science, engineering and health studies

Research Article

Science, Engineering and Health Studies
https://lio1.tci-thaijo.org/index.php/sehs
ISSN (Online): 2630-0087

Time series data enrichment using semantic information for dengue incidence forecasting

Wanarat Juraphanthong[1] *and Kraisak Kesorn*[2*]

[1] Department of Computer Engineering, Faculty of Industrial Technology, Pibulsongkram Rajabhat University, Phitsanulok 65000, Thailand
[2] Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand

ABSTRACT

Forecasting the incidence of dengue diseases as time series models facilitates public health anticipation and preparation for managing an outbreak and reducing morbidity. Previous works have indicated that many potential predictors are significant factors for improving the accuracy and effectiveness of the prediction. However, these factors are usually used as dependent variables and are rarely used to identify and utilize data relationships in time series approaches. Therefore, the purpose of this study was to enrich time series data with semantic information and knowledge from a dengue fever ontology model, in order to improve the capability of time series methods to forecast dengue incidence in the provinces of Thailand. In this paper, a new technique, named auto regressive integrated moving average (ARIMA) with semantic data (ARIMAS) was introduced and compared with classical time series approaches such as ARIMA and ARIMAX. The root mean squared error (RMSE) of ARIMA and ARIMAX was 25.97 and 27.45, respectively, whereas that of ARIMAS was 24.29. The results showed that the predicted values of ARIMAS were closer to the observed data than the values obtained from traditional time series techniques. In addition, the forecast performance of unusual periods with fluctuant incidence improved significantly. In 2013 and 2015, the RMSE of ARIMAS was 39.78 and 48.09, respectively, whereas ARIMA had an RMSE of 61.32 and 71.66, respectively, both years witnessed a large epidemic of dengue fever and have been explored in previous studies.

Keywords: time series; ontologies; semantic information; dengue; forecast; ARIMAS

***Corresponding author**:
Kraisak Kesorn
kraisakk@nu.ac.th

***Received**: 2 February 2021*
***Revised**: 16 February 2021*
***Accepted**: 19 February 2021*
***Published**: 14 December 2021*

***Citation**:
Juraphanthong, W., and Kesorn, K. (2021). Time series data enrichment using semantic information for dengue incidence forecasting. Science, Engineering and Health Studies, 15, 21050013.

1. INTRODUCTION

Dengue fever is a well-known and deadly tropical disease, the incidence of which increased from 2.4 million cases to over 4.2 million cases, from 2010 to 2019 (World Health Organization, 2020). This long-term, homogeneous data can be used to forecast future outbreaks, and may provide public health authorities with sufficient advanced warning to allow an early response and planning that would decrease morbidity and death rates.

Time series analysis is a statistical technique that is used in many studies of dengue diseases. This has enabled researchers to create dengue forecasting models. Time series models, especially auto regressive integrated moving averages (ARIMA), are popular method used in several previous works. Cortes et al. (2018) used ARIMA analysis to

forecast dengue incidence in two Brazilian cities while Nayak and Narayan (2019) deployed ARIMA to predict outbreaks in India. Somboonsak (2019) used ARIMA and Gaussian distribution to predict the incidence of dengue fever in northeast Thailand. Recently, Polwiang (2020) employed ARIMA to extracted seasonal patterns of dengue cases in Thailand.

However, these methods have a weakness in that they consider only the dengue-specific data such as a number of cases but did not include other, potentially useful, dependent factors in their forecasting processes. Hence, the studies are inconsistent with many other studies that have found that influencers of outbreaks are usually derived from various dependent factors. ARIMAX, an extension of ARIMA, is an alternative approach to integrating external time series variables that can assist the forecasting model for better time series fitting. Many researchers have used ARIMAX in dengue fever predictions. Anggraeni and Aristiani (2016) appended Google trends into ARIMAX in order to forecast the number of outbreaks. Jing et al. (2018) incorporated external regressors such as imported cases, mosquito densities, and temperature into ARIMAX to predict dengue transmission in China. Nguyen (2018) predicted dengue spread using ARIMA and ARIMAX with climate variables based on the datasets of two cities. Thiruchelvam et al. (2018) investigated relationships between dengue cases and air quality using ARIMAX. Typically, the dependent factors of ARIMAX are added to each item of continuous data to assist the forecast process. The main limitation of ARIMAX is that it ignores semantic information that exists in the data, which might help the time series model provid a better forecast. In the past decade, there are a few studies showing that these factors contain useful semantic information identifying relationships in the time series data. As such, our hypothesis is that using a knowledge base (an ontology model) containing semantic information can enhance the performance and effectiveness of time series analysis models.

An ontology is a framework that represents the semantic description of information using concepts and relationship abstractions. This can provide the knowledge to infer the relevance of associations between information concepts. Dengue fever ontologies have been constructed for many aspects of the biomedical field. For example, Herdiani et al. (2012) proposed the dengue hemorrhagic fever ontology (DHFO), which contains general and epidemiological information that can help in the formulation of control policy initiatives. Similarly, Mitraka et al. (2015) constructed IDODEN, an ontology for dengue fever, which covers all the important information of dengue fever, to assist researchers and medical personnels and identify new developments that may arise. In this research, IDODEN and DHFO were modified by adding information from our dataset.

