Engineering Journal issn: 0125-8281

Article

Weekly Forecasting Model for Dengue Hemorrhagic Fever Outbreak in Thailand

Akeamorn Puengpreeda[a], Suphawit Yhusumrarn[b], and Surapong Sirikulvadhana[c,*]

Department of Industrial Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand
E-mail: [a]akeamorn.puengpreeda@gmail.com, [b]suphawit.yhusumrarn@gmail.com, [c]surapong.email@gmail.com (Corresponding author)

Abstract. A dengue virus causes diseases, including dengue hemorrhagic fever (DHF) which induces several sicknesses and deaths in Thailand. DHF is categorized as one of the most dangerous communicable diseases by the Ministry of Public Health Thailand (MoPH); moreover, the MoPH also sets strict protocols and encourages forecasting techniques for monitoring and dealing with the outbreaks. This research aims to utilize the data that were gathered from external sources, e.g. Google Trends data and meteorology data, to forecast the number of cases that will occur within the 7 day-interval in the next 1–4 weeks. Six provinces—including Chiang Rai, Mukdahan, Pattani, Phichit, Ayutthaya, and Ratchaburi—were selected as they represent the unique patterns of dengue outbreaks in Thailand. The machine learning models—including Random Forest, AdaBoost, Extra-Trees, and Regularized Regressions—were used to forecast the number of the cases. The performances of these models were compared to the performances of the traditional time series model including Naïve model and Moving Average. The proposed machine learning models for Chiang Rai, Mukdahan, and Pattani yield better results than those of the traditional models.

Keywords: Dengue hemorrhagic fever, machine learning, disease forecasting, Google Trends.

ENGINEERING JOURNAL Volume 24 Issue 3
Received 2 July 2019
Accepted 25 March 2020
Published 31 May 2020
Online at https://engj.org/
DOI:10.4186/ej.2020.24.3.71

DOI:10.4186/ej.2020.24.3.71

1. Introduction

Dengue is a crucial mosquito-borne disease, causing illnesses and deaths, and mainly transmitted by *Aedes aegypti* mosquitoes. Dengue quickly spreads throughout several tropical countries, including Thailand. According to the report from 2014 to 2018 [1], dengue caused, on average, 90,471 DHF cases and 91 deaths per year in Thailand. Due to its severity, DHF is categorized as one of the most dangerous communicable diseases by MoPH. There were several imposed protocols in order to handle the outbreaks. However, these protocols depend on a surveillance system, i.e. they are reactive. Therefore, the Thai Government has encouraged the uses of forecasting techniques for anticipating the upcoming short- and long-term outbreaks.

To forecast the situation of the outbreak that will occur in the future, there are multiple objectives depend on the time horizon that the models forecast. For example, for long-term forecasting, the models might assist in foreseeing the overall picture of a situation that will occur in the next few months or maybe years. Nevertheless, the short-term forecasting can provide other aspects to help related parties plan to intervene in the upcoming outbreak. There are many challenges in order to forecast the situation occurring in a short time horizon, e.g. 1–4 weeks ahead. Thereby, the promising traditional models for forecasting the short-term situation of the outbreak tend to rely on the concept of time-series forecasting that the recent data gain more weight.

Beside of the forecasting models that are based on the previous occurring situation, there are many pieces of research using different variables and models to forecast dengue outbreaks. One study showed that weather conditions seem to affect the abundance of the vectors and also the risk of dengue [2]. Hence, some studies utilized the meteorology data—such as rainfall, relative humidity, and temperature—as features of the dengue forecasting models [3], [4]. Furthermore, the search query data on the search engines (e.g., Google, Yahoo, and Baidu) seem to have a high correlation with the severity of some disease outbreaks. Thereby, they were used as the input variables for disease forecasting models [5].

Nevertheless, in Thailand, there is no research that used meteorology data, the lagged terms of the number of illness cases, and the search query data to forecast the short-term situation. Thereby, the objective of this research is to integrate the weather data, the lagged terms of the number of DHF cases, and the search query volume into machine learning models for forecasting the number of DHF cases. The models are designed to forecast the incidences in 1–4 weeks horizon for the selected provinces. The provinces that their performance of short-term forecasting models—based on only lagged terms like Naïve and Moving Average—can be improved by integrating these external variables into the machine learning models are also investigated.

2. Literature Review

2.1. Basic Knowledge of Dengue

Dengue is a tropical mosquito-borne infection especially widespread in tropical and subtropical areas. Dengue virus is mainly transmitted by female *Aedes aegypti* mosquitoes [6]. Dengue usually causes flu-like illness (known as dengue fever, DF) such as a high fever, headache, joint pain, or skin rash. In a particular case, the infection develops into life-threatening complications called dengue hemorrhagic fever (DHF), leading to bleeding and blood plasma leakage, or into a critical stage named dengue shock syndrome (DSS), causing dramatically low blood pressure and risk of death.

DEN-1, DEN-2, DEN-3, and DEN-4 are four serotypes of the virus that cause dengue. Due to shared antigen determinants among these four serotypes, temporary cross-reaction and cross-protection can occur. Recovery from one particular serotype grants permanent immunity against that serotype but only provides partial immunity against the other three serotypes for about 6–12 months. The risk of developing DHF increases with consecutive infections by the other serotypes [7].

As the primary vector of dengue, female *Aedes aegypti* mosquitoes are able to transmit the virus for the rest of its life once infected. Humans, bitten by infected mosquitoes, become a virus source for uninfected mosquitoes. The virus, in humans, can be transmitted for 4–12 days once the first symptom showed [7].

Some meteorology variables (e.g. temperature, precipitation, and humidity) can lead to an abundance of the mosquitoes, as well as the risk of dengue infection [2]. Because of proper conditions, the number of patients infected by dengue in the rainy season is high compared to the other seasons.

2.2. Dengue-related Data

Previous studies on disease forecasting used various types of information to obtain disease insights based on its nature or characteristic. Different kinds of data described below are going to be implemented in proposed forecasting models as they are naturally related to dengue incidences in many aspects.

2.2.1. Search Query Volume

Google Trends is a tool for investigating the trend of a keyword that was searched in a particular time-frame and location. It provides a result in the form of Google Trends score. The score is scaled from 0 to 100 according to the popularity of keywords across different locations within a fixed time range, or across different periods of time within a particular area. It was shown that Google Trends data could capture the peak of the flu incidences in Taiwan [8]. Google query volume data can be used to accurately predict influenza epidemics at the regional level in the U.S. [9]. Furthermore, the

Google Trends data could also be used as a feature for a tuberculosis prediction model at the state and national levels in the U.S. [10].

In order to find the proper search terms to be used as forecasting model features, the Google Correlate platform seems to be a useful site. Since it can list up to 100 words that have a similar search volume trend to the interested keyword. It was used to find 164 search terms that may have some correlation with 64 infectious diseases to create the prediction models [11].

2.2.2. Meteorology Data

The number of dengue infection cases seems to increase rapidly in the rainy season. Because the vector can reproduce in rainwater that accidentally trapped in containment site or forested swamp. Moreover, other characteristics of the weather also have effects on both human and mosquito behavior [7]. Hence, it is necessary to use the meteorology data as features to dengue prediction models [3], [4], [12].

2.2.3. Incidence Data

Since the sets of disease incidence data are in a time-series form and some disease outbreak data seem to have seasonal and/or trend patterns [7]. The incidence data from the past, therefore, are very crucial information to create a prediction or forecasting model [10], [13].

2.3. Machine Learning Models for Disease Forecasting

Machine learning models are effective methods for various kinds of analysis tasks. Thanks to its diversity and benefits, many machine learning techniques were used to forecast disease outbreaks by several researchers. Some machine learning algorithms were selected to be applied to this research as explained in the following subsections.

2.3.1. Tree-based Models

Decision tree model can be used to build a prediction model using sequential decision steps in order to minimize the loss function to find the best possible model called strong learners. The decision steps begin with the root, then decide which branch should be selected, based on criteria on each step, and keep deciding until reaching the last leaf. However, there are other approaches, called ensemble learning, that use many weak learners based on a majority vote or average outcome in order to predict the result.

Random forest is one of the most popular ensemble tree-based models that seems to have great performance. It uses many decision trees that have different sets of features and steps to average the results in order to predict the target [14]. However, some hyperparameters, such as the number of estimators that indicates how many trees should be used to create a model, the maximum number of features used in a tree, and the minimum number of leaves required to split an internal node, have to be tuned.

In Guangdong, China, [12] used a gradient boosted regression tree algorithm to forecast weekly dengue cases. In Manila, the Philippines, [15] developed a random forest model to predict weekly dengue incidences. In Singapore, [16] mapped a spatial dengue risk using a random forest model. Globally, [17] also implemented and random forest models to map a Zika transmission risk.

Beside of the models mentioned above, there are some ensemble concepts of the tree-based model that usually promise the accurate predicting results, Adaptive Boosting (AdaBoost) and extremely randomized trees (Extra-Trees) regressor [18].

2.3.2. Regularized Regression Models

Different from Linear Regression Models, Regularized Regression Models combine the approach of important feature identification to the models, i.e. penalization, which can alleviate an overfitting problem and extract significant variables from the models. These models are simple yet robust and applicable to several forecasting tasks.

Ridge Regression and Least Absolute Shrinkage and Selection Operator (LASSO) are kinds of regularized regression models which were implemented to forecast dengue incidences. For example, [19] developed LASSO to forecast weekly dengue incidences in Singapore and also the result can be beneficial to the country's dengue control program.

3. Methodology

This section describes processes for developing DHF forecasting models as illustrated in Fig. 1. The processes can be summarized as a framework with four components: Data Preparation and Preprocessing, Traditional Forecasting Models Creation, Machine Learning Models Development, and Model Selection and Feature Importance Finding. The detailed methodology for each component will be explained in the following subsections.

3.1. Data Preparation and Preprocessing

3.1.1. Data Gathering

The datasets used in this project, for feature and target values creation, were acquired from several sources.

3.1.1.1. The Data of the Number of DHF Cases Occurred within Each Province in Thailand

The weekly data of the number of DHF cases occurred within each province in Thailand were provided by the "Report 506 (RP506)" of the department of

disease control Thailand. The data range is from January 2007 to November 2018. Unfortunately, the data from July 2017 to December 2017 were missing.

Although there are 77 provinces in Thailand, only some provinces were used to represent all major unique patterns of the outbreaks in Thailand. K-means clustering model was chosen as a method to group the provinces. Before being clustered, the standardize of the number of DHF cases was calculated by normalizing the raw data along the time (2007–2018) within each province. Then, these values of each province were subtracted by the average value of the whole country in same period to demonstrate the pattern that each province had the outbreak situation under or beyond the situation of the whole country. Consequently, the 6 groups of provinces were clustered, as shown in Fig. 2, because the inertia score after clustering seemed to change with a lower rate after clustering more than 6 groups. Hence, 6 groups clustering was the optimal choice. Since the provinces in each group have similar historical standardized the number of DHF cases to others in the group, it could be assumed that the number of cases occurring in the future of those provinces can be forecasted with the resemble models of the representative province of each group. Chiang Rai, Ayutthaya, Ratchaburi, Phichit, Pattani, and Mukdahan were decided to be the representative of each group due to their high correlation of historical standardized of the number of DHF cases with the average values of their groups.

As shown in Fig. 3, the six groups of provinces were demonstrated as six different colors. Furthermore, the provinces clustered to the same group tend to locate nearby the others in the same group. This might be because the immunities of the people live within each area are the same. Since the immunity of dengue virus depends on the previous infection, the people live within the same neighborhood might experience the prior infection during the earlier peak of the outbreak within that area in the same time.

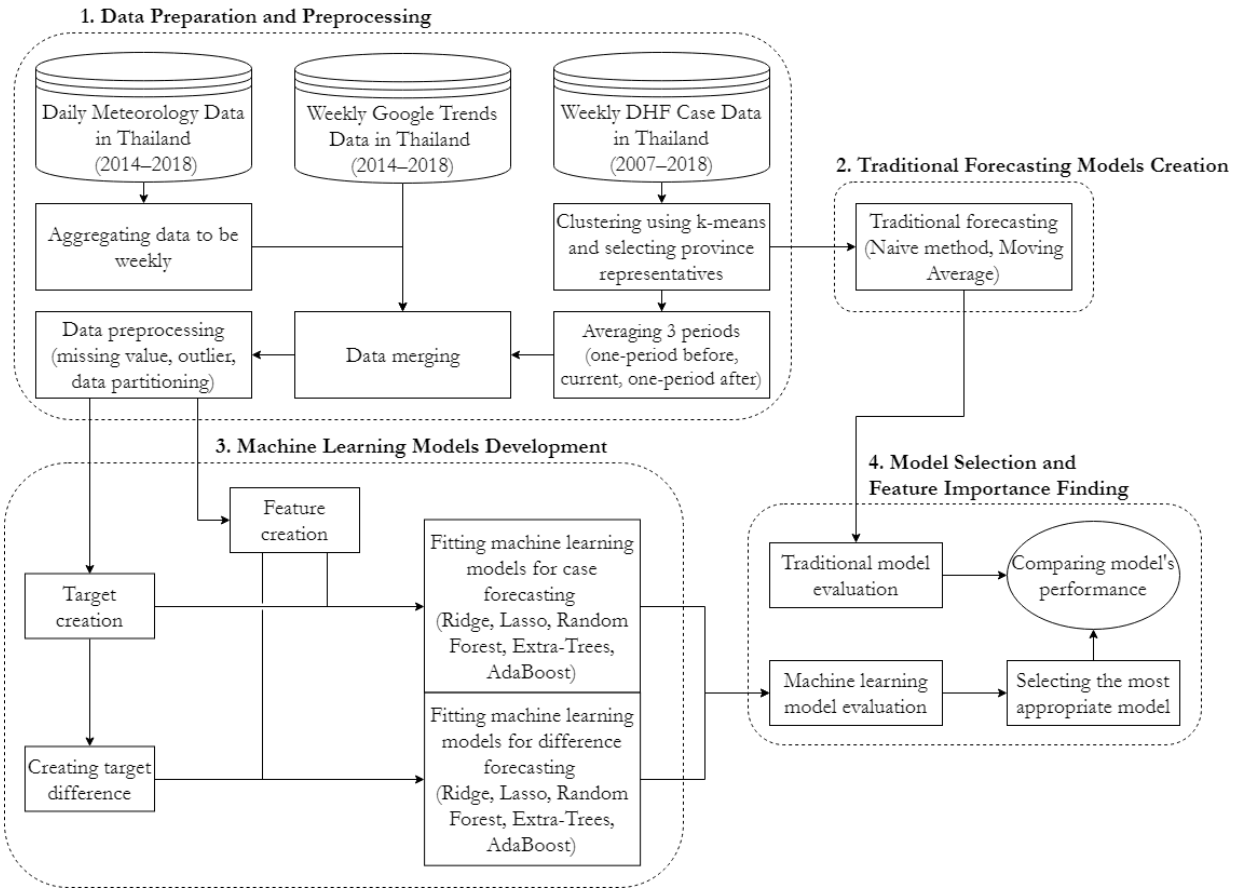

Fig. 1. DHF forecasting models development framework.

