second high incidence village. Figure 3(a) shows that the BN fits the three consecutive weeks of first two

Table 1. Prediction Accuracy of BN, ARIMA, ARIMAX, (BN+ARIMA) and (BN+ARIMAX) for one- and two-week prediction.

	Mean Absolute Error									
Models	BN		ARIMA		ARIMAX		(BN+ARIMA)		(BN+ARIMAX)	
Set of villages	W_1	W_2	W_1	W_2	W_1	W_2	W_1	W_2	W_1	W_2
13-high	2.310	3.033	2.504	2.883	2.461	2.967	2.114	2.562	2.007	2.527
13-med	1.421	1.951	1.504	1.730	1.483	1.720	1.259	1.581	1.228	1.485
14-low	0.323	0.461	0.160	0.211	0.130	0.163	0.189	0.232	0.122	0.157
66-all	1.098	1.417	1.102	1.217	1.074	1.251	0.963	1.121	0.911	1.068

W_1 = First Week and W_2 = Second Week

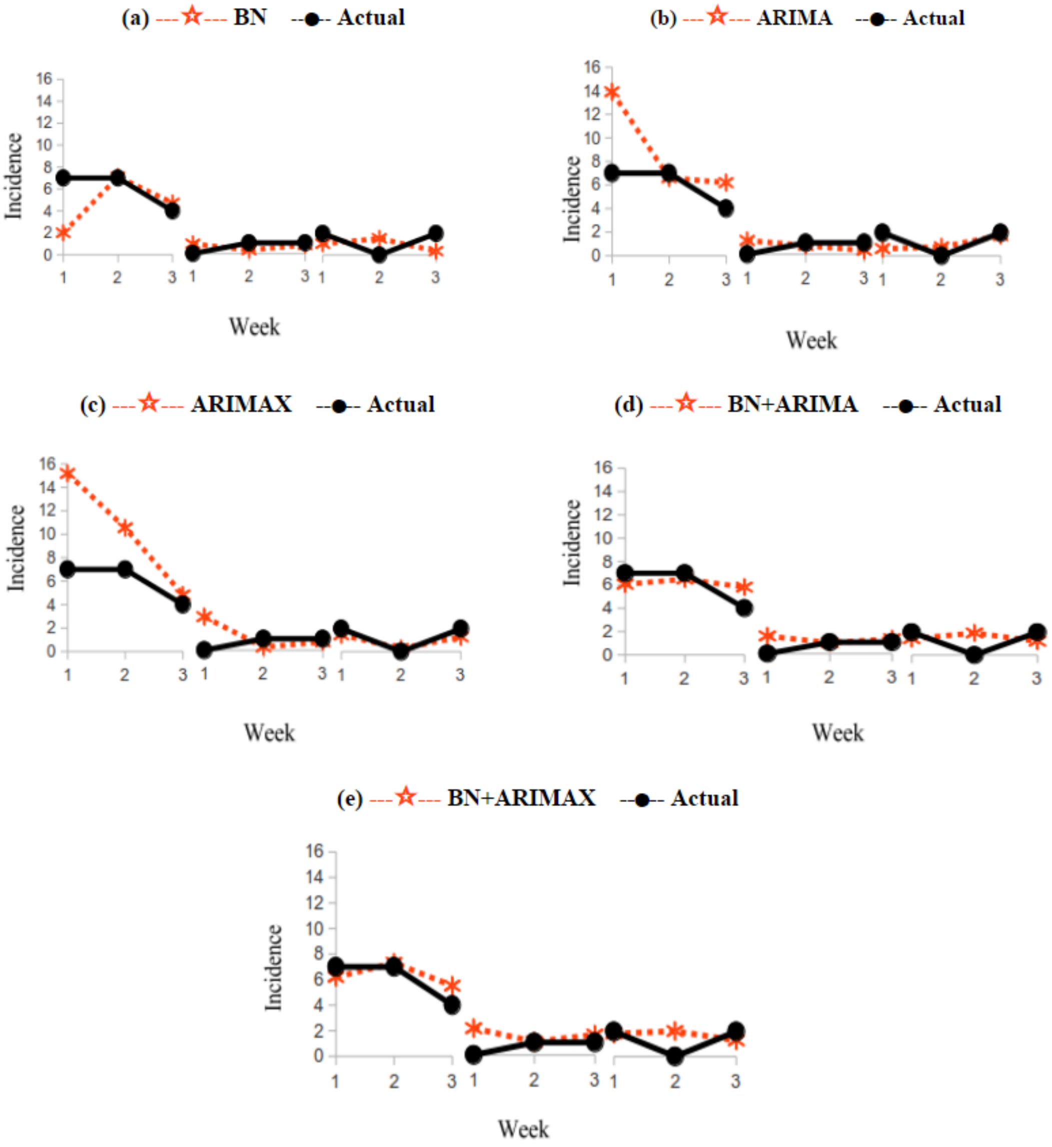

Figure 2. One week ahead malaria prediction versus actual for BN(a), ARIMA(b), ARIMAX(c), BN+ARIMA(d) and BN+ARIMAX(e) over a period of 3 weeks of 3 different months of a year for Village-1.