Climate is one of the factors that are directly correlated with the dengue incidence cases (Siriyasatien et al., 2016). Climatic conditions and seasonal weather affect the mosquito population and transmission of the disease (Nagao et al., 2003). In previous work, this factor was often used as a predictor variable (Lu et al., 2009; Polwiang, 2020), but its semantic relationships with other factors were usually ignored. Since we used the knowledge obtained from the ontology, the semantic relationship between dengue fever time series data was considered to be useful to filter out irrelevant data, leading to improvements in the quality of the data and, thus, the forecasting power of ARIMA.

The aim of this research was to integrate an ontology model with time series data analysis (i.e., ARIMAS) to improve the quality of the data input into the time series analysis.

2. MATERIALS AND METHODS

2.1 Simplified dengue ontology

We built a simplified schema of a dengue ontology, necessary because existing ontologies only consider information from the biomedical domain such as disease data, host, diagnosis, therapy, and prevention. The design of our ontology is based on the DHFO (Herdiani et al., 2012) and IDODEN (Mitraka et al., 2015) models, and combines simple biomedical concepts with other crucial concepts necessary for the modeling process such as direct and indirect factors. The important parts of the simplified dengue ontology are shown in Figure 1.

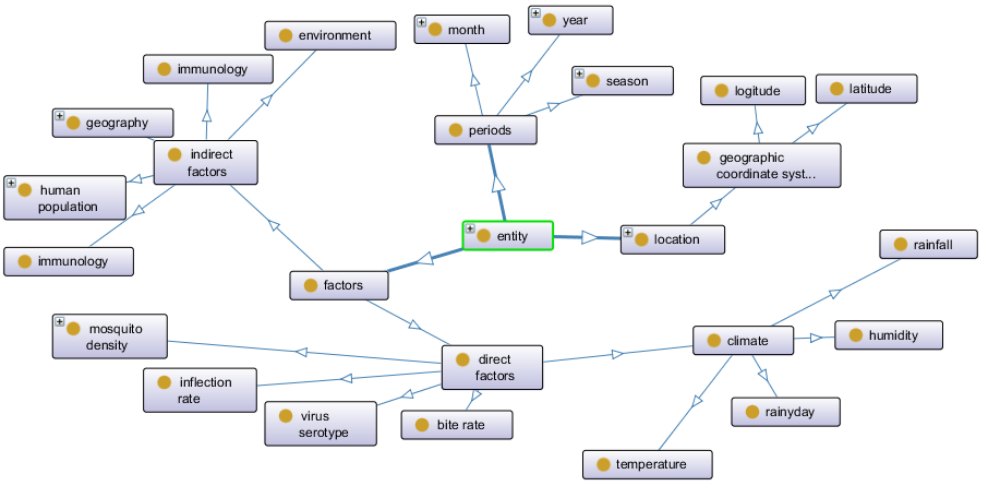

Figure 1. A simplified overview of the important parts of the dengue ontology

2.2 Data and study areas

Climate is one of the major factors associated with dengue incidence in Thailand (Johansson et al., 2009). To infer this knowledge, we have used aggregated monthly dengue cases (dengue fever and dengue hemorrhagic fever) and climate data of 76 provinces in Thailand over the period 2003 to 2017 as the source of the enrichment process. However, at this stage of implementation, our research was focused only on Phitsanulok; it would be extended to other provinces in the future. Phitsanulok is a province in the central region of Thailand and was selected because the observed incidence of dengue fever during specific times was favorable to our proposed approach. The dengue-related data were retrieved from the Health Data Center of the Thai Ministry of Public Health, and the meteorological data for the same period were collected from the Thai Meteorological Department.

2.3 Time series models

To model the dengue fever time series data, we worked mostly with time series models, including ARIMA and its extension ARIMAX.

An ARIMA model is a classical approach that produces forecasts based on historical data. This model is denoted as ARIMA *(p, d, q)*, with three combination components: the autoregressive (AR) component *p*, which is the number of autoregressive lags; the integration (I) component *d*, which is the number of differences required to obtain stationarity; and the moving average (MA) component *q*, which is the number of moving averages lags. Following the ARIMA model, the observed value of time series y_t is given by Equation 1:

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d y_t = \left(1+\sum_{j=1}^{q}\theta_j L^j\right)\varepsilon_t \quad (1)$$

with lag operator L, self-regression parameter φ_i, moving average parameter θ_j, and random error ε_t.

An ARIMAX model is an extended version of ARIMA that can add explanatory time series data as predictor variables. Therefore, the ARIMAX *(p, d, q)* model with the exogenous variable x_t is represented by Equation 2:

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d y_t = \Theta(L)x_t + \left(1+\sum_{j=1}^{q}\theta_j L^j\right)\varepsilon_t \quad (2)$$

2.4 Methodology

To develop the model for forecasting dengue incidence, we proposed five main processes, as shown in Figure 2:

2.4.1 Semantic time series enrichment

In this step, knowledge was enhanced to create a normalized time series data based on the simplified dengue ontology. The meta-information corresponding to the ontology was added to the time series by the semantic processor, as shown in Figure 3. An example of a semantic time series is shown in Table 1.