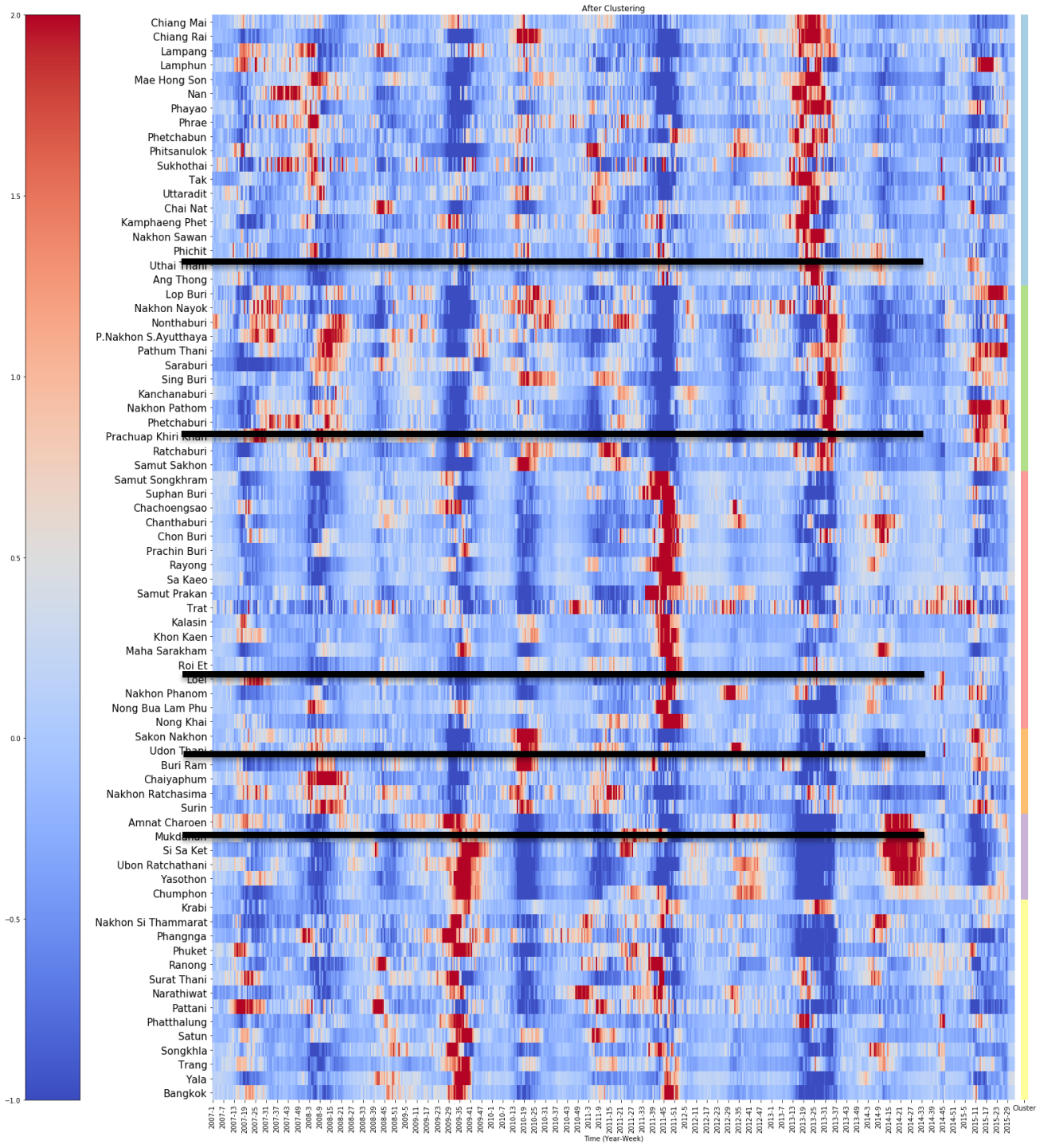

Fig. 2. Heatmap of standardized the number of DHF cases occurring within each province in Thailand grouped by K-means (K=6).

DOI:10.4186/ej.2020.24.3.71

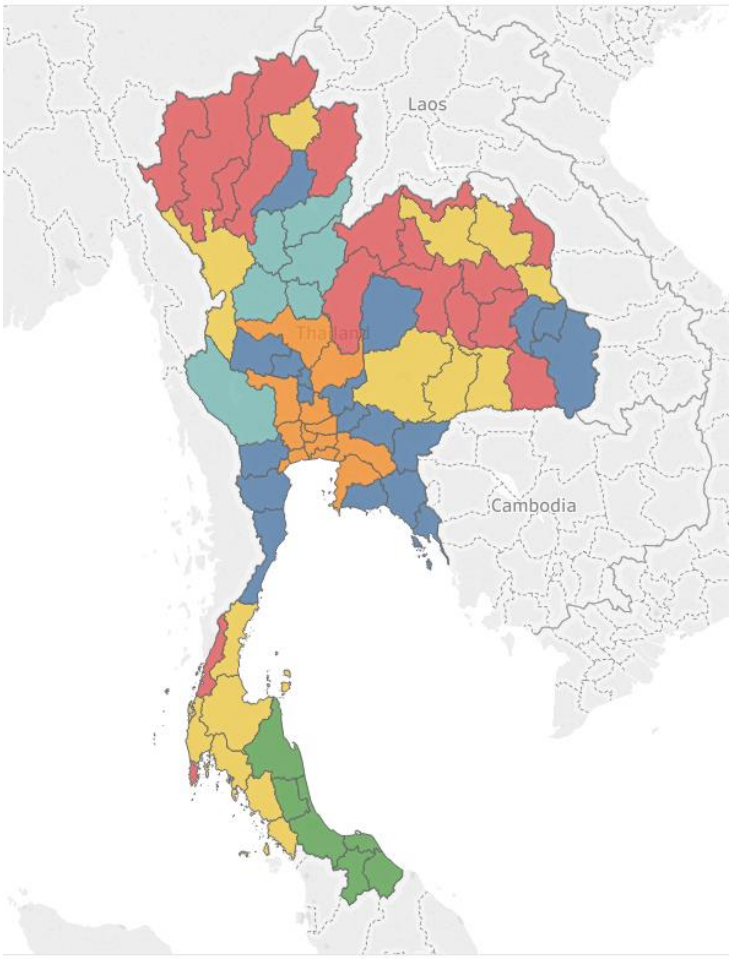

Fig. 3. Thailand map demonstrating the clustered provinces within each group.

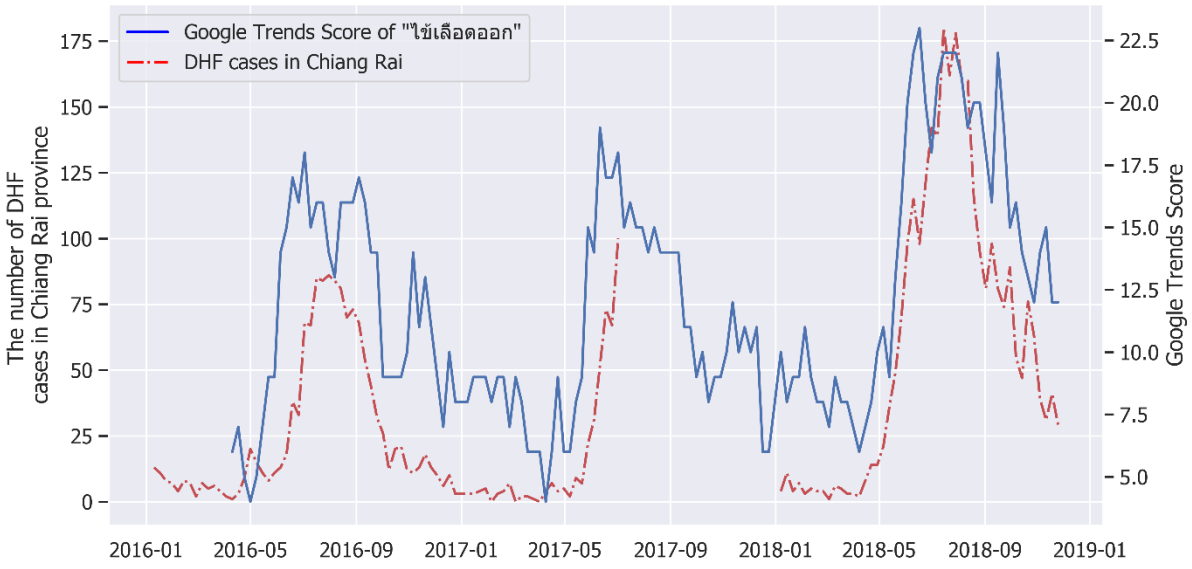

Fig. 4. The number of DHF occurred in Chiang Rai province and the Google Trends score of the keyword "ไข้เลือดออก", which means "dengue" in Thai.