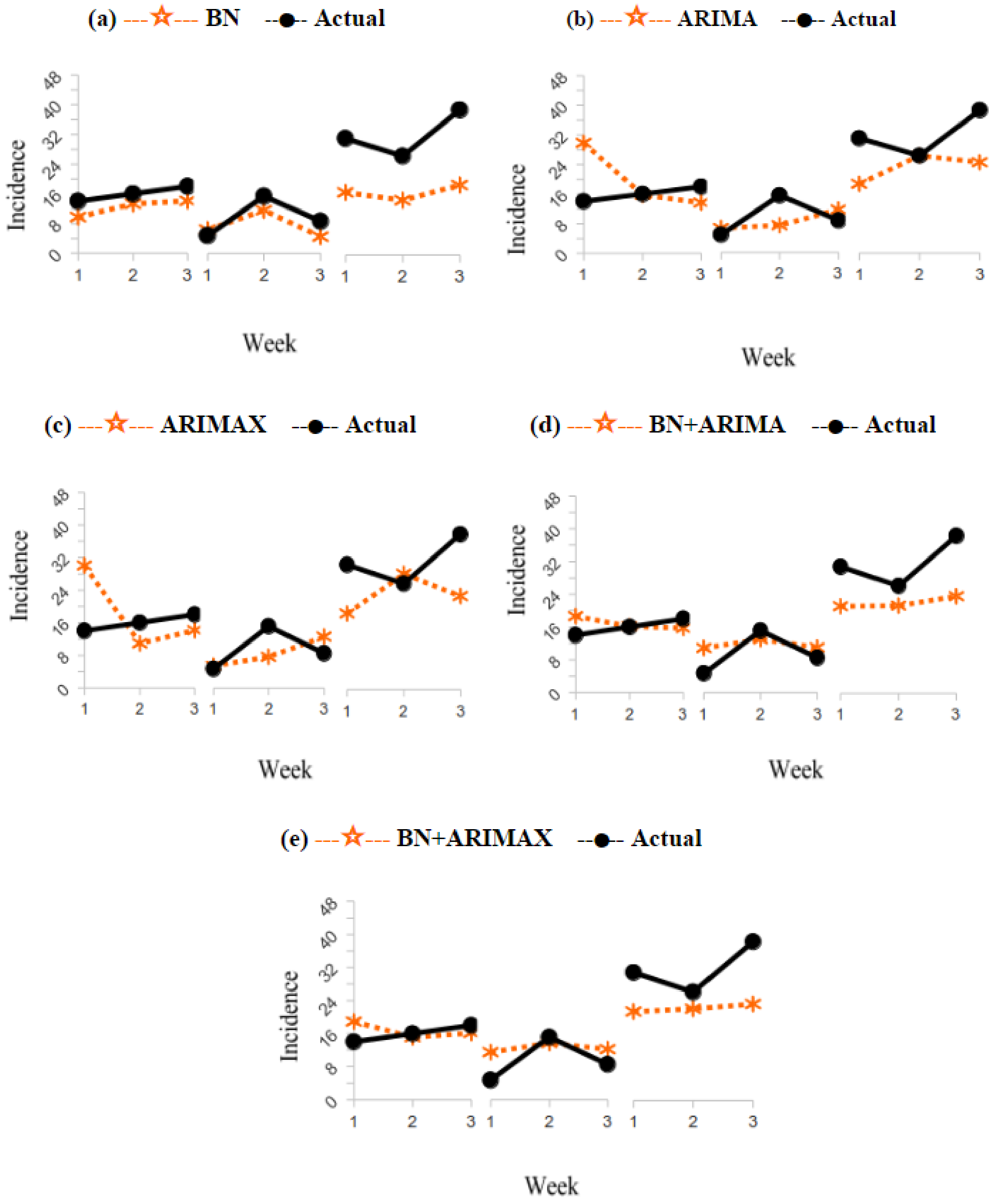

Figure 3. Two week ahead malaria prediction versus actual for BN(a), ARIMA(b), ARIMAX(c), BN+ARIMA(d) and BN+ARIMAX(e) over a period of 3 weeks of 3 different months of a year for Village-2.

months well but underestimates all three weeks of the last month. The ARIMA and ARIMAX models (Figure 3(b,c)) both overestimate the first week of the first month and underestimate weeks 1 and 3 of the last month. Again BN+ARIMA and BN+ARIMAX do better than all three individual models by correcting for the largest errors in the other models.

Table 2 shows the percentage improvement of performance of the combined BN+ARIMA model compared to BN, ARIMA, and ARIMAX for high, medium, and low incidence villages, as well as all 66 villages overall. Statistical significance was evaluated using a 2-talied t-test. All values are statistically significant ($p < 0.05$) except where indicated. The combined model outperforms the BN model with the difference statistically significant except for one week prediction for high incidence villages. The model performs worse than ARIMA and ARIMAX alone for one and two-week prediction for the low incidence villages. For all 66 villages BN+ARIMA outperforms the single models in all cases with the differences statistically significant except for ARIMAX W_2.