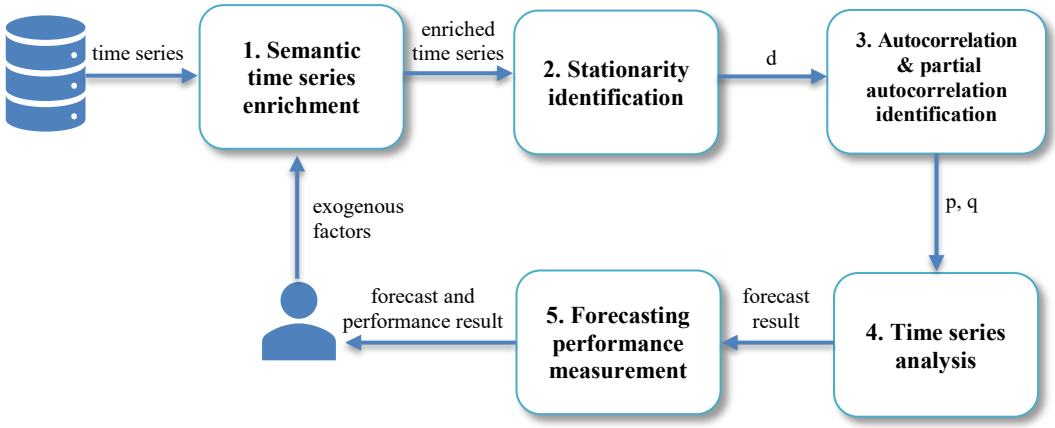

Figure 2. The framework of the main processes

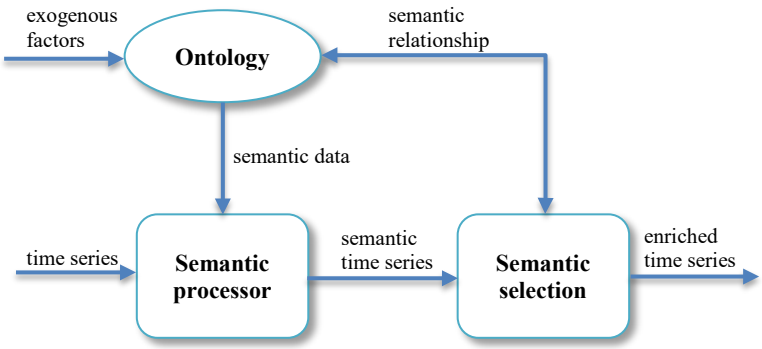

Figure 3. Semantic time series enrichment process

Table 1. Example for semantic time series of Phitsanulok, Thailand

Date	Number of incidences	Concept of semantic time series (C_{TS})		
		Normalized rainfall	Latitude	xxx
01/01/2003	7	0.0022	0.1131	...
01/02/2003	1	0.0097	0.1131	...
...	...	...	...	...
01/12/2017	14	0.0000	0.1131	...

Using the knowledge that a user provided, the semantic selection process focused on semantic selection and filtering of the related items in the classical time series data. The semantic relationships between data were measured for the semantic similarity using a combination of ontology-based and distance-based approaches. In the ontology-based method, the idea of measurements from Leacock and Chodorow (1998) that relate the path between concepts and the depth of the ontology was used. In the distance-based method, the traditional measure of Euclidean distance was used.

For a better understanding of Algorithm 1, we provide an example scenario in which we seek to enrich the data in the 1st record of Table 1. TS_{loc} is the semantic time series of Phitsanulok and the TS_{all} is the semantic time series of the other 75 provinces of Thailand included in the analysis. The variable x in Algorithm 1 is defined as a set of exogenous factors from a user. In this approach, the semantic weight between the exogenous factors and the semantic time series was computed using the path of their least common subsumer (LCS) with the concept of exogenous factors (C_x) and the concept of semantic time series (C_{TS}) in the dengue ontology, as shown in Equation 3.

$$Weight(C_X, C_{TS}) = \frac{path(LCS, C_X) + path(LCS, C_{TS})}{2 \times depth} \quad (3)$$

where *path(x,y)* indicates the distance between a concept x and a concept y in the ontology model. *Depth* refers as the height of the ontology model.

If X is a "climate" factor specified by a user (C_x = climate), the semantic time series in Table 1 is "rainfall" and "latitude" and C_{TS} = rainfall, latitude. Based on Figure 1, we can obtain LCS (climate, rainfall) = climate and LCS (climate, latitude) = entity. Consequently, we computed the semantic weight and obtain *Weight* ($C_{climate}$, $C_{rainfall}$) = 0.1 and *Weight* ($C_{climate}$, $C_{latitude}$) = 0.6 where the ontology depth = 5. Thereafter, the distance between each row between TS_{loc} and TS_{all} was computed using the Euclidean distance (Dis_i). Then, all distances were used to find the semantic similarity per Equation 4:

$$Sim(List_k, List_l) = \sum_{i=1}^{m} Dis_i \times Weight_i \quad (4)$$

where $List_k$, $List_l$ are the rows of TS_{loc} and TS_{all}, respectively.

Finally, the row of TS_{all} that has a semantic similarity value less than the threshold was selected and inserted to TS_{loc} ordered by time. The threshold was determined by the characteristic of the value in the semantic weight and the distance, and then the optimal threshold for the enriched time series data was selected.