Table 1. Top 10 keywords that are most correlated with keyword "ไข้เลือดออก" on Google Correlate.

Keywords	Correlation	Meaning in English
ยุง ลาย	0.9109	*Aedes aegypti* mosquito
โรค ไข้เลือดออก	0.9046	dengue fever
กำจัด ยุง ลาย	0.8197	kill *Aedes aegypti* mosquito
การ กำจัด ยุง ลาย	0.8078	*Aedes aegypti* mosquito killing
dengue	0.8070	Dengue
dengue fever	0.7973	Dengue fever
ข่าว ไข้เลือดออก	0.7875	dengue fever NEWS
ยุง	0.7838	Mosquito
ป้องกัน ไข้เลือดออก	0.7737	dengue fever prevention
ป้องกัน ยุง	0.7707	mosquito prevention

3.1.1.2. Google Trends Data

Weekly Google Trends scores of dengue-related keywords were obtained in a whole country scale. Unfortunately, the province-level data were not used in this research. In some provinces, the search volume of some selected keywords was too small within weekly interval; therefore, Google does not show the actual score of search volume and demonstrate it as zero to secure the privacy of its users.

Like other researches [9], [13], the search volume of disease name in the native language, "ไข้เลือดออก" that means "dengue" in Thai, have a high correlation with the number of cases occurred in the particular areas. For example, as depicted in Fig. 4, the search score of "ไข้เลือดออก" has high correlation with the number of DHF cases occurred in Chiang Rai. Therefore, the Google Trends score of this keyword is accounted one of features in forecasting models. Moreover, other 10 keywords (see Table 1) that have a high correlation with the keyword "ไข้เลือดออก" in search volume on Google, based on Google Correlate platform [20], were also considered and treated as the model features. The model features considered only 10 additional words because some words that were ranked beyond the tenth seemed to be unrelated to the outbreak. To illustrate, the 11th ranked word, "ผู้ ก่อตั้ง ลูกเสือ" which means "the founder of boy scout organization", seems to not relate the dengue in any aspect.

3.1.1.3. Daily Thailand Meteorology Data

The daily meteorology data—including rainfall, relative humidity, sunshine length, daily maximum temperature, daily minimum temperature, daily average temperature, wind direction, and wind speed—were provided by the information service of Thai Metrological Department. The data range is from 2014 to 2018. Each field of data was aggregated to the weekly interval by averaging the values within each week.

However, the data of sun duration of Pattani province, the selected province, were not be measured and recorded by the Thai Metrological Department; hence, only the rest meteorology data fields were used to build the forecasting model for Pattani.

3.1.2. Data Merging

All datasets involving selected provinces were merged into one weekly interval dataset, range from January 2014 to November 2018.

3.1.3. Data Preprocessing

The missing rainfall values of meteorology data were set to zero. Other missing values, e.g. sunshine duration of Mukdahan, were replaced by the average of sunshine duration in the same week of the year.

From preliminary analysis, the outliers of Google Trends scores were observed. Any extreme score (excess 2.7SD. of each keyword) was set to its 2.7SD. The zero values of Google Trends scores were set to the average value between the minimum value that is not zero and zero.

With machine learning techniques, data partitioning process is required. The dataset was split into two datasets. First, the dataset, range from January 2014 to June 2017, was prepared for training the models. Second, the dataset from January 2018 to November 2018 was prepared for the final evaluation of each model.

3.2. Traditional Time-series Forecasting Model

In this research, the Naïve model and Moving Average (time window is 2 weeks) were used as baseline models because the forecast results of these models are mainly weighted by the very recent data. These baseline models are the models that tailor-made models tried to overcome. However, Exponential Smoothing with Linear Trend was not included in this research as from a preliminary experiment, the best parameters (alpha and

beta) of Exponential Smoothing with Linear Trend seemed to be close to one and zero, respectively (strongly weighting based on the most recent data point resembling Naïve method). For example, the best alpha and beta of the 1-week ahead model for Ratchaburi are 0.94 and 0.0051, respectively, and the 4 weeks ahead model are 0.98 and 0, respectively. Moreover, the same situation occurred in other provinces. Therefore, only the Naïve method and Moving Average would be discussed in this research.

As these models are typically used for a very short forecasting step, i.e. one period ahead, the models were proposed under a practicality assumption. To illustrate, if a forecast occurs in week T, the forecasted incidence values in the next 4 weeks (week T+1, T+2, T+3, and T+4) by the Naïve model will be the same, which is the number of DHF incidences in week T. Likewise, the forecasted incidences from Moving Average will also be the same, which is the average value between the number of incidences in week T-1 and T.

3.3. Machine Learning Models

This subsection explains a machine learning procedure that this work follows. The procedure covers why new target values for the models are intentionally crafted, how model features are engineered, and how to fit, fine-tune, and evaluate the models.

3.3.1. Target Value Creation

In order to forecast the number of DHF cases that will occur in the next 1–4 weeks accurately, the model had to learn the pattern from historical data. However, the historical data seems to be noisy and have fluctuating trends (see Fig. 5) that potentially cause overfitting and lagging issues. Therefore, new target values that remedy noises and trends by averaging and differentiating between two smoothed terms were created (discussed later in 3.3.1.1 and 3.3.1.2, respectively). By these target creation methods, it was believed that the models can capture major trends, without excess noises, and forecast DHF incidences more accurately.

3.3.1.1. First Target

As mentioned above, the average value of 3 periods was considered as a target value for training. Since the model could train with any data points in a training dataset, the target value was calculated by averaging the number of the cases from the previous week, current week, and the next week. To illustrate, the models will learn features in week T to forecast the average value of the number of DHF cases occurred in week T-1, T, and T+1. As shown in Fig. 5, the target values (orange triangle point) are made by averaging three red diamond points, i.e. actual data.

The reasons why future data points are included in modifying the targets are the trend and seasonality. It is noticeable that the incidence data show possible trends and a seasonal manner as they are nature-related. Averaging only current and past records may result in a lagged manner and not represent the actual disease epidemic. On the other hand, including some future data in the target creation process can solve this issue. By learning these target values, the machine learning models could forecast more accurately compared to those that tried to forecast the actual data that were full of noises and fluctuating trends.

3.3.1.2. Second Target

Besides the created first target values mentioned in 3.3.1.1, another form of target value was crafted, called the second target. Although the models, built based on learning the first target values, could resolve some overfitting problems, most models seemed to face the major problem that they could not forecast the target values being out of the range of the training dataset. For example, their forecasting values were not able to reach the peaks of some outbreaks. Hence, another target value form that is more stationary was developed.

By forecasting the difference of the first targets accurately, the number of cases that tends to occur in the future could be forecasted more precisely. The differences of the first target points between weeks vary within limited range compared to the actual number of cases or the first target values. It is supposed that the machine learning models might perform better in forecasting these values. As demonstrated in Fig. 6, the second target value (red triangle point) was created by calculating the difference between two blue diamond points. By adding the forecasted second target to the previous first target, the results of these models might be better. To summarize, the patterns of actual incidences, the first targets, and the second targets for selected provinces can be found in Fig. 7.

