Prediction of Reference Evapotranspiration with Missing Data in Thailand

Kitsuchart Pasupa
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand
Email: kitsuchart@it.kmitl.ac.th

Ek Thamwiwatthana
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand
Email: nupippo@gmail.com

***Abstract*—Artificial Neural Networks (ANNs) has been used in prediction of reference evapotranspiration for a recent decade. Its performance is competitive to a widely used method the so-called "Penman-Monteith" method. In this study, we aim to estimate the crop evapotranspiration by ANNs from climatic data in Thailand and compare the performance with the Penman-Monteith method. As missing data is inevitable, we also included the missing data situation into the study. This can be solved by expectation-maximization algorithm. The accuracy of the prediction decreases when the amount of missing values increases. Furthermore, we exploit the feature selection in the study. It shows that sunshine duration is the most important feature followed by temperature and wide speed, respectively.**

***Index Terms*—reference evapotranspiration; missing data; neural network; feature selection**

I. Introduction

Agriculture is an important mean of food production, and food is a basic necessity of human life. Nowadays, information and communication technology (ICT) begins to play a role in agriculture in order to reduce the cost and time of production. This is the so-called "precision farming". It can influence the agricultural market competitiveness. Moreover, ICT can be utilized to water supply management in order to cope with drought and rain risks.

Thailand is an agricultural country. According to the recent summary of the labor force survey in Thailand reported by National Statistical Office, Ministry of Information and Communication Technology in March 2013, 35.15% of Thailand's labor force is employed in agriculture [1]. Moreover, Thai agriculture is very competitive and highly required water supply management, therefore precision farming is needed.

One of the approaches to water supply management is to schedule irrigation for high water use efficiency which can be done by estimating the crop evapotranspiration (ET_c). ET_c can be calculated from a multiplication between crop coefficient and reference evapotranspiration (ET_0). