An Analytics Prediction Model of Monthly Rainfall Time Series: Case of Thailand

Wattana Punlumjeak, Jedsada Arunrerk, Nachirat Rachburee
Department of Computer Engineering, Faculty of Engineering,
Rajamangala University of Technology Thanyaburi, Pathumthani, Thailand.
wattana.p@en.rmutt.ac.th

***Abstract*—Rainfall prediction is regarded as a challenging task in an agricultural country like Thailand. A time series data especially rainfall and temperature needs analytics technologies to return a valuable knowledge. It has been recognized that a high accuracy of rainfall prediction model will be helpful for agriculturist and water management. The study area of this research is located in Thailand, which the daily rainfall and temperature time series data collected from five regions of Thailand were taken by Meteorological Department of Thailand from years 2000 to 2015. In this research, analytics method is proposed in the preprocessing steps, which are composed of data cleansing and data transform. Principal component analysis in feature selection step and weighted moving average are applied. In the prediction modeling, support vector regression (SVR) and artificial neural network (ANN) are employed. The results of the experiment showed the comparison of overall accuracy between ANN and SVR in five data sets over the area of study. The results of the experiment showed that the two prediction models gave a high overall accuracy, although SVR plays an important advantage in less computational time than ANN. This experiment is extremely useful not only as the most effective way to manage the amount of rainfall in water management for Thai agriculturist, but the proposed model can also become a representative in the monthly rainfall prediction model used in Thailand.**

***Index Terms*—Monthly Rainfall Prediction; Time Series Prediction; Support Vector Regression; ANN; Moving Average.**

I. INTRODUCTION

Thailand is an agricultural cornucopia country [1], which is one of the world's largest exporters of rice [2]. Rice and other agricultural crops production are most likely rely on weather and climate. One of the key factors in crop production is an effective water management. Rainfall is one of the precipitation in natural climatic phenomena in hydrology and meteorology. The amount of rainfall affects not only the quality and quantity of agricultural product, but also results in bad situations like flood, drought and traffic-jams in big cities. An accurate rainfall prediction has a great potential to prevent some damages caused by natural disaster.

The prediction of rainfall time series depends on many complex factors, which are affected by global warming, temperature, wind, monsoon, relative humidity, and others. Predicting rainfall is regarded as a challenging task due to its characteristics as well as its non-linear and deterministically chaotic system. Although many researchers have conducted research in this topic for a long time, there is still a need for sophisticated modeling for more accurate prediction.

A time series data is a collection of data or observations, which are taken at equal intervals and recorded over a period of time [3]. In real world, time series is found in many applications, including stock market price, click-stream processing, and sensor data. The study and publications of time series analysis was started from Yule. After that, it was carried out by many statisticians like Waller, Box and Jenkins [4-6] and so on. The variation components represented in time series are trend, seasonal, cyclical and irregular pattern. An analytic in time series concerns a specific quantity varied over time. The windowing function is one of the ways to transform any time series to cross-sectional format or attribute data set before feeding the data into a learning model. The window size determines how many attributes are created. Meanwhile, the horizon determines how far the forecast and step size is determined to advance the window. The windowing process with window size will predict the next unknown value. The idea can be described as shown in Figure 1.