Algorithm 1: Semantic selection algorithm
1: Input : TS_{all} , TS_{loc} , $\{X\}$
2: Compute Eq.(2)

$$Weight(C_X, C_{TS}) = \frac{path(LCS, C_X) + path(LCS, C_{TS})}{2 \times depth}$$

3: For each $List_k$ of TS_{loc}
4: For each $List_l$ of TS_{all}
5: Compute $Dis_i = (List_k - List_l)^2$
6: Compute

$$Sim(List_k, List_l) = \sum_{i=1}^{m} Dis_i \times Weight_i$$

7: If *Sim* ($List_k$, $List_l$) < threshold:
8: add $List_l$ to TS_{loc} order by Time
9: Output : enriched TS_{loc}

2.4.2 Time series stationarity identification

The stationarity of time series data is an important property offering statistical equilibrium of variables. In this step, the stationarity of the time series data was identified using the augmented Dickey-Fuller (ADF) test (Dickey and Fuller, 1979). The unsteady variance of data was considered to be removed by differencing and repeating the test to ensure that the test was not at a root unit. Then, a suitable order of difference d to both the ARIMA and the ARIMAX models was employed. Examples of non-stationarity and stationarity are shown in Figures 4(a) and (b), respectively. The traditional time series were not stationary because there were trends that affected the value of the time series at different times. Hence, the trends were eliminated using their first order difference to identify the stationarity.

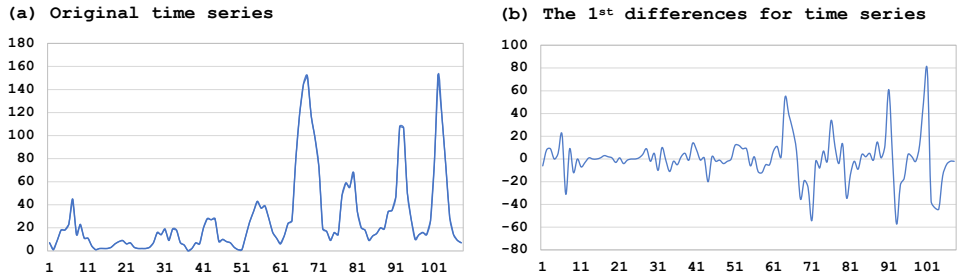

Figure 4. Line plot of original and difference dengue incidence data of first order (d=1)

2.4.3 Time series autocorrelations and partial autocorrelations identification

To select the optimal value of q and p for each model, the autocorrelation function (ACF) and partial autocorrelations function (PACF) (Box et al., 2015) were used to identify the possible order of MA and AR. The plots of the autocorrelation and partial autocorrelation for orders 1 to 20 of dengue fever cases in Phitsanulok are shown in Figures 5 and 6. Interpreting the correlogram, the values of autocorrelation at lags 1 and 2 were over the significance limits and were likely to reduce to zero after lag 2. The values at lags 10 and 11 have reached the upper limit, but the remainder values were well within the limits. The partial correlogram was also interpreted to mean that the values of the partial autocorrelation was over the significance limits at both lags 1 and 2 and were likely to be reduced to zero after lag 2. Since both values reduced to zero after lag 2, it is assumed that the possible ARMA model for this time series data was the possible order of autoregressive p and moving average q are 2.

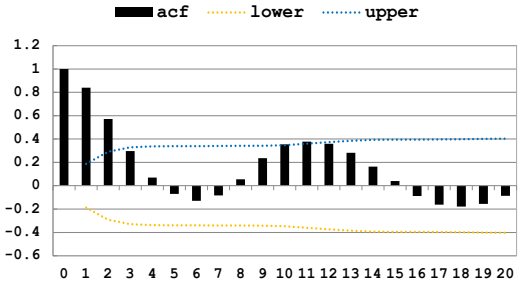

Figure 5. Autocorrelations (ACF) of dengue incidence in Phitsanulok, Thailand

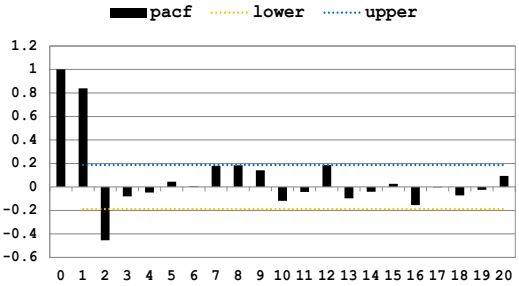

Figure 6. Partial autocorrelations (PACF) of dengue incidence in Phitsanulok, Thailand

2.4.4 Time series analysis

In this analysis process, ARIMA and ARIMAX models were used to forecast the dengue incidence in the study area. The ARIMA model used both an original dataset and an enriched dataset from the semantic time series processing. ARIMAX was used for the classical time series data with exogenous variables that were the same as the semantic variables in the semantic processor. In each experiment, the most suitable variable was selected from the candidate models. These candidate models were generated from a suitable order of difference d, a possible order of autoregression p, and a moving average q in previous processes. Then, the fitted model with the lowest Bayesian information criterion (BIC) and Akaike information criterion (AIC) values was selected. For example, candidate ARIMA models of classical time series of dengue fever cases in Phitsanulok are shown in Table 2. The ARIMA (2,1,0) model with (p=2, d=1 and q=0) was the best predictive model.