3.3.2. Feature Creation

Due to the life cycle of mosquitoes, the lagged terms of meteorology data seem to affect the propagation of the mosquitoes and also dengue virus [21]. Moreover, many previous pieces of research showed that some lag terms of temperature have a high correlation with the recent situation of dengue. For example, [22] found that 2 months lag term of rainfall and minimum temperature has a high correlation with the abundance of the vector in the north region of Thailand. [23] found that the minimum temperature since the last two weeks is also a strong input variable for dengue prediction. Therefore, the lagged terms (up to 60 lags), 2–8 weeks averages, 2–8 weeks minimums, 2–8 weeks maximums, and differences between recent week data and the last 2–8 weeks for all fields of meteorology data were created. This feature engineering method was also applied to other data, e.g. Google Trends score and smoothed the number of the DHF cases. All created features were listed in Table 2.

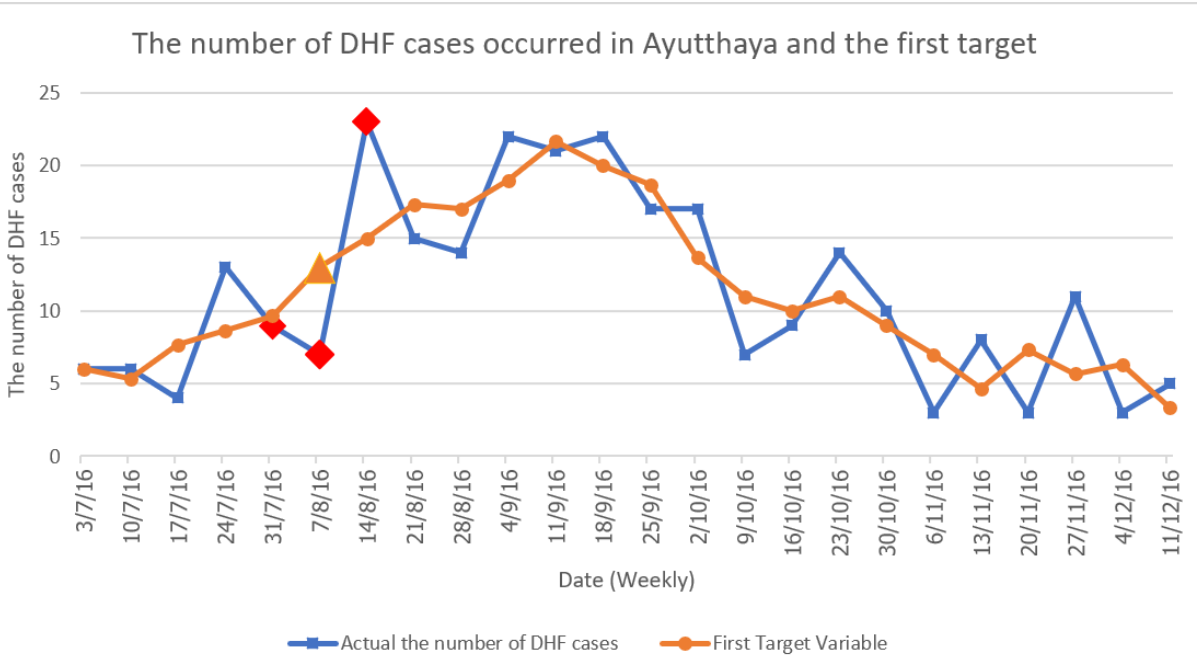

Fig. 5. The examples of the actual number of DHF cases (rectangle mark on the blue line) and created first target values for model training (circle mark on the orange line) from training dataset of Ayutthaya province.

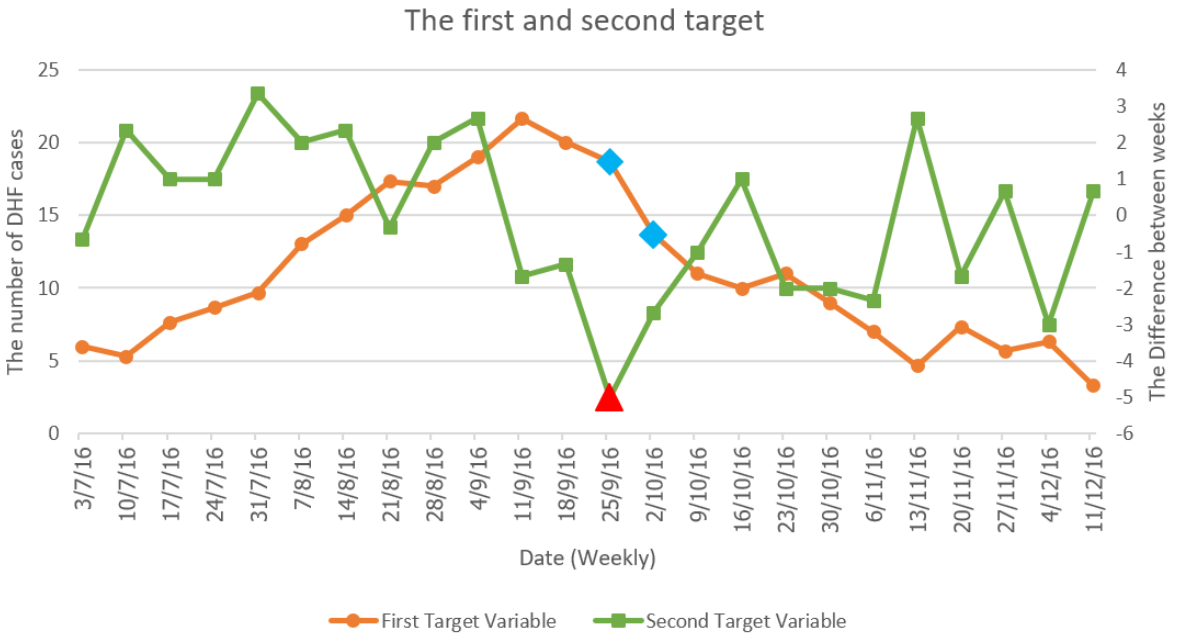

Fig. 6. Second targets (rectangle mark on the green line) created by differentiation of first target (circle mark on the orange line) from training dataset of Ayutthaya province.

DOI:10.4186/ej.2020.24.3.71

Fig. 7. the number of DHF cases, first target, and second target of Chiang Rai (a), Mukdahan (b), and Pattani (c).

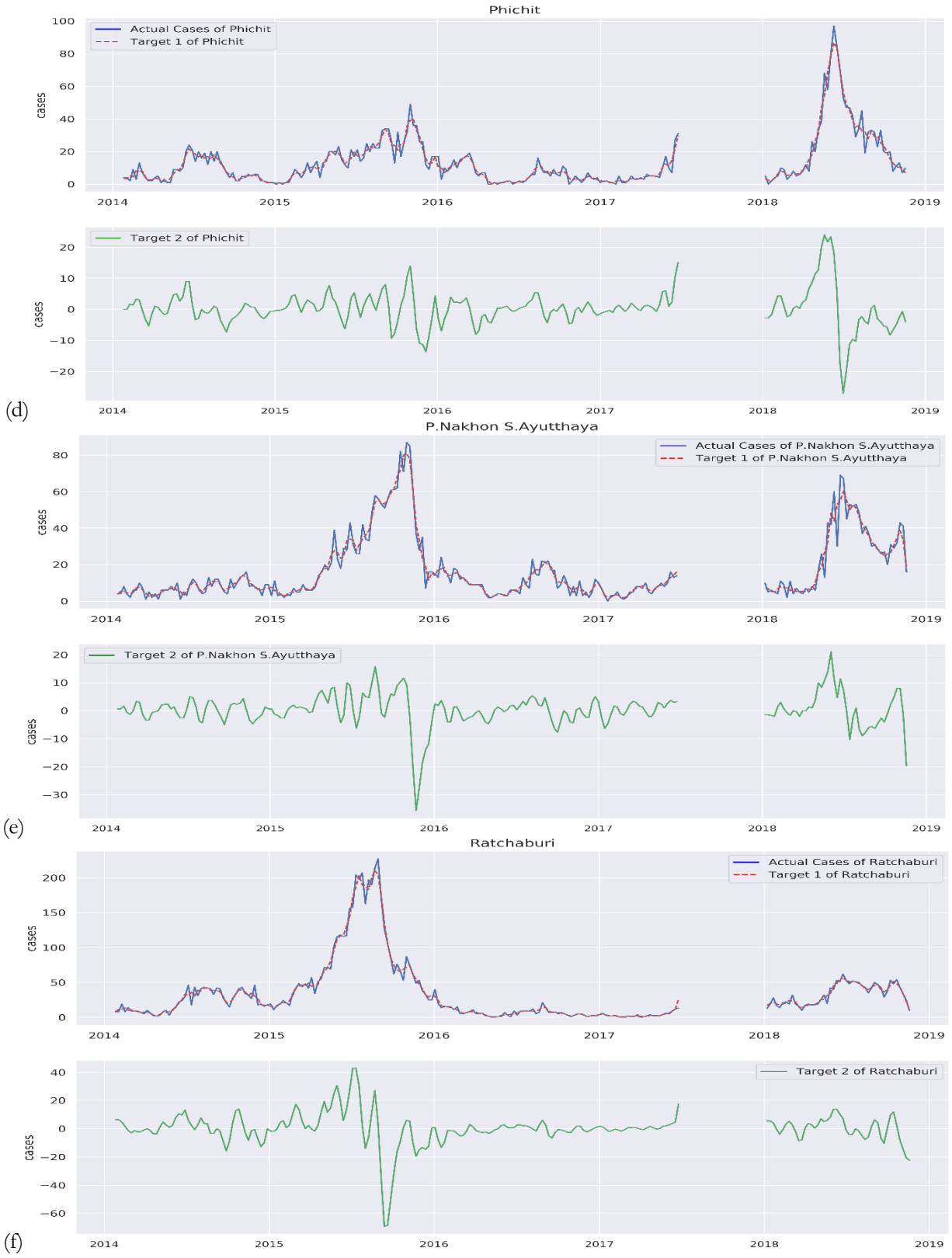

Fig. 7. (continue) the number of DHF cases, first target, and second target of Phichit (d), Ayutthaya (e), and Ratchaburi (f).

Table 2. Feature creation summary.

Original Features	Lags	Difference	Max Min and Average values
Google Trends scores of interested keywords (the most recent data while forecasting is 1 lag)	2-60 lagged terms	The Difference between the value of the previous week and those of the other 8 previous weeks (difference between week T-1 and T-2, T-3, …, T-9)	Average, minimum, and maximum values of the time-window since the previous week to the other previous T weeks (for T in [1, 2, 3, 4, 5, 6, 7])
Meteorology Data (the most recent data while forecasting is 0 lag)	1-60 lagged terms	The Difference between the value of the current week and those of the other 8 previous weeks (difference between week T and T-1, T-2, …, T-8)	Average, minimum, and maximum values of the time-window since the recent week to the other previous T weeks (for T in [1, 2, 3, 4, 5, 6, 7])
The smoothed number of DHF cases (the most recent data while forecasting is 0 lag)	1-60 lagged terms	The Difference between the value of the current week and those of the other 8 previous weeks (difference between week T and T-1, T-2, …, T-8)	Average, minimum, and maximum values of the time-window since the recent week to the other previous T weeks (for T in [1, 2, 3, 4, 5, 6, 7])

3.3.3. Fitting Machine Learning Models

3.3.3.1. Fitting Machine Learning Models for Case Forecasting

The first target values, discussed in subsection 3.3.1.1, along with created features, discussed in subsection 3.3.2, were fitted into Ridge Regression, LASSO, Random Forest regressor, Extra-Trees regressor, and AdaBoost. By using these models, the weekly number of cases that will occur in the next 1–4 weeks can be forecasted.

3.3.3.2. Fitting Machine Learning Models for Difference Term Forecasting

The difference of the first target (second target), discussed in subsection 3.3.1.2, along with created features, discussed in section 3.3.2, were fitted into Ridge regularized regression, LASSO, Random Forest regressor, Extra-Trees regressor, and AdaBoost. By using these models, the differentiation between the situation of this week and the next week could be forecasted. In order to forecast the case for the next T week, the forecasting result could be obtained by adding the term of forecasted T-period-difference to the most recent first target.

3.3.4. Model Evaluation

The metric used for tuning and model selecting is mean squared error (MSE) because the model that yield extremely high errors should be penalized severely. This is because, in practical uses, if the errors between forecasting values and actual values are too high, it might cost a lot of things. For example, it might make the government underestimate the situation, that might cause low prevention, or overestimate the situation, that might cause the unnecessary usage of the precious resources. Each machine learning was tuned by selecting the best combination of the importance hyperparameters that yield the minimum average MSE score of validation sets using time-series split, i.e. expanding window (five-fold, randomly searching throughout the grid of the selected ranges of hyperparameters). The random searching throughout the pre-determined grid was used because of time limitation while receiving promising results. However, if the best-found combination of hyperparameters seemed to be near the bound of the pre-selected range, the bound was expanded in order to validate that the best-found model is close to the local optimum of the searched area.

Other metrics, including mean absolute error (MAE) and R-squared (R^2), were also used to understand other views of the model performance. Nevertheless, mean absolute percentage error (MAPE) was not used in this research. This is because some points of the data, the number of DHF case of each week, are very small. For example, some of each week of some province had zero DHF occurred; therefore, the MAPE score might be very high close to infinity. Thus, it cannot be useful to demonstrate anything about the model performance.

4. Results & Discussion

4.1. First Target Models and Traditional Models

The results of the traditional model against the first target model are demonstrated in Table 3. Some traditional models of some provinces seem to yield promising result since the number of occurring DHF cases usually has a high correlation to the situation of the recent weeks. As a result, the models that forecast the first target seem to lose to the traditional models while forecasting the near future (1 week ahead). However, some of the first target models, for Mukdahan, seem to have promising result in predicting the further future (2–4 weeks ahead). Nevertheless, in some provinces, there are the situations that there are some peaks, which is extremely high compared to the data of the test set, in training dataset. For example, Pattani and Ratchaburi (see Fig. 7 (c) and (f)), that have high peaks of the number of

DHF cases data that cause the first target model to perform inefficiently. Moreover, the unpleasant results of some first target model might occur because the relations between created features and the first target, the smoothed cases, is too slightly for machine learning model to capture and use these insights to make the forecasting accurately. Therefore, the first target models performed worse than the traditional models.

Table 3. The performance of the best models for each province and each forecast step among the first target models and traditional time-series models.[1, 2]

Ahead	Metrics	Province					
		Chiang Rai	Mukdahan	Pattani	Phichit	Ayutthaya	Ratchaburi
1	MSE	281.73 (N)	6.39 (MA)	19.02 (MA)	111.68 (N)	92.11 (MA)	67.50 (N)
	MAE	12.86 (N)	1.84 (MA)	2.80 (MA)	7.88 (N)	6.70 (MA)	6.74 (N)
	R-Squared	0.91 (N)	0.78 (MA)	0.76 (MA)	0.81 (N)	0.75 (MA)	0.67 (N)
2	MSE	609.88 (N)	9.76 (L)	17.07 (MA)	200.34 (N)	113.12 (MA)	98.88 (N)
	MAE	18.10 (N)	2.20 (L)	3.12 (MA)	10.75 (N)	7.71 (MA)	8.00 (N)
	R-Squared	0.81 (N)	0.66 (L)	0.77 (MA)	0.65 (N)	0.69 (MA)	0.53 (N)
3	MSE	1026.66 (N)	11.89 (L)	25.11 (MA)	343.05 (N)	158.08 (MA)	123.87 (N)
	MAE	23.95 (N)	2.43 (L)	3.84 (MA)	13.78 (N)	8.76 (MA)	8.87 (N)
	R-Squared	0.97 (N)	0.58 (L)	0.66 (MA)	0.40 (N)	0.56 (MA)	0.40 (N)
4	MSE	1729.00 (N)	13.60 (L)	32.86 (MA)	539.46 (N)	233.71 (N)	147.14 (N)
	MAE	33.38 (N)	2.67 (L)	4.17 (MA)	17.47 (N)	12.32 (N)	9.26 (N)
	R-Squared	0.45 (N)	0.52 (L)	0.56 (MA)	0.07 (N)	0.36 (N)	0.30 (N)

Table 4. The MSE improvement percentage of the best first target models compared to the best traditional models for each province and forecast step.

Ahead	Province					
	Chiang Rai	Mukdahan	Pattani	Phichit	Ayutthaya	Ratchaburi
1	-19.42%	-6.89%	-19.54%	-55.94%	-44.25%	-17.00%
2	-54.70%	0.50%	-143.92%	-65.87%	-12.07%	-50.31%
3	-72.18%	22.12%	-226.17%	-24.34%	-38.47%	-30.67%
4	-25.19%	34.77%	-171.29%	-2.29%	-9.85%	-3.19%

[1] Grey cells represent that the first target models perform better compared to the traditional ones and light grey cells represent that the traditional models perform better compared to the first target ones.

[2] (N), (MA), and (L) refer to Naïve, Moving Average, and LASSO models respectively.

Additionally, the improvement percentage of MSE when forecasting with the first target models is listed in Table 4. Similar to the result in Table 3, most of the first models showed negative improvement, which means the traditional models outperform, and showed positive improvement only in 2–4 weeks ahead forecasts for Mukdahan. However, among the positive improvement, it can be found that only 3 and 4 weeks ahead forecasts for Mukdahan presented noticeable reduces in MSE, i.e. more than 10% in improvement.

4.2. Second Target Models and Traditional Models

The results of traditional models against the second target models are demonstrated in Table 5. The results showed that the models predicting the second target yielded more promising results. They could forecast the outbreak situations accurately. For example, in Mukdahan cases, the second target models performed better than traditional and first target models. Furthermore, Pattani models showed significant improvement in forecasting 3 and 4 weeks ahead with the second target models. Moreover, Chiang Rai models showed significant improvement for every time horizon with the second target models. It is supposed that the relationship between the features and the second targets can be more detected. Therefore, the improvements in accuracy for some provinces could be discovered. Unfortunately, in Phichit, Ayutthaya, and Ratchaburi cases, that the first target models underperformed the traditional models for every time horizons (ahead 1–4 weeks), the second target models also could not alleviate this issue.

Furthermore, Table 6 presents the improvement in MSE when forecasting with the second target models. The overall improvement was clearly better than those in Table 4 and as the percentages increased. For Mukdahan, one-week ahead forecasting could now provide noticeable reduction in MSE. Fortunately, 1–4 weeks ahead for Chiang Rai as well as 3 and 4 weeks ahead forecasting for Pattani illustrated the major recovery as their performance was substantially augmented. Four weeks ahead forecasting for Phichit and Ratchaburi also gave important decreases in MSE. Nevertheless, the second target models could not perform considerably better than the traditional models for Ayutthaya.

Table 5. The performance of the best models for each province and each forecast step among the second target models and traditional time-series models.[3, 4]

Ahead	Metrics	Province					
		Chiang Rai	Mukdahan	Pattani	Phichit	Ayutthaya	Ratchaburi
1	MSE	237.98 (RF)	5.52 (AB)	17.13 (ET)	111.68 (N)	92.11 (MA)	67.50 (N)
	MAE	10.98 (RF)	1.61 (AB)	2.83 (ET)	7.88 (N)	6.70 (MA)	6.74 (N)
	R-Squared	0.92 (RF)	0.81 (AB)	0.78 (ET)	0.81 (N)	0.75 (MA)	0.67 (N)
2	MSE	543.87 (R)	6.91 (ET)	17.07 (MA)	200.34 (N)	113.12 (MA)	98.88 (N)
	MAE	16.44 (R)	1.80 (ET)	3.12 (MA)	10.75 (N)	7.71 (MA)	8.00 (N)
	R-Squared	0.82 (R)	0.76 (ET)	0.77 (MA)	0.65 (N)	0.69 (MA)	0.53 (N)
3	MSE	847.23 (R)	9.18 (L)	18.20 (AB)	343.05 (N)	155.30 (RF)	123.87 (N)
	MAE	21.27 (R)	2.05 (L)	3.20 (AB)	13.78 (N)	9.34 (RF)	8.87 (N)
	R-Squared	0.72 (R)	0.68 (L)	0.78 (AB)	0.40 (N)	0.56 (RF)	0.40 (N)
4	MSE	1193.56 (R)	10.12 (RF)	22.16 (L)	464.72 (RF)	233.71 (N)	116.05 (AB)
	MAE	25.65 (R)	2.02 (RF)	3.37 (L)	14.91 (RF)	12.32 (N)	8.56 (AB)
	R-Squared	0.61 (R)	0.65 (RF)	0.73 (L)	0.18 (RF)	0.36 (N)	0.47 (AB)

[3] Dark grey cells represent that the second target models perform better compared to the traditional ones and light grey cells represent that the traditional models perform better compared to the second target ones.
[4] (N), (MA), (RF), (L), (AB), (R) and (ET) refer to Naïve, Moving Average, Random Forest, LASSO, AdaBoost, Ridge Regression, and Extra-Trees models respectively

Table 6. The MSE improvement percentage of the best second target models compared to the best traditional models for each province and forecast step.

Ahead	Province					
	Chiang Rai	Mukdahan	Pattani	Phichit	Ayutthaya	Ratchaburi
1	15.53%	14.56%	9.95%	-24.36%	-7.22%	-16.46%
2	10.82%	29.51%	-14.54%	-16.86%	-2.84%	-12.28%
3	17.48%	39.84%	27.49%	-1.31%	1.76%	-5.50%
4	30.97%	51.50%	32.55%	13.85%	-0.62%	21.13%

4.3. Further Discussion

In addition, some provinces, e.g. Phichit, Ayutthaya, and Ratchaburi, that the machine learning models could not perform better than traditional models in both initiatives were explored more. It was found that the data in 2018 (test data) for Phichit (see Fig. 7 (d)) are dramatically different from the data for model training and the pattern of incidences shows no sign of reappearance for both first and second target. Consequently, the models struggle to forecast unseen patterns and only the second target model for 4-week ahead forecasting achieves better performance than the traditional.

For Ayutthaya (see Fig. 7 (e)), the dengue situation seems to differ along the time. The actual number of DHF cases that occurs within each week appears not to repeat itself. Thereby, the first and second targets are fluctuating and hard to predict. As a result, the traditional models perform better than the others. Only few second target models that can perform slightly better than traditional model for the 3 weeks ahead forecasting. This might be because the effect of used features gives some valuable insights when forecasting; therefore, the performance of models is better than the traditional models that only rely on the previous data of itself.

Other reasons to explain the unpleasant results might be a surveillance system problem, short forecasting intervals, and unknown relationship between the data. As the number of dengue incidences in this research mainly bases on the surveillance system, with an inconsistent report, the data could be fuzzier than an expectation. Especially in the provinces with a small population like Phichit, variation from the system could significantly affect the pattern of an outbreak record and made the trend less predictable. Moreover, the nature of patients within each province—whether they usually go to the hospital right after they got sick or they usually to rest at home until the symptoms go bad—might affect the nature of the surveillance record and outbreak nature pretty much. Furthermore, there might not be a strong correlation between created features and the outbreak pattern. Thereby, these issues should be explored more in the research. And other feature should be tried to use as input variables for these provinces.