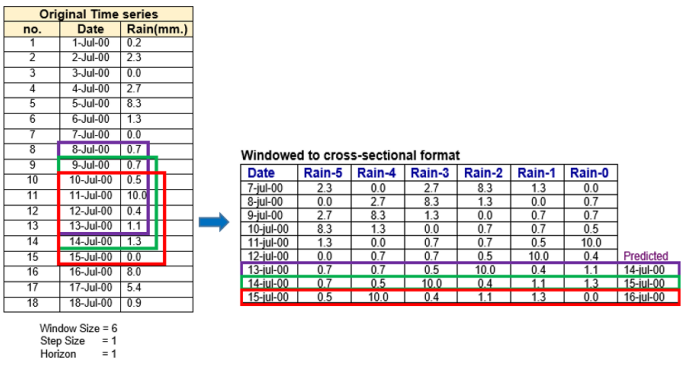

Figure 1: The Idea of Windowing Function

The goal of this research is to develop an efficient monthly rainfall prediction model, which is a representative rainfall prediction in Thailand. Because of the many complex factors which affect the rainfall prediction, the temperature is chosen to be one factor which involves in this experiment. This paper is organized as follows: first, an introduction and then related works will be presented in Section II. The study area and data sets in the proposed method is described in Section III. Section IV shows the experimental results and finally, Section V provides the conclusion of this research.

II. Related Work

We begin this section with the methodology and work related to the method in the preprocessing step and modeling step as follow:

Principal Component Analysis (PCA) is a classical multivariate factor analysis used in feature reduction approach to reduce the dimensions in a feature space. The first principal component with the highest variance is the linear combination. The following is the linear combination with maximum variance in an orthogonal direction to the first principal component. The most relevant combination of the features is projected to the eigenvector with the highest eigenvalue. This means that the highest eigenvalue is significant to identify each feature. The linear combination can be expressed in the kth dimension of the projected feature vector *u* is

$$u = \sum_j w_j^k x_j \quad (1)$$

Many researchers used PCA in the preprocessing step to reduce the dimension of multivariate and find a significant feature subset [7-9]. In time series research, such as EEG data from meditators, they also used PCA to get the distinct pattern from the statistic feature in EEG signal [10]. Further, satellite Image Time series applied the PCA to reduce their feature set and improve the multi-temporal classification [11].

Moving average (MA) is a basic method used in data smoothing in the preprocessing step. MA is used to filter a white noise or short-term fluctuations from the data. In time series data, MA aims to estimate the trend of the data. Simple moving average is applied to smoothen the data by computing the average of the past, which can be defined as:

$$M = \frac{1}{N}\sum_{i=1}^{N} V_i \quad (2)$$

where *M* is the average value, *V* is actual value and *N* is a number of periods in the moving average.

Weighted moving average (weighted MA) filters a white noise by giving a weighted value. A weighted MA can be calculated by adding the weighting factor, which implies in different ways such as Gaussian function, Hann function, and etc. To compute the forming the average of the past can be defined as:

$$M = \frac{1}{N}\sum_{i=1}^{N} W_i * V_i \quad (3)$$

where *M* is the average value, *V* is the actual value, *W* is weighting factor and *N* is a number of periods in a weighting group.

One of the most popular approaches used in many time series researches are the Moving Average. In Tamil Nadu India, MA is used in preprocess to smoothen the rainfall time series [12]. In addition, rainfall time series and the other researches used MA to reduce the noise from the data as well as estimate the trend of the data [13, 14].

Artificial Neural Network (ANN) is a non-linear methodology, which is also called as the multi-layer perceptron. This method is inspired by the biological central nervous systems in a human brain. The artificial neural network consists of three main layers: an input layer, hidden layer, and output layer. The architectural graph is shown in Figure 2. Each layer is composed of neurons, which receives signals through synapses located on the dendrites of the neuron. They are interconnected with each other when the signals are received, which is called as synaptic weight that is strong enough or it meets a threshold. ANN has been popular and widely acknowledged as a method to classify the complex data set in many researches. ANNs have used in rainfall-runoff time series forecasting in Thailand [15, 16] and it also used to forecast rainfall in Japan [17] and China [18, 19].

