Application of Remote Sensing Data for Dengue Outbreak Estimation Using Bayesian Network

Chanintorn Ruangudomsakul, Apinya Duangsin, Kittisak Kerdprasop, and Nittaya Kerdprasop

Abstract—**Dengue is an epidemic that is a major endemic health problem found in many countries in tropical and worm area. Weather conditions are important factors directly influence the degree of dengue outbreak. Current practice in the dengue outbreak forecasting relies on the meteorological reports. This study shows an alternative way to estimate dengue outbreak level by using Bayesian network (BN). We employ the satellite based remote sensing data to generate the probability model that can estimate dengue outbreak level. We use publicly available satellite based remote sensing data from the NOAA STAR. The data consist of weekly SMN, SMT, VCI, VHI, TCI indexes as factors for estimating the dengue outbreak in the northeast region of Thailand. In this study, 3 BN models had been generated using expert knowledge, greedy thick thinning algorithm, and combination of expert and greedy thick thinning algorithm. All 3 models are validated with the 10-fold cross-validation and ROC Analysis. The experimental results on real data show that the model automatically generated by greedy tick tinning algorithm performs well on overall estimation of dengue outbreak levels. But for an abnormal situation that the outbreak level is significantly higher than usual, the BN model with combination of expert and greedy thick thinning algorithm perform the best in such situation.**

Index Terms—**Bayesian network, remote sensing data, dengue outbreak, epidemiology estimation model.**

I. Introduction

Dengue is an epidemic that is a major health problem and it is widely found in more than 110 countries along the tropical and sub-tropical areas [1]. Dengue outbreak in Thailand has been found more than 50 year and the first report about dengue outbreak began to appear in 1958 [2]. In the past 15 years, the outbreak of dengue is continuously rising. In 2015, there were 144,952 cases found in all regions of Thailand, accounting for 222.58 per 100,000 population [3]. Such high prevalence indicates very high incident rate of infection.

Factors that are important for the spread of dengue fever vary in each area. These factors include the immunity of the population, dengue virus serotype, population density, migration of population, climatic condition and global worming [4]. These factors also affect the change pattern of dengue epidemic. Because dengue is a disease that many factors involved, prevention and warning process are focus on analysis of data from many sources to produce forecasting model to forecast future dengue outbreak incident. Climate conditions are important factors directly related to the dengue outbreak [5]-[8]. In Thailand dengue prediction model was generate using univariate forecasting model such as ARIMA model [9]. Climatic factors such as rainfall, average temperature and humidity are sued to analyze their relation to the spread of dengue in human [5]-[8].

Many researchers present the idea to use remote sensing data, such as sensing index related to vegetation and climate observation, to replace local rainfall data from ground base station [10]-[13]. Satellite based remote sensing data have more advantage than local ground based data because of their cost effective, real-time access, and more coverage. In previous study of Kerdprasop and Kerdprasop [14], they found that vegetation and climate indexes from the NOAA Star can be used to estimate precipitation in the studied area.

In this work, we present probabilistic model using Bayesian network (BN) for predicting outbreak level of dengue. Our models are built from the satellite based remote sensing data as prediction factors. BNs use a combination of graph theory and probability theory based on Bayes' theorem [15]. BNs have been widely used in many fields, especially in medical and industrial [2], because of their simplicity and easy to understand. BNs in our work are generated using both expert knowledge and automatic construction by the inherent algorithm of the Bayesian network software.

II. Background

A. Dengue

Dengue is a mosquito-borne viral disease. Female mosquitoes transmit dengue virus. Dengue Hemorrhagic Fever (DHF) or severe dengue was first recognized in the 1950s during dengue epidemics in the Philippines and Thailand. Today dengue has rapidly spread in all regions of Asian and Latin American countries.

Aedes aegypti mosquito is the primary vector of dengue. Dengue transmits to human by bite of infectious female mosquitos. Virus takes 4-10 days for incubation in infectious female mosquitos after the bite of infectious dengue patients. After that, female mosquitoes are capable of transmitting virus to human for the rest of their lives [16].

In human, dengue virus has incubation period for 4–10 days after the bite from an infected mosquito. Dengue can cause high fever about 40^{o} C or 104^{o} F and has some severe symptoms such as headache, pain behind eye, pain of muscle and joint, vomiting, swollen glands, rash, and nausea. These symptoms may last for 2–7 day. Severe dengue is a potentially deadly complication due to plasma leaking, fluid

Manuscript received May 17, 2018; revised July 20, 2018. This work has been supported by grants from Thailand Toray Science Foundation, Suranaree University of Technology, and Sisaket Rajabhat University.

C. Ruangudomsakul is with the Department of Software Engineering, Faculty of Liberal Art and Science,Sisaket Rajabhat University. 319 Thai-panta Avenue, Muang District, Sisaket 33000, Thailand (e-mail: Chanintorn.r@sskru.ac.th).

A. Duangsin was with Epidemiology and Intelligence Group, Office of Disease Prevention and Control 10 Ubon Ratchathani 34000.

K. Kerdprasop and N. Kerdprasop are with the School of Computer Engineering, Suranaree University of Technology, Thailand.

doi: 10.18178/ijmlc.2018.8.4.718

accumulation, respiratory distress, severe bleeding, or organ impairment [17]. The important issue is that there is no specific treatment for dengue fever.

B. *Aedes aegypti* Life Cycle

The life cycle of *Aedes aegypti* divide into 4 periods: egg, larva, pupa, and adult. Life cycle of mosquitoes from egg to adult takes about 8-10 days [16]. Therefore, when calculating time from the start of *Aedes aegypt* life cycle until it turns into adult mosquito will take around 8-10 days. If takes into account the incubation time of dengue virus in infectious female mosquitos (4-10 days) and in human before the presence of symptoms (4-10 days), the overall incubation period is 16-30 days or 3–4 weeks.

C. Application of Remote Sensing for Rainfall Estimation

This study uses satellite based remote sensing data from the NOAA STAR [18]. NOAA is a satellite in Polar-Orbiting Operational Environmental Satellites (POES) project of the National Aeronautics and Space Administration's (NASA). It installs the Advanced Very High Resolution Radiometer (AVHRR) to detect humidity and heat radiated from the Earth's surface. In this study, five indexes from the NOAA have been used in our analysis to determine the rainfall and temperature of the dengue outbreak area in Thailand. These five indexes are SMN, SMT, VCI, TCI, and VHI.

Smoothed and normalized difference vegetation index (SMN), or Smoothed NDVI, is satellite index that can be used to evaluate the greenness of the vegetation. SMN values are in the range [+1.0, -1.0]. The SMN value at 0.1 or lower means rock, sandy or sandy substrates. SMN value around 0.2-0.5 can be interpreted as sparsely cultivated areas such as shrubs or farmland after harvest. High SMN value around 0.6-0.9 can be interpreted as forest areas or cultivated areas fully grown.

Smoothed brightness temperature index (SMT) is the brightness temperature (BT) with completely removed high frequency noise. This index can be used for the estimation of thermal condition. High SMT values indicate dryness of vegetation.

Vegetation condition index (VCI) is a proxy for moisture condition. VCI varies between 0 and 100 percent, which relates to the change of vegetation condition. VCI can be used to estimate moisture condition. The VCI < 40 means lack of moisture, whereas VCI > 60 means moisture is in good condition. VCI can calculated from NDVI as in (1).

$$VCI(\%) = \frac{(NDVI - NDVI_{min})}{(NDVI_{max} - NDVI_{min})} \times 100 \quad (1)$$

Temperature condition index (TCI) represents vegetarian condition that responses to temperature. TCI indicates stress of vegetation, which relates to temperature. TCI can be calculated from BT as in (2).

$$TCI(\%) = \frac{(BT_{max} - BT)}{(BT_{max} - BT_{min})} \quad (2)$$

Vegetation health index (VHI) is a combination of VCI and TCI. VHI is used to monitor the integrity of vegetation, humidity and heat dissipation. VHI indicates greenness of vegetation. Higher VHI implies greener vegetation.

From previous study of Kerdprasop and Kerdorasop [14], they found that SMN with 2-month lag and VHI with 3-month lag can be used to estimate range of precipitation in the current month [14]. In this work, we thus employ the same set of remote sensing indexes to forecast rainfall, which is one important factor for estimating dengue outbreak with Bayesian network.

III. Material and Method

A. Bayesian Network

Bayesian network (BN) is a graphical representation of the joint probability distributions over a set of random variables. BN consists of a series of nodes representing variables connected by arrows forming a graph that has no cycles. Each node of the network is associated with a set of probability tables [19]. Both the structure and the numerical parameters of BN can be learned entirely from data. The joint probability distribution formula in BN can be expressed as in (3).

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | X_{i+1}, \dots, X_n) \quad (3)$$

When considering using the local distribution defined by the BN, the joint probability distribution can be rewritten as in (4).

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Parent(X_i)) \quad (4)$$

While $P(X_1, \dots, X_n)$ refers to the probability of a specific combination of values $X_1, \dots, X_n$ from the set of variables $(X_1, \dots, X_n)$. $Parent(X_i)$ refers to closet parent nodes to the set of $X_i^{'}$. The conditional probability $P(X_i|Parent(X_i)$ is the probability relative to the node X_i based on its parent nodes.

There are large number of algorithms that can learn the structure and the parameters of BN from data. In this study, we use Greedy Thick Thinning algorithm [20], which can modify the structure and scoring the result from fully connected network by subsequently removing arcs between nodes based on conditional independences tests [20]. The Greedy Thick Thinning algorithm is able to identify the best scoring network.

To assess model performance, error rate and predictive value of the BN were estimated using a 10-fold cross validation procedure. In this work, BNs are implemented with GeNIe version 2.0 [20].

B. Dengue Surveillance Dataset

This study uses dengue surveillance dataset of Sisaket province collected by the 10th office of disease prevention and control, Thailand. The dataset contains list of confirmed dengue patients recorded from week 1 of 2007 until week 45 of 2015. All data have been clean and calculate incident rate (I.R.) per week using the following equation:

$$I.R. = \frac{n}{X} \times 10^5 \quad (5)$$

While n refer to new cases of dengue found in the region. X refers to population at risk in the same time period.

Because BN is suitable for categorical data. Numeric data are needed to be discretized. Dengue case data are discretized into 3 categories using the following criteria.

- *Low* is the incident rate at which the I.R. is lower than the mean, that is the range $(0 \; to \; \bar{x})$.
- *Intermediate* is the incident rate at which the I.R. is higher *than* the average up to 2 times of the standard deviation, that is the range $(\bar{x} \; to \; (\bar{x} + 2SD))$.
- *High* is the *incident* rate when I.R. is higher than the intermediate level.

The satellite index data from the NOAA project will also be divided into 5 ranges of uniform widths. Table I shows all discretization range of data.

TABLE I: DISCRETIZATION OF NUMERIC DATA INTO RANGE VALUES

Variable	Category	Range of values
CASE	Low	< 4.093
	Intermediate	4.093- 14.695
	High	> 14.695
SMN	Low	< 0.695
	Lower Intermediate	0.2744-0.25638
	Intermediate	0.25638 – 0.30532
	Upper Intermediate	0.30532 – 0.35426
	High	> 0.35426
SMT	Low	< 282.345
	Lower Intermediate	282.345 – 288.037
	Intermediate	288.037 – 293.729
	Upper Intermediate	293.729 – 299.421
	High	> 299.421
VCI	Low	< 40.556
	Lower Intermediate	40.556 – 53.902
	Intermediate	53.902 – 67.248
	Upper Intermediate	67.248 – 80.594
	High	> 80.594
VHI	Low	< 33.742
	Lower Intermediate	33.742 – 45.404
	Intermediate	45.404 – 57.066
	Upper Intermediate	57.066 – 68.728
	High	> 68.728
TCI	Low	< 23.228
	Lower Intermediate	23.228 – 40.676
	Intermediate	40.676 – 58.124
	Upper Intermediate	58.124 – 75.572
	High	> 75.572

C. *Lagged Correlation Analysis of Dengue Incidence*

Cross correlation function has been used to evaluate the appropriate lagged time between satellite remote sensing indexes and the dengue I.R. Table II indicates most appropriate lagged time between dengue I.R. and satellite based indexes.

TABLE II: APPROPRIATE LAGGED TIME OF SATELLITE BASED INDEXES

index	SMN	SMT	VCI	VHI	TCI
Lag(week)	-13	-13	-30	-30	-1

IV. RESULT

A. *BN Construction and Result*

In this work, the BNs were created in three ways: (1) modeling with the expert knowledge, (2) modeling with expert knowledge and some guidance form automatic method using the greedy thick thinning algorithm, and (3) modeling with automatic greedy thick thinning algorithm.

Fig. 1 shows BN model created from expert knowledge. The most important factor to the dengue outbreak forecasting is the number of patients who are host of dengue virus. The model depends on the rate of I.R in the past 4-8 weeks.

Fig. 2 shows a BN model firstly created by expert knowledge and then modified by an automatic algorithm. In this model, we can see that satellite based remote sensing data with appropriate lagged time were taken into account for forecasting dengue outbreak.

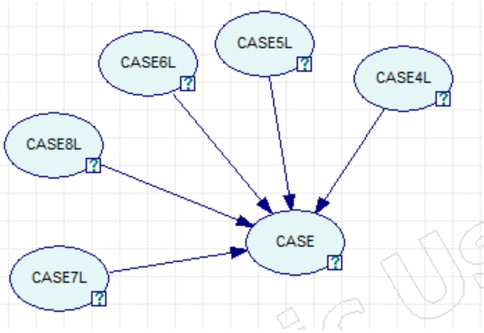

Fig. 1. BN modeling with the expert knowledge.

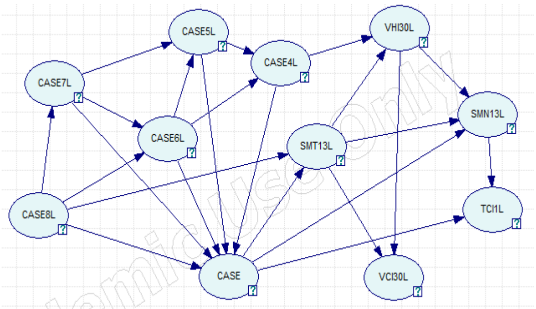

Fig. 2. BN modeling with expert knowledge and guidance from automatic method.

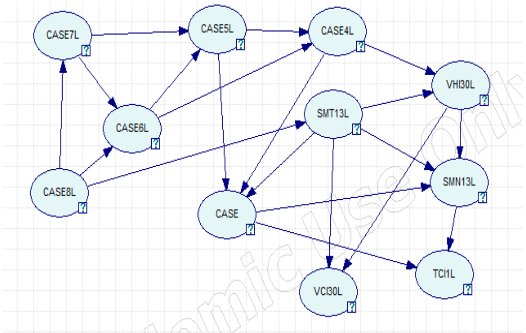

Fig. 3. Automatic BN modeling using greedy thick thinning algorithm.

Fig. 3 shows BN model automatically generated by the algorithm. In this model, the greedy thick thinning (GTT) algorithm learns the model based on the input data alone. It can be seen that the generated model is quite similar to Fig. 2, but the connection is less complicated.

To validate the above 3 models, we perform 10-fold cross-validation to assess two aspect: (1) which model is the most accurate one to predict range of dengue I.R., and (2) which model can high incident events that indicate risk of dengue outbreak.

The result of first aspect is shown in Fig. 4. The highest gray line is the accuracy of the model 3, which is the one automatically generated by the GTT algorithm. Average accuracy of this model form running 10-fold cross-validation is 0.906. The next best one is model 2 (expert + GTT) that has average accuracy at 0.889. The worst is model 1 (expert) with average accuracy at 0.846.

The result of the second aspect that emphasizes on high incidence of dengue cases is shown in Fig. 5. For this kind of event, model 2 (expert + GTT) can accurately predict the event with average accuracy at 0.705. The second accurate model is the model 3 (GTT) with the average accuracy at 0.621. The worst one is model 1 (expert) with the accuracy rate as low as 0.421.

B. ROC Analysis

Receiver operating characteristic (ROC) is a metric commonly used to measure the performance of binary classification. ROC plots a sensitivity (true-positive rate) by false positive rate (1-specificity) in every possible range. ROC can represent the overall efficiency of the model. Calculation of area under curve (AUC) from ROC plot can measure performance of the model. AUC = 1 indicates the best performance, whereas AUC below 0.5 indicating that the model performs no better than random guess.