Table 3 shows the percentage improvement of performance of the combined BN+ARIMAX model compared to BN, ARIMA, and ARIMAX. This ensemble model now significantly outperforms the other three models in all cases except for three where the difference is not statistically significant. In particular, the improvement over ARIMAX for one and two week prediction is not statistically significant. Comparing the ensemble with the BN and ARIMAX models we see the largest improvement over the BN model for low incidence villages and the largest improvement over the ARIMAX model for the high incidence villages, which is reflective of the complementary strengths of the two models. For all entries in the tables, the BN+ARIMAX model outperforms the BN+ARIMA model.

Table 2. Performance Improvement(%) of BN+ARIMA over BN, ARIMA and ARIMAX.

	Percentage Improvement					
Models	BN		ARIMA		ARIMAX	
Set of Villages	W_1	W_2	W_1	W_2	W_1	W_2
13-high	8.48*	15.53	15.58	11.13*	14.10*	13.65
13-med	11.40	18.96	16.29	8.61*	15.10	8.08*
14-low	41.49	49.67	-18.13	-9.95	-45.38	-42.33
66-all	12.30	20.89	12.61	7.89	10.34	10.39*

W_1 = First Week, W_2 = Second Week
All values are statistically significant ($p < 0.05$) except where indicated by *.

Table 3. Performance Improvement (%) of BN+ARIMAX over BN, ARIMA and ARIMAX.