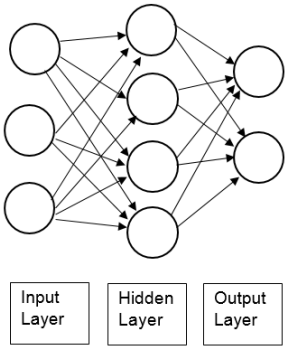

Figure 2: The Artificial Neural Network Architecture

The support Vector Regression (SVR) is known as support vector machine, which is developed to solve regression problem. SVR transforms the non-linear regression to linear regression in the high dimension feature space by using ε-insensitive loss function introduced by Vapnik in 1995. The input of non-linear is mapped onto m-dimensional feature space. The linear model is constructed as follow:

$$f(x) = \omega \cdot \varphi(x) + b \quad (4)$$

To reduce the complexity of the model, SVR is formulated as the minimization of the following functional.

$$\frac{1}{2} \| \omega \|^2 + C \sum_i^n (\xi_i + \xi_i') \quad (5)$$

$$s.t. ((\omega \cdot \varphi(x_i) + b) - y_i \leq \varepsilon + \xi_i, i = 1,2,\dots,n,$$
$$y_i - ((\omega \cdot \varphi(x_i) + b) \leq \varepsilon + \xi_i, i = 1,2,\dots,n,$$
$$\xi_i' \geq 0, i = 1,2,\dots,n,$$

This optimization problem is based on Lagrange function and the dual problem and its solution is given by:

$$\frac{1}{2}\sum_{i,j=1}^{n}(a_i' - a_i)(a_j' - a_j)K(x_i, x_j) + \varepsilon\sum_{i=1}^{n}(a_i' + a_i) - \sum_{i=1}^{n} y_i(a_i' - a_i)$$
$$s.t. \sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0,$$
$$0 \leq \alpha_i^{(*)} \leq C, i = 1,2,\dots,n \quad (6)$$

SVR is widely used in non-linear regression problems. The research [20] showed that SVR can perform a higher accuracy and computational time performance of rainfall

forecasting model in Bangladesh. SVR is proposed based on the wavelet kernel function to ensemble forecasting with radial basis function and neural network and obtain a greater accuracy on monthly rainfall in Guangxi, China [21].

III. THE PROPOSED METHOD

The main ideas of our work can be depicted in Figure 3, as follows.

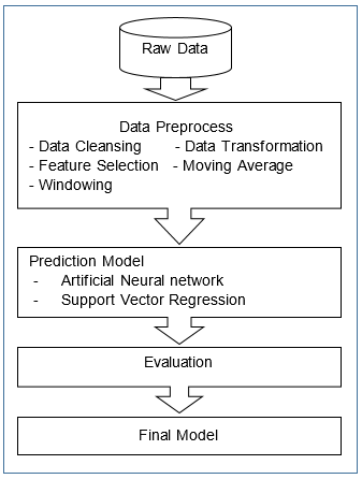

Figure 3: The Proposed Model

The overview of the proposed model consists of five main steps. First, we described the source of the raw data and their characteristics. Secondly, we conducted the data preprocessing stage aimed to prepare the satisfied data as the input data to the prediction model in the next step. Thirdly, we used two popular methods to predict the amount of rainfall. Fourth, we compared and evaluated the model. Finally, the rainfall prediction model for Thailand was proposed.

In this research, we used the rainfall time series data compiled by Meteorological Department of Thailand from year 2000 to 2015. The study area is located in Thailand with the latitude 5°37'- 20°28' north, longitude 97°21'-105°37' east. In relation to the large amount of raw data, our interest was 29,220 records collected from five regions of Thailand as shown a Figure 4. Each of the regions illustrates the name and code of the station name in which the data were collected as shown Table 1.

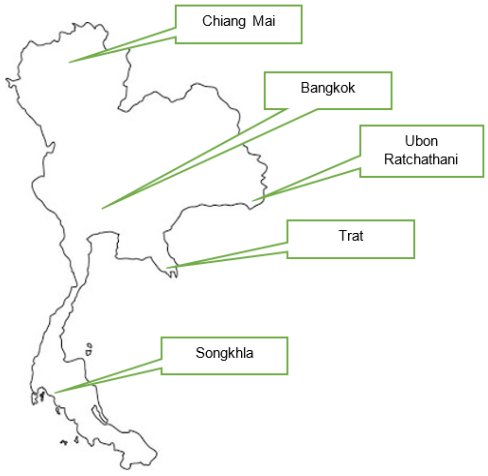

Figure 4: The Area of Study

Table 1
The Source of Data

Region	Station name	Station code
Northern	Chiang Mai	D1
Northeastern	Ubon Ratchathani	D2
Central	Bangkok	D3
Eastern	Trat	D4
Southern	Songkhla	D5

At first, the actual time series data we considered in the day period was composed of two components: rainfall and temperature. Thus, five attributes included in the raw data set were date, rain, minimum temperature, average temperature, and maximum temperature. Figure 5 shows an example of the actual rainfall in a one-day period in Ubon Ratchathani station (D2) from year 2000-2015.

The duration of the data set used is sixteen years (2000-2015). However, thirteen years (2000-2012) of data sets or about 70% of data set were selected to be included in the training data set. Meanwhile data sets within the duration of three years (2013-2015) or about 30% was set as the testing data set.

The proposed method was experimented in RapidMiner Studio version 7.2.001, on a Window 8.1 Pro operating system with Intel core i5-2400 processor and 12GB of RAM.

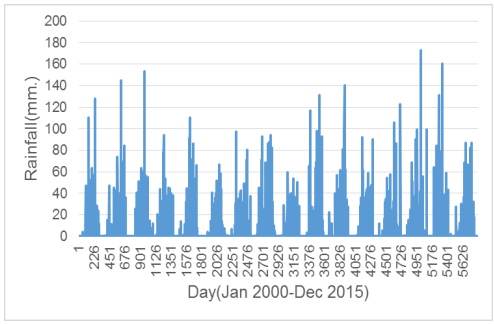

Figure 5: Example of Actual Rainfall of Ubon Ratchathani

IV. EXPERIMENTAL RESULT

In the data preprocessing steps, the exploration of the data involves the way we used to examine the raw data. When we looked insight at the deck of the data, the outlier is not shown, but we found that some data in some periods of time had been missing. Data cleansing was the first method we used to handle the missing data, although many approaches were chosen as a choice to handle the data, such as listwise data deletion, mean substitution, hot deck imputation, and so on. In this case, hot deck imputation was selected by filling the missing data with a posterior data in the same time series. After cleansing the data, data transformation was processed based on six attributes. The first attribute was transformed from day to month. We calculated a summation of rainfall monthly for the second attribue. The third attribue was based on the maximum value of rainfall in each month. For the fourth and the fifth attribute, we found a maximum and minimum temperature in every month. Lastly, the sixth attribute is the average temperature in monthly. Hence, the original time series data, which were collected in the day period, had been transformed to monthly period time series data. Therefore, the example of the new data set for each station had six attributes: month (atr1), sum_rain (atr2),

max_rain (atr3), max_tmp (atr4), min_tmp (atr5), and avg_tmp (atr6), as shown as Table 2.

Table 2
The Example of the Data in Data Set

ATR1	ATR2	ATR3	ATR4	ATR5	ATR6
Jan-00	0	0	10.3	33.2	23.10
Feb-00	37.4	45.6	12.5	36.5	24.39
Mar-00	21.8	41.2	16.3	37.2	26.27
Apr-00	107.7	83.5	20.8	37.8	28.89

The method of the feature selection is to find a significant feature subset from feature space. In this research, PCA is used to find the correlation of temperature (max_tmp, min_tmp, and avg_tmp) with the amount of rainfall and to reduce the computational time. Table 3 show the eigenvalue of each feature in each station after PCA is employed.