From the ROC analysis of the three models, we found that model 3 had the highest overall efficiency with an average AUC of 0.954, followed by model 2 with an average AUC of 0.951, and model 1 had the lowest overall efficiency with average AUC=0.870. Models 2 and 3 show very little deference in their overall efficiency.

Considering only the high level of outbreak (Fig. 6), we found that model 2 (AUC=0.962) was better than model 3 (AUC=0.958).

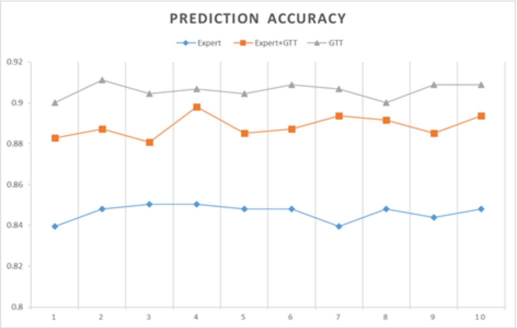

Fig. 4. Overall accuracy of running 10-fold cross-validation.

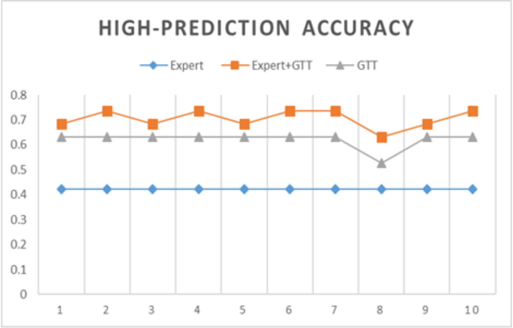

Fig. 5. Prediction accuracy at high occurrence of dengue cases.