Table 2. AIC and BIC values of candidate ARIMA models for classical time series of dengue incidences in Phitsanulok, Thailand

Candidate model	AIC	BIC
ARIMA (1,1,0)	458.0282	462.4964
ARIMA (1,1,1)	461.6702	470.6066
ARIMA (1,1,2)	463.6319	474.8024
ARIMA (2,1,0)	453.1814	459.8399
ARIMA (2,1,1)	453.6230	464.7205
ARIMA (2,1,2)	449.9905	463.3075

2.4.5 Forecast performance measurement

In the previous process, we estimated forecast performance and compared different models using the root mean squared error (RMSE) and mean absolute percentage error (MAPE). These measures are defined in Equations 5 and 6, respectively.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(observed_t - forecast_t)^2}{n}} \quad (5)$$

$$MAPE = \frac{\sum_{t=1}^{n}\left|\frac{observed_t - forecast_t}{observed_t}\right|}{n} \quad (6)$$

3. RESULTS

In order to compare the proposed approach and the classical approaches, three experiments were undertook to apply the classical time series data and the enriched time series data from the proposed method. In the first trial, the classical time series data of dengue cases were used with the ARIMA model. For the second trial, the classical time series of dengue cases, together with climate as an exogenous variable, were used with ARIMAX. In the final trial, the ARIMA model with enriched time series data was performed based on the knowledge derived from the climate factors (using the same exogenous variables as in ARIMAX) as semantic information. The model developed was called "ARIMA with semantic data (ARIMAS)." To satisfy the forecasts, the optimal parameters were selected for each trial.

3.1 Optimal parameters

For parameter optimization, the order of difference d was identified to ensure that the classical and enriched time series data were stationary. The number of monthly dengue cases in Phitsanulok from 2003 to 2017 was used as the training set of classical time series. The modeling of the enriched time series data used dengue cases from the same period together with the lag time of their related data from each month. The ADF test provides Tau-stat and Tau-crit values to check the stationarity property of the time series. In Table 3, the Tua-stat value of the classical time series (d=0) is greater than the Tau-crit value. This implied that the classical time series data were not stationary. The first order of difference (d=1) was

then considered to be the optimal variable of the classical time series, because of the greater Tau-crit value. The same test was also applied to the enriched time series, where the result achieved, indicating that it is stationary at d=0.

Table 3. The optimal order of d from stationarity identification processing of classical and enriched time series of dengue incidences in Phitsanulok, Thailand

Time series data	d = 0			d = 1		
	Tau-stat	Tau-crit	Stationary	Tau-stat	Tau-crit	Stationary
Classical time series	-1.3382	-2.8884	no	-4.6468	-2.8886	yes
Enriched time series	-3.7658	-2.8921	yes			

In order to identify the possible order of p and q for the AR and MA of the models, the ACF and PACF values of the stationary time series were examined using their lag and the significance limit. In Figure 7, the ACF and PACF values of the classical time series indicated that the possible order of p and q were 2. Moreover, the ACF and PACF values of the enriched time series indicated that the possible order of p was 3 and of q was 1.

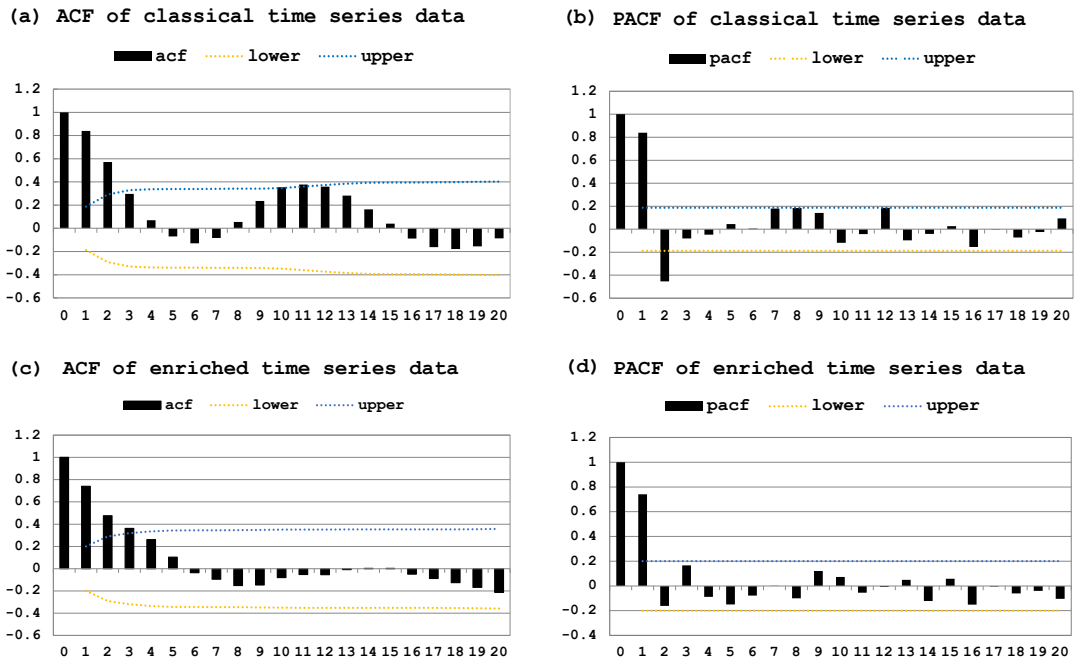

Figure 7. ACF and PACF of classical and enriched time series data

From the selected parameters, a candidate of each model was created and then the best fitted model was selected using BIC and AIC values. In Table 4, the candidate model of classical time series in both ARIMA and ARIMAX employed the same parameter with d = 1, p in the range of 1 to 2 and q in the range of 0 to 2. The candidate models for the enriched time series were the ARIMAS with d = 0, p = 1 and q in the range of 0 to 3. The candidates that had the lowest BIC and AIC were considered to be the most suitable models. Therefore, the optimal model for ARIMA with classical time series was ARIMA (2,1,0). The optimal model for ARIMAX with classical time series and exogenous data was ARIMAX (1,1,1). In the enriched time series model, ARIMAS (1,0,0) and ARIMAS (1,0,1) both achieved the lowest BIC and AIC. Hence, both models were selected for forecasting in the next process.