Table 3
The Eigenvalue of Each Feature

Feature	D1	D2	D3	D4	D5
min_tmp	0.846	0.889	0.864	0.974	0.191
avg_tmp	0.445	0.419	0.393	0.227	0.365
max_tmp	0.293	0.182	0.313	-0.014	0.911

From the experiment in the feature selection, min_tmp returned the highest eigenvalue. Thus, two attributes (max_tmp and avg_tmp) had been removed from the data set. Thus, the attributes in the new data set were composed of the month, sum_rain, max_rain, and min_tmp as shown the Table 4.

Table 4
The Example of the New Data Set

MONTH	SUM_RAIN	MAX_RAIN	MIN_TMP
Jan-00	0	0	33.2
Feb-00	37.4	45.6	36.5
Mar-00	21.8	41.2	37.2
Apr-00	107.7	83.5	37.8

After that, the weighted moving average with Hann function was employed to smooth the data. And then, the windowing function was used. The parameter that we used in the windowing operator are: window size = 3, step size = 1, and horizon = 1. The windowing operator transformed the time series data into a cross-sectional data set and fed this data into the learning algorithm. At this point, two learning methods of prediction were applied: backpropagation neural network and support vector regression with polykernel function. In RapidMiner software, the hidden layer in ANN was set to one layer and the number of neurons in the hidden layer was computed from:

$$((number\ of\ attributes + number\ of\ classes)/2) + 1 \quad (7)$$

The other parameters were learning rate = 0.3, momentum = 0.2 and training cycle = 500. In this research, the goal is to predict the amount of rainfall in one step ahead, that is, one month. Thus, the attribute "month" was set as ID, max_rain, and min_tmp was a regular attribute, and sum_rain was set to be a label attribute. The training data set and the testing data set in the five areas were done in the same process as in the conduct of the experiment.

The experiment compared the performance of each prediction techniques by evaluating the accuracy of the predictors. The computational time in a second is shown in Table 5. The overall accuracy of ANN and SVR for five data set is shown in Table 6 and the results are depicted graphically in Figure 6.

Table 5
The Computational Time (second)

	D1	D2	D3	D4	D5
ANN	9.0	9.0	9.0	9.0	9.0
SVR	1.0	1.0	<1.0	1.0	<1.0

Table 6
The Overall Accuracy

	D1	D2	D3	D4	D5
ANN	0.879	0.863	0.903	0.907	0.952
SVR	0.895	0.895	0.907	0.899	0.935

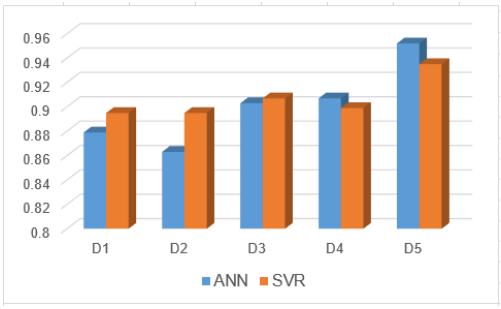

Figure 6: The Comparative Result of ANN and SVR

V. Conclusion

This research presented an analytics prediction model applied to monthly rainfall time series in Thailand. The study is done with five data sets and compared the overall accuracy between ANN and SVR model. In terms of accuracy, SVR learning model takes a great advantage in comparison to ANN, especially in the north and central of Thailand. Based on the data drawn from the south of Thailand, ANN learning model produced higher accuracy over SVR. In relation to the goal of the research, SVR plays an efficient model, which represents a monthly rainfall prediction model in Thailand in terms of less computational time than ANN. This proposed model provides significant benefits to the Thai agriculturist and the water management responsibilities.

In future, it is aspired to discover an alternate method to improve the prediction quality and obtain a greater accuracy for the monthly rainfall prediction.

Acknowledgments

Thanks to Meteorological Department of Thailand for providing the rainfall and temperature data.