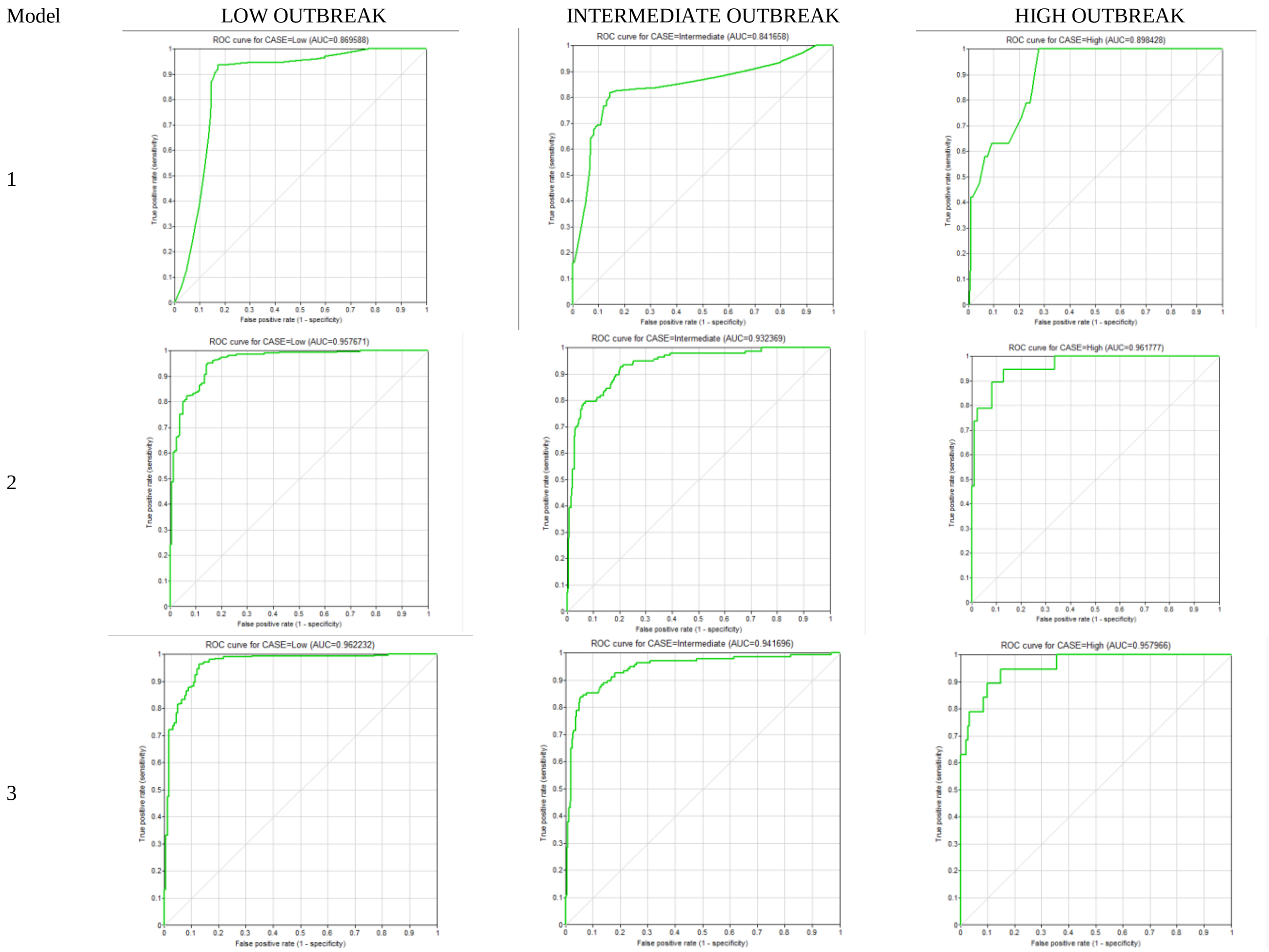

Fig. 6. ROC curves present AUC for all Bayesian network models (low outbreak, intermediate outbreak, high outbreak). Model 1 BN modeling with the expert knowledge (AUC = 0.870, 0.842, 0.898), Model 2 BN modeling with expert knowledge and guidance from automatic method (AUC = 0.958, 0.932, 0.962), Model 3 Automatic BN modeling using greedy thick thinning algorithm (AUC = 0.962, 0.942, 0.958)

V. CONCLUSION

Dengue is a world-wide epidemic health problem found in more than 110 countries in tropical and sub-tropical areas. In this study, we present probabilistic model based on Bayesian network for predicting dengue outbreak levels using remote sensing data from the NOAA project as prediction factors.

On generating the BN model, we consider the association of dengue outbreaks with change of vegetation and climate. By using satellite based remote-sensing indices, we firstly analyze correlation between dengue incident rate and all satellite indexes to identify most appropriate lagging time. The study found that SMN, SMT, and TCI with the lag time 8-9 weeks were relative to dengue incident rate in the study area. After that the data with lag-time were use as factor for analyzing dengue outbreak with three different BN models: the one built from expert (named model 1), network built from expert plus modification by the greedy thick thinning (GTT) algorithm (named model 2), and the network built from only the GTT without any intervention from the expert (named model 3).

We found from the experimental results that generally model 1 was the most accurate model with average accuracy 0.906 and average AUC 0.954. For the rare situation that dengue cases occur significantly high than usual, we found that the BN model which was generate by using combination of expert knowledge and GTT algorithm yielded the best predictive result with average accuracy 0.705 and average AUC 0.962.

When can conclude from our findings that the BN model for estimating dengue outbreak level built from the combination of expert knowledge and GTT algorithm perform well in terms of dengue forecasting at high outbreak level and perform acceptably on predicting cases at overall different outbreak levels.

REFERENCES

[1] S. Ranjit and N. Kissoon, “Dengue hemorrhagic fever and shock syndromes,” *Pediatric Critical Care Medicine*, vol. 12, no. 1, pp. 90-100, 2011.

[2] Bureau of Epidemiology, Department of Disease Control. (2017). [Online]. Available: http://www.boe.moph.go.th/fact/Dengue_Haemorrhagic_Fever.htm

[3] National Trustworthy and Competent Authority in Epidemiological Surveillance and Investigation. (2017). [Online]. Available: http://www.boe.moph.go.th/boedb/surdata/index.php

[4] J. Farrar and J. Whitehorn, “Dengue,” *Br Med Bull*, vol. 95, pp. 161-173, 2010.

[5] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, and J. Rocklöv, “Forecast of dengue incidence using temperature and rainfall,” *PLoS Neglected Tropical Diseases*, vol. 6, no. 11, 2012.

[6] A. L. Buczak, P. T. Koshute, S. M. Babin, B. H. Feighner, and S. H. Lewis, “A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data,” *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, 2012.

[7] N. C. Dom, A. A. Hassan, Z. A. Latif, & R. Ismail, 2013, “Generating temporal model using climate variables for the prediction of dengue cases in Subang Jaya, Malaysia,” *Asian Pacific J of Tropical Disease*, vol. 3, no. 5, 352-361.

[8] H. G. Gu *et al*., “Meteorological factors for dengue fever control and prevention in South China,” *Int J of Environmental Research and Public Health*, vol. 13, no. 9, p. 867, 2016.

[9] S. Wongkoon, M. Jaroensutasinee, and K. Jaroensutasinee, “Development of temporal modeling for prediction of dengue infection in Northeastern Thailand,” *Asian Pacific J of Tropical Medicine*, vol. 5, no. 3, pp. 249-252, 2012.

[10] V. Boken, G. Hoogenboom, F. Kogan, J. Hook, D. Thomas, and K. Harrison, “Potential of using NOAA-AVHRR data for estimating irrigated area to help solve an inter-state water dispute,” *Int J Remote Sens*, vol. 25, no. 12, pp. 2277–2286, 2004.

[11] M. Jalili, J. Gharibshah, S. Ghavami, M. Beheshtifar, R. Farshi, “Nationwide prediction ofdrought conditions in Iran based on remote sensing data,” *IEEE Trans Comput*, vol. 63, no. 1, pp. 90–101, 2014.

[12] A. Karnieli, N. Agam, R. Pinker *et al*., “Use of NDVI and land surface temperature for drought assessment: merits and limitations,” *J Clim*., vol. 23, pp. 618–633, 2010.

[13] S. Quiring and S. Ganesh, “Evaluating the utility of the Vegetation Condition Index (VCI) for monitoring meteorological drought in Texas,” *Agric for Meteorol*, vol. 150, pp. 330–339, 2010.

[14] K. Kerdprasop and N. Kerdprasop, “Remote sensing based model induction for drought monitoring and rainfall estimation,” in *Proc. Int Conf on Computational Science and Its Applications*, pp. 356-368, July 2016.

[15] T. D. Nielsen, & F. V. Jensen, 2009, *Bayesian Networks and Decision Graphs*, Springer Science & Business Media.

[16] Centers for Disease Control and Prevention, *Mosquitoes’ Life Cycle*, National center for emerging and zoonotic infectious diseases, 2014.

[17] WHO, D, *Dengue Haemorrhagic Fever*, Fact Sheet No. 117, 2009.

[18] NOAA STAR Center for Satellite Applications and Research. STAR-Global Vegetation Health Products, U.S.A. (2015). [Online]. Available: http://www.star.nesdis.noaa.gov/smcd/emb/vci/VH

[19] F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer, 2001.

[20] Decision Systems Laboratory. GeNIe 2.0. (2006). [Online]. Available: http://www.sis.pitt.edu/~genie/

Chanintorn Ruangudomsakul is a lecturer in School of Software Engineering, Faculty of Faculty of Liberal Art and Science, Sisaket Rajabhat University. He is a member of Knowledge Engineering Research. His current research is advanced development in the multidisciplinary field of machine learning techniques and computational epidemiology.

Apinya Duangsin is currently a professional in Public Health of Epidemiology and Intelligence Group, Office of Disease Prevention and Control 10, Ubon Ratchathani province, Thailand. She received her master degree in Public Health from Khon Kaen University in 2016. Her current research of interest is on the field of Epidemiology.

Kittisak Kerdprasop is an associate professor at the School of Computer Engineering, Chair of the School, and the head of Knowledge Engineering Research Unit, SUT. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, MS in computer science from the Prince of Songkla University, Thailand, in 1991 and PhD in computer science from Nova Southeastern University, U.S.A., in 1999. His current research includes machine learning and artificial intelligence

Nittaya Kerdprasop is an associate professor and the head of Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. She received her B.S. in radiation techniques from Mahidol University, Thailand, in 1985, M.S. in computer science from the Prince of Songkla University, Thailand, in 1991 and Ph.D. in computer science from Nova Southeastern University, U.S.A., in 1999. Her research of interest includes data mining, artificial intelligence, logic and constraint programming.