	Percentage Improvement					
Models	BN		ARIMA		ARIMAX	
Subset of villages	W_1	W_2	W_1	W_2	W_1	W_2
13-high	13.14	16.70	19.87	12.36*	18.47	14.84
13-med	13.58	23.87	18.35	14.14	17.19	13.65
14-low	62.14	65.97	23.56	25.64	5.92*	3.74*
66-all	17.08	24.63	17.38	12.24	15.22	14.63

W_1 = First Week, W_2 = Second Week
All values are statistically significant ($p < 0.05$) except where indicated by *.

8 CONCLUSION

In this paper we have taken the first essential step towards a system to support targeted malaria intervention using the data in Thailands E-Malaria Information system by developing a high resolution prediction model. We developed a Bayesian network model that represents the effect of environmental variables and captures nonlinear effects. Comparison with traditional ARIMA models showed that the two types of models have complementary strengths. Leveraging these complementary strengths, we developed an ensemble predictor that has significantly better accuracy than either predictor alone. Our results were obtained for one district in northern Thailand. A next step will be to test the generality of the model by applying it to districts with varying endemicity and environmental characteristics.

The structure of our Bayes net model can be used to provide causal explanations but by creating an ensemble in the way we did, we lose some of this explanatory power. It would thus be of potential benefit to seek to integrate the ARIMA model directly into the Bayes net. This might be done by including a node for the ARIMA result and a node that computes the weighted average of the models.

We intend to use this model as part of a decision support tool for targeted intervention of malaria. We are currently working to integrate it with a GIS to facilitate interaction and more intelligibly display results. An additional long-range goal is to apply our techniques the modeling of dengue.

ACKNOWLEDGEMENTS

This paper is based upon work supported by the U.S. Army International Technology Center Pacific (ITC-PAC) under Contract No. FA5209-15-P-0183. This work was also partially supported through a fellowship from the Hanse-Wissenschaftskolleg Institute for Advanced Study, Delmenhorst, Germany to Haddawy for collaborative work with the University of Bremen. We thank Saranath Lawpoolsri for providing the malaria data, Patiwat Sa-angchai for providing the environmental data, and Ngamphol Soonthornworasiri for advice on the ARIMA model.

REFERENCES

[1] R. Adhikari, 'A neural network based linear ensemble framework for time series forecasting', *Journal of Neurocomputing*, **157**, 231–242, (2015).

[2] R. Adhikari and R.K. Agrawal, 'A novel weighted ensemble technique for time series forecasting', *Lecture Notes in Computer Science*, **7301**, 38–49, (2012).

[3] W. Anggraeni, R.A. Vinarti, and Y.D. Kurniawati, 'Performance comparisons between arima and arimax method in moslem kids clothes demand forecasting: Case study', *Procedia Computer Science*, **72**, 630–637, (2015).

[4] G.E.P Box and G.M. Jenkins, *Time series Analysis: Forecasting and Control*, Holden-Day, California, 3rd edn., 1970.

[5] G.F. Cooper, D.H. Dash, J.D. Levander, W. Wong, W.R. Hogan, and M.M. Wagner, 'Bayesian biosurveillance of disease outbreaks', *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence (UAI '04)*, 94–103, (2004).

[6] L.A. Diaz-Robles, J.C. Ortega, J.S. Fu, G.D. Reed, J.C. Chow, J.G. Watson, and J. Moncada-Herrera, 'A hybrid arima and artificial neural network model to forecast particulate matter in urban areas: The case of temuco,chile', *Atmospheric Environment*, **42**, 8331–8340, (2008).

[7] W.M. Dlamini, 'A bayesian belief network analysis of factors influencing wildfire occurrence in swaziland, environmental modelling & software', *Environmental Modelling & Software*, **25**, 199–208, (2010).

[8] A. Gomez-Elipe, A. Otero, M. van Herp, and A. Aguirre-Jamie, 'Forecasting malaria incidence based on monthly case reports and environmental factors in karuzi, burundi,1977-2003', *Malaria Journal*, **6**, (2007).

[9] K.W. Hipel and A.I. McLeod, *Time Series Modelling of Water Resources and Environmental Systems*, ELSEVIER, Amsterdam, 1994.