References

[1] http://thailand.prd.go.th/ebook2/kitchen/ch1.html (accessed August 18, 2016).
[2] http://www.fao.org/docrep/010/ah994e/AH994E01.htm (accessed August 18, 2016).
[3] Chatfield C. The analysis of time series: an introduction, CRC press. 2016.
[4] Mills, T. C. The foundations of modern time series analysis, Palgrave Macmillan. 2011.
[5] GE Box, GM Jenkins, GC Reinsel, GM Ljung. Time series analysis: forecasting and control, John Wiley & Sons. 2015.
[6] Rojas I, Pomares H. Time series analysis and forecasting: selected contributions from the ITISE Conference. Contributions to statistics.

2016.

[7] N. A Biswas, F. M Shah, W. M. Tammi, and S.Chakraborty. FP-ANK: An improvised intrusion detection system with hybridization of neural network and K-means clustering over feature selection by PCA. 18th International Conference on Computer and Information Technology (ICCIT), (2015) 317-322.

[8] I. Siegert, R. Böck, A.Wendemuth, and B.Vlasenko (2015, September). Exploring dataset similarities using PCA-based feature selection. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), (2015), 387-393.

[9] N.Shaltout, M.Moustafa, A.Rafea, A. Moustafa, M.ElHefnawi. Comparing PCA to information gain as a feature selection method for Influenza-A classification. ICIIBMS 2015, International Conference on Intelligent Informatics and Biomedical Sciences, Okinawa, Japan, (2015), 279-283.

[10] Shaw, L., and Routray, A. Statistical features extraction for multivariate pattern analysis in meditation EEG using PCA. In Student Conference (ISC), 2016 IEEE EMBS International. Orlando, Florida, USA, (2016) 1-4.

[11] S. Réjichi, F. Chaabane. Feature extraction using PCA for VHR satellite image time series spatio-temporal classification. 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), (2015) 485-488.

[12] M. Nirmala. Integrated soft computing approach for modeling rainfall prediction in Tamilnadu. 9th International Conference on Intelligent Systems and Control (ISCO), (2015) 1-6.

[13] F. Nhita, D.Saepudin, U. N. Wisesty. Comparative Study of Moving Average on Rainfall Time Series Data for Rainfall Forecasting Based on Evolving Neural Network Classifier. 3rd International Symposium on Computational and Business Intelligence (ISCBI), (2015) 112-116.

[14] W. Sulandari, Y. Yudhanto. Forecasting trend data using a hybrid simple moving average-weighted fuzzy time series model. 2015 International Conference on Science in Information Technology (ICSITech), (2015) 303-308.

[15] S. Areerachakul, P. Junsawang. Rainfall-Runoff relationship for streamflow discharge forecasting by ANN modelling. 2014 World Congress on Sustainable Technologies (WCST), (2014) 27-30.

[16] J. Kajornrit, K. W. Wong, C. C. Fung, Y. S Ong. An integrated intelligent technique for monthly rainfall time series prediction. 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), (2014) 1632-1639

[17] S. S. Monira, Z. M. Faisal, H. Hirose. A neural network ensemble incorporated with dynamic variable selection for rainfall forecast. 12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), (2011) 7-12.

[18] J. Wu. An Effective Hybrid Semi-parametric Regression Strategy for Artificial Neural Network Ensemble and Its Application Rainfall Forecasting. 2011 Fourth International Joint Conference on Computational Sciences and Optimization (CSO), (2011) 1324-1328.

[19] Z. L. Wang, H. H. Sheng. Rainfall prediction using generalized regression neural network: case study Zhengzhou. 2010 International Conference on Computational and Information Sciences (ICCIS), (2010) 1265-1268.

[20] N. Hasan, N. C. Nath, R. I. Rasel. A support vector regression model for forecasting rainfall. 2015 2nd International Conference on Electrical Information and Communication Technology (EICT), (2015) 554-559.

[21] L.Wang, J. Wu. Application of hybrid RBF neural network ensemble model based on wavelet support vector machine regression in rainfall time series forecasting. 2012 Fifth International Joint Conference on Computational Sciences and Optimization (CSO), (2012) 867-871.