3.2 Forecast performance

After the best predictive models were selected with optimal parameters, the dengue incidences were forecasted using a test set. The differences between the observed data and predicted value were used as a reference to compare the forecast performance of each model. To obtain the most accurate method, RMSE and MAPE were used to measure the prediction accuracy. In Table 5, the ARIMA (2,1,0) model with classical time series achieved 25.9706 and 0.7879 for RMSE and MAPE, respectively. The ARIMAX (1,1,1) model with classical time series and exogenous variables achieved 27.4462 and 0.7791 for RMSE and MAPE, respectively. For the enriched time series, the ARIMA (1,0,0) model achieved 24.2866 and 0.7665 for RMSE and MAPE, respectively, and the ARIMAS (1,0,0) model achieved 25.5012 and 0.7769 for RMSE and MAPE, respectively. The results showed that both models of enrich time series data had lower errors, compared to the other models, in all performance measurements.

Figure 8 illustrates the forecasting of the incidence of dengue in the study area. An unusual number of dengue incidences occurred in two periods of the study. The first period was in 2013, when a large epidemic of dengue fever

coincided with the rainy season. There were a fluctuant number of incidences from May to September. The second period was in 2015, when fluctuations appeared during August and December, indicating an irregular period. These unusual situations may have caused a significant prediction error. Therefore, we focused on the unusual period and measured the performance of all models, as illustrated in Table 6. In the first period, the ARIMA (2,1,0) model with classical time series achieved an RMSE of 50.7804 and a MAPE of 2.1040; the ARIMAX (1,1,1) model achieved an RMSE of 52.7405 and a MAPE of 2.1601; while the ARIMAS (1,0,0) model with enriched time series achieved an RMSE of 48.4837 and a MAPE of 0.0739. In the second period, the ARIMA (2,1,0) model with classical time series achieved 52.6731 and 0.4825 for RMSE and MAPE, respectively; the ARIMAX (1,1,1) model achieved 49.8187 and 0.4826 for RMSE and MAPE, respectively; and the ARIMAS (1,0,0) model with enriched time series achieved 34.2573 and 0.0588 for RMSE and MAPE, respectively. The enriched time series model had lower errors in the unusual period and a varying number in both the first and second periods. These results indicated that ARIMAS, was an improved dengue incidence prediction model that can lead to improvements in forecasting performance.

Table 4. AIC and BIC values of optimal ARIMA models for classical and semantic time series of dengue incidences in Phitsanulok, Thailand

Time series data	Candidate model	AIC	BIC
Classical time series	ARIMA (1,1,0)	458.0282	462.4964
	ARIMA (1,1,1)	461.6702	470.6066
	ARIMA (1,1,2)	463.6319	474.8024
	ARIMA (2,1,0)	453.1814	459.8399
	ARIMA (2,1,1)	453.6230	464.7205
	ARIMA (2,1,2)	449.9905	463.3075
Classical time series with exogenous variables	ARIMAX (1,1,0)	-416.4794	-414.2599
	ARIMAX (1,1,1)	-447.7051	-443.2661
	ARIMAX (1,1,2)	-443.4150	-436.7565
	ARIMAX (2,1,0)	-408.9969	-404.5875
	ARIMAX (2,1,1)	-434.7727	-428.1587
	ARIMAX (2,1,2)	-424.5995	-415.7808
Enriched time series	ARIMAS (1,0,0)	624.0043	629.1120
	ARIMAS (1,0,1)	621.7150	631.9305
	ARIMAS (1,0,2)	623.2749	636.0443
	ARIMAS (1,0,3)	625.2721	640.5953

Table 5. RMSE and MAPE of optimal models for classical and enriched time series of dengue incidence forecasts in Phitsanulok, Thailand

Time series data	Model	RMSE	MAPE
Classical time series	ARIMA(2,1,0)	25.9706	0.7879
Classical time series with exogenous variables	ARIMAX(1,1,1)	27.4462	0.7791
Enriched time series	ARIMAS(1,0,0)	24.2866	0.7665
	ARIMAS(1,0,1)	25.5012	0.7769