[10] R.J. Hyndman. Forecasting functions for time series and linear models. `http://CRAN.R-project.org/package=forecast`, 2014.

[11] S. Johnson, F. Fielding, G. Hamilton, and K. Mengersen, 'An integrated bayesian network approach to lyngbya majuscula bloom initiation', *Marine Environmental Research*, **69**, 27–37, (2010).

[12] S. Johnson, K. Mengersen, A. de Waal, K. Marnewick, D. Cillers, A.M. Houser, and L Boast, 'Modelling cheetah relocation success in southern africa using an iterative bayesian network development cycle', *Ecological Modeling*, **221**, 641–651, (2010).

[13] A. Khamsiriwatchara, P. Sudathip, Sawang, S. Vijakadge, Potithavoranan T, A. Sangvichean, W. Satimai, C. Delacollette, P. Singhasivanon, S. Lawpoolsri, and J. Kaewkungwal, '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).

[14] M. Khashei, M. Bijari, and G.A.R. Ardali, 'Hybridization of autoregressive integrated moving average(arima) with probabilistic neural networks(pnns)', *Journal of Computer & Industrial*, **63**, 37–45, (2012).

[15] C. Kongcharoen and T. Kruangpradit, 'autoregressive integrated moving average with explanatory variable(arimax) model for thailand export', *Internation Journal of forecasting*, (2013).

[16] M.A. Kulkarni, R.E. Desrochers, and J.T. Kerr, 'High resolution niche models of malaria vectors in northern tanzania: A new capacity to predict malaria risk?', *PLoS One*, **5**, (2010).

[17] L. Liu, R.S. Luan, F. Yin, X.P. Zhu, and Q. Lu, 'Predicting the incidence of hand, foot and mouth disease in sichuan province, china using the arima model', *Epidemiology and Infection*, **144**, 144–151, (2015).

[18] Thailand Ministry of Public Health. Report on weekly malaria situation. `http://www.thaivbd.org/n/home`, 2015.

[19] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan-Kaufmann, San Francisco, 1st edn., 1988.

[20] S. Promprou, M. Jaroensutasinee, and K. Jaroensutasinee, 'Forecasting dengue haemorrhagic fever cases in southern thailand using arima models', *Dengue Bulletin*, **30**, 99–106, (2006).

[21] A.M. Tompkins and V. Ermert, 'A regional-scale, high resolution dynamical malaria model that accounts for population density, climate and surface hydrology', *Malaria Journal*, **12**, (2013).

[22] K. Wangdi, P. Singhasivanon, T. Silawan, S. Lawpoolsri, N.J. White, and 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).

[23] World Health Organization (WHO). World malaria report 2015. `http://apps.who.int/iris/bitstream/10665/200018/1/9789241565158_eng.pdf?ua=1`, 2015.

[24] J.D. Wichard, 'Forecasting the nn5 time series with hybrid models', *International Journal of Forecasting*, **27**, 700–707, (2011).

[25] L. Wilkinson, Y.E. Chee, A.E. Nicholson, and P. Quintana-Ascencio, '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).

[26] I. H. Witten and E. Frank, *Data Mining, Practical Machine Learning Tools and Techniques*, Morgan-Kaufmann, San Francisco, 2nd edn., 2005.

[27] D.H. Wolpert, 'Stacked generalization', *Neural Networks*, **5**, 241–259, (1992).

[28] G.P. Zhang, 'Time series forecasting using a hybrid arima and neural network model', *Journal of Neurocomputing*, **50**, 159–175, (2003).

[29] Y. Zhang, P. Bi, and J. Hiller, 'Meteorological variables and malaria in a chinese temperate city: A twenty-year time-series data analysis', *Environment International*, **36**, (2010).

[30] K. Zinszer, A.D. Verma, K. Charland, T.F. Brewer, J.S. Brownstein, Z. Sun, and Buckeridge, 'A scoping review of malaria forecasting: past works and future directions', *BMJ Open*, **2**, (2012).