The short period ahead forecasting usually experiences noisy data as the variation is separated into periods. The task becomes difficult as the forecasting model must detect and capture trend that lies under the fluctuation and if the nature of the data is itself considerably fluctuated, it becomes even more problematic. With this characteristic, some sets of data can be explained well using only the simplest and easiest forecasting model, Naïve forecasting, since it is nearly impossible to extract the relation between data from the pool of variation. Therefore, the most recent record appears to be the best forecast.

Although the pattern of some provinces, e.g. Ratchaburi (see Fig. 7 (f)), seems to be forecasted acceptably with the second target model as Pattani do, the traditional models dominated the proposed models. The possible reason might from insufficient data relationship. Despite the current feature engineering process, the machine learning models could need more data sources to learn and find the relationship. With more sources, the models are believed to improve their performance. Moreover, due to Google Trends and meteorology data limitation, the overall data range used in this research (about 4 years and a half) is relatively low compared to other researches. Increasing in a data pool can enhance the power of the models, both traditional and machine learning.

5. Conclusion

This research aims to integrate the analysis of meteorology and Google Trends data to forecast the number of DHF incidences of provinces, representing major outbreak trends in Thailand. In order to forecast the very near situation of DHF outbreak of noisy data in some provinces, the models that depend only on recent data, i.e. Naïve and Moving Average models, seem to be good at it. However, for these provinces, if the further time-horizon situations are required to be forecasted, machine learning models that learn the relation between the pattern of related features and the number of the DHF cases yield the better performance.

In some provinces, whose outbreak pattern usually repeats itself and varies within the same-limited range along the time-horizon, the machine learning models seem to be good for these cases. Furthermore, by

forecasting the difference of the situation—the difference of smoothed values of the number of the cases like the performed task in the second target models—seem to be better than forecasting the number of occurring cases, like the task performed in the first target models.

For further works, it is recommended to provide both traditional and complex forecasting models with practical and sufficient data sources. Specifically, the process of feature creation for machine learning can be researched further to craft more beneficial and dengue-related features which also improve forecasting accuracy. Moreover, there are the areas that should be explored more in order to increase the accuracy of forecasting. For example, the micro-components of time-series data such as, trends and seasonal effect, should be extracted to be target values in order to gain the models that yield better performance.

References

[1] The Ministry of Public Health Thailand, The Bureau of Vector-Borne Disease, "Official dengue situation report," Nonthaburi, 2018.

[2] E. P. Pliego, J. Velázquez-Castro, and A. F. Collar, "Seasonality on the life cycle of Aedes aegypti mosquito and its statistical relation with dengue outbreaks," *Applied Mathematical Modelling*, vol. 50, pp. 484–496, 2017.

[3] N. Rachata, P. Charoenkwan, T. Yooyativong, K. Chamnongthal, C. Lursinsap, and K. Higuchi, "Automatic prediction system of dengue haemorrhagic-fever outbreak risk by using entropy and artificial neural network," in *2008 International Symposium on Communications and Information Technologies*, 2008.

[4] S. A. Lauer, K. Sakrejda, E. L. Ray, L. T. Keegan, Q. Bi, P. Suangtho, S. Hinjoy, S. Iamsirithaworn, S. Suthachana, Y. Laosiritaworn, D. A. Cummings, J. Lessler, and N. G. Reich, "Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010–2014," in *Proceedings of the National Academy of Sciences*, 2018, vol. 115, no. 10.

[5] S. Yang, S. C. Kou, F. Lu, J. S. Brownstein, N. Brooke, and M. Santillana, "Advances in using Internet searches to track dengue," *PLOS Computational Biology*, vol. 13, no. 7, 2017.

[6] "Dengue and severe dengue," *World Health Organization* [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[7] S. Sangtharathip and T. Boonma, *Ordinary Edition of Dengue Fever Education*, 2nd ed. Nonthaburi: The Publisher of the Agricultural Co-operative Federation of Thailand, 2002.

[8] T.-H. Chen, Y.-C. Chen, J.-L. Chen, and F.-C. Chang, "Flu trend prediction based on massive data analysis," in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 2018.

[9] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[10] X. Zhou, J. Ye, and Y. Feng, "Tuberculosis surveillance by analyzing Google Trends," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 8, pp. 2247–2254, 2011.

[11] G. J. Milinovich, S. M. R. Avril, A. C. A. Clements, J. S. Brownstein, S. Tong, and W. Hu, "Using internet search queries for infectious disease surveillance: Screening diseases for suitability," *BMC Infectious Diseases*, vol. 14, no. 1, 2014.

[12] P. Guo, T. Liu, Q. Zhang, L. Wang, J. Xiao, Q. Zhang, G. Luo, Z. Li, J. He, Y. Zhang, and W. Ma, "Developing a dengue forecast model using machine learning: A case study in China," *PLOS Neglected Tropical Diseases*, vol. 11, no. 10, 2017.

[13] W. Anggraeni and L. Aristiani, "Using Google Trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia," in *2016 International Conference on Information & Communication Technology and Systems (ICTS)*, 2016.

[14] G. Bonaccorso, *Machine Learning Algorithms: Reference Guide for Popular Algorithms for Data Science and Machine Learning*. Packt Publishing., 2017.

[15] T. M. Carvajal, K. M. Viacrusis, L. F. T. Hernandez, H. T. Ho, D. M. Amalin, and K. Watanabe, "Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines," *BMC Infectious Diseases*, vol. 18, no. 1, 2018.

[16] J. Ong, X. Liu, J. Rajarethinam, S. Y. Kok, S. Liang, C. S. Tang, A. R. Cook, L. C. Ng, and G. Yap, "Mapping dengue risk in Singapore using Random Forest," *PLOS Neglected Tropical Diseases*, vol. 12, no. 6, 2018.

[17] D. Jiang, M. Hao, F. Ding, J. Fu, and M. Li, "Mapping the transmission risk of Zika virus using machine learning models," *Acta Tropica*, vol. 185, pp. 391–399, 2018.

[18] S. Bouktif, A. Fiaz, A. Ouni, and M. Serhani, "Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches," *Energies*, vol. 11, no. 7, p. 1636, 2018.

[19] Y. Shi, X. Liu, S.-Y. Kok, J. Rajarethinam, S. Liang, G. Yap, C.-S. Chong, K.-S. Lee, S. S. Tan, C. K. Y. Chin, A. Lo, W. Kong, L. C. Ng, and A. R. Cook, "Three-month real-time dengue forecast models: An early warning system for outbreak alerts and policy decision support in Singapore," *Environmental Health Perspectives*, vol. 124, no. 9, pp. 1369–1375, 2016.

[20] M. Mohebbi, D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi and S. Kumar, "Google Correlate Whitepaper," 2011.

[21] S. Deb, C. M. L. Acebedo, G. Dhanapal, and M. C. H. Chua, "An ensemble prediction approach to weekly Dengue cases forecasting based on climatic and terrain conditions," *Journal of Health and Social Sciences*, vol. 2, no. 3, pp. 257–272, 2017.

[22] Y. Nagao, U. Thavara, P. Chitnumsup, A. Tawatsin, C. Chansang, and D. Campbell-Lendrum, "Climatic and social risk factors for Aedes infestation in rural Thailand," *Tropical Medicine and International Health*, vol. 8, no. 7, pp. 650–659, 2003.

[23] L. Lu, H. Lin, L. Tian, W. Yang, J. Sun, and Q. Liu, "Time series analysis of dengue fever and weather in Guangzhou, China," *BMC Public Health*, vol. 9, no. 1, 2009.

Akeamorn Puengpreeda, photograph and biography not available at the time of publication.

Suphawit Yhusumrarn, photograph and biography not available at the time of publication.

Surapong Sirikulvadhana, photograph and biography not available at the time of publication.