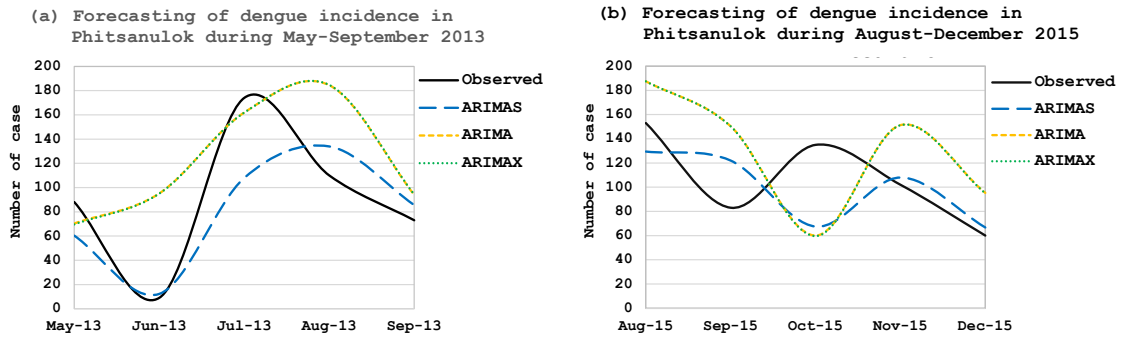

Figure 8. Forecasting of dengue incidence in Phitsanulok in which high and variant outbreak

Table 6. RMSE and MAPE of dengue incidence forecasted in significant periods of high and varying number of cases

Time Series Data	Model	Period: May to September 2013		Period: September to December 2015	
		RMSE	MAPE	RMSE	MAPE
Classical time series	ARIMA(2,1,0)	50.7804	2.1040	52.6731	0.4825
Classical time series with exogenous variables	ARIMAX(1,1,1)	52.7405	2.1601	49.8187	0.4826
Enriched time series	ARIMAS(1,0,0)	48.4837	0.0739	34.2573	0.0588

4. DISCUSSION

Figure 9 shows that the traditional ARIMA provided the highest RMSE because its algorithm considered only the previous data and ignored other relevant factors that affected the spreading of the outbreak. Although ARIMAX takes several features into account to create a model and predict dengue incidence in the future, its performance is slightly lower than the ARIMA model because some continuous data of exogenous variables are not stationary; thus, the covariate coefficient of exogenous and response variables is inconsistent and arrives at a spurious regression situation. ARIMAS achieved the lowest RMSE in every year because the quality of the input data into the model was higher than for the others, leading to better prediction quality with a very low RMSE each year. This proves that the wise saw of "*garbage in, garbage out*" is definitely true in the data analytics arena. Therefore, if the quality of the input data was enhanced, better prediction performance can be obtained.

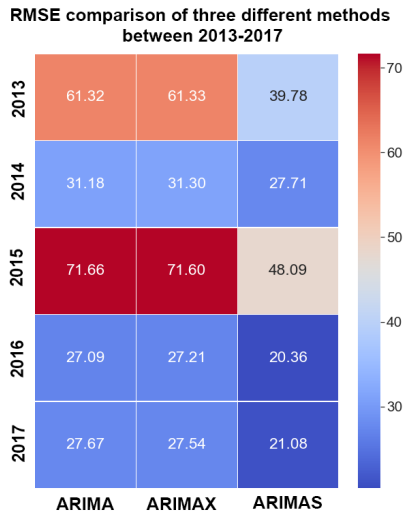

Figure 9. Comparison matrix of RMSE of three different methods

In recent decades, machine learning has become an area in which researchers have introduced several approaches involving time series analysis. For dengue incidence forecasting (Chakraborty et al., 2019; Guo et al., 2017; Zhao et al., 2020; Benedum et al., 2020; Xu et al., 2020), machine learning algorithms are mainly used as predictive models such as the support vector machine (SVM), step-down linear regression, gradient boosted regression tree (GBM), negative binomial regression (NBM), least absolute shrinkage and selection operator linear regression (LASSO), generalized additive (GAM), random forest (RF), artificial neural networks (ANN), and deep learning. However, those works usually ignored the relationships among data, which may introduce limitations that lead to low prediction power of the model. For example, in exceptional cases such as natural disasters, the number of cases in a region can be abnormally high or low. Therefore, using only previous data of that region may cause the forecasting model to return a huge error. As opposed to those existing models, our approach considers semantic information existing in the data. Then, the presented system intelligently selects previous data (factors) that are similar and relevant to the meta-data of the study area. When only relevant data is used for the prediction, noisy data is eliminated and, consequently, the quality of input data is higher than the original data. As a result, this can enhance the prediction power of the model.

5. CONCLUSION

The goal of this study was to develop a novel method that used dengue fever factors together with their semantic relationships in order to identify and include related data in the observation. We used data from several provinces in Thailand to assist the prediction of dengue incidences in the study area, such as Phitsanulok. We proposed a method of using the knowledge of dengue fever represented in an ontology model to be incorporated with the ARIMA technique to enhance its efficiency, that we termed "ARIMAS." Based on the results, we are confident that our approach improves the forecast performance even when unusual events occur that affect the number of cases. Our findings can be used to inform public health policy makers and planners how to provide more timely and accurate warnings to public health staff and citizens on the likelihood of imminent outbreaks. Our findings can also be guidelines for effective public health resource management.

Our future work goals are to expand the ARIMAS technique to be able to determine the best *p, d,* and *q* for ARIMA. We would like to extend our work to provide support for ARIMAX and extend the technique to more study areas.

ACKNOWLEDGMENT

This research was supported by Computer Science and Information Technology Department, Science Faculty, Naresuan University (Grant No. R2565E047), Health Systems Research Institute (Grant No. 64-156), the Program

Management Unit for Human Resources & Institutional Development, Research and Innovation - CU (Grant number B16F630071), Thailand Science Research Innovation (TSRI)-CU (Grant No. FRB640001), and Thailand National Science, Research and Innovation (Fundamental Fund-NU: Grant No. R2565B063). The manuscript was edited by Mr. Roy I. Morien of the Naresuan University Graduate School.

REFERENCES

Anggraeni, W., and Aristiani, L. (2016). Using Google trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia. In *Proceedings of the 2016 International Conference on Information Communication Technology and Systems*, pp. 114-118. Surabaya, Indonesia.

Benedum, C. M., Shea, K. M., Jenkins, H. E., Kim, L. Y., and Markuzon, N. (2020). Weekly dengue forecasts in Iquitos, Peru; San Juan, Puerto Rico; and Singapore. *PLOS Neglected Tropical Diseases*, 14(10), e0008710.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5th edition), Milton, Queensland: John Wiley & Sons, Inc., pp. 64-66.

Chakraborty, T., Chattopadhyay, S., and Ghosh, I. (2019). Forecasting dengue epidemics using a hybrid methodology. *Physica A: Statistical Mechanics and its Applications*, 527, 121266.

Cortes, F., Turchi Martelli, C. M., Arraes de Alencar Ximenes, R., Montarroyos, U. R., Siqueira Junior, J. B., Gonçalves Cruz, O., Alexander, N., and Vieira de Souza, W. (2018). Time series analysis of dengue surveillance data in two Brazilian cities. *Acta Tropica*, 182, 190-197.

Dickey, D. A., and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366), 427-431.

Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., Luo, G., Li, Z., He, J., Zhang, Y., and Ma, W. (2017). Developing a dengue forecast model using machine learning: A case study in China. *PLOS Neglected Tropical Diseases*, 11(10), e0005973.

Herdiani, A., Fitria, L., Hayurani, H., Wibowo, W., and Sungkar, S. (2012). Hierarchical conceptual schema for dengue hemorrhagic fever ontology. *International Journal of Computer Science,* 9(4), 53-58.

Jing, Q. L., Cheng, Q., Marshall, J. M., Hu, W. B., Yang, Z. C., and Lu, J. H. (2018). Imported cases and minimum temperature drive dengue transmission in Guangzhou, China: Evidence from ARIMAX model. *Epidemiology & Infection*, 146(10), 1226-1235.

Johansson, M. A., Cummings, D. A. T., and Glass, G. E. (2009). Multiyear climate variability and dengue—El Niño southern oscillation, weather, and dengue incidence in Puerto Rico, Mexico, and Thailand: A longitudinal data analysis. *PLOS Medicine*, 6(11), e1000168.

Leacock, C., and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. In *WordNet: An Electronic Lexical Database* (Fellbaum, C., and Miller, G., eds.), pp. 265-283. Massachusetts: MIT Press.

Lu, L., Lin, H., Tian, L., Yang, W., Sun, J., and Liu, Q.-Y. (2009). Time series analysis of dengue fever and weather in Guangzhou, China. *BMC Public Health*, 9(1), 395.

Mitraka, E., Topalis, P., Dritsou, V., Dialynas, E., and Louis, C. (2015). Describing the breakbone fever: IDODEN, an ontology for dengue fever. *PLOS Neglected Tropical Diseases*, 9(2), e0003479.

Nagao, Y., Thavara, U., Chitnumsup, P., Tawatsin, A., Chansang, C., and Campbell-Lendrum, D. (2003). Climatic and social risk factors for Aedes infestation in rural Thailand. *Tropical Medicine & International Health*, 8(7), 650-659.

Nayak, M. S. D. P., and Narayan, K. A. (2019). Forecasting dengue fever incidence using ARIMA analysis. *International Journal of Collaborative Research on Internal Medicine and Public Health*, 11(6), 924-932.

Nguyen, N. (2018). Predicting dengue spread using seasonal ARIMAX model and meteorological data. *Towards Data Science.* [Online URL: https://towardsdatascience.com/predicting-dengue-spread-using-seasonal-arimax-model-on-meteorology-data-3f35979ec5d] accessed on September 13, 2018.

Polwiang, S. (2020). The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017). *BMC Infectious Diseases*, 20(1), 208.

Siriyasatien, P., Phumee, A., Ongruk, P., Jampachaisri, K., and Kesorn, K. (2016). Analysis of significant factors for dengue fever incidence prediction. *BMC Bioinformatics*, 17, 166.

Somboonsak, P. (2019). Forecasting dengue fever epidemics using ARIMA model. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference,* pp. 144-150, Kobe, Japan.

Thiruchelvam, L., Dass, S. C., Zaki, R., Yahya, A., and Asirvadam, V. S. (2018). Correlation analysis of air pollutant index levels and dengue cases across five different zones in Selangor, Malaysia. *Geospatial Health*, 13(1), 102-109.

World Health Organization. (2020). *Dengue and severe dengue.* [Online URL: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue] accessed on June 23, 2018.

Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L., and Liu, Q.-Y. (2020). Forecast of dengue cases in 20 Chinese cities based on the deep learning method. *International Journal of Environmental Research and Public Health*, 17, 453.

Zhao, N., Charland, K., Carabali, M., Nsoesie, E. O., Maheu-Giroux, M., Rees, E., Yuan, M., Balaguera, C. G., Ramirez, G. J., and Zinszer, K. (2020). Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. *PLOS Neglected Tropical Diseases*, 14(9), e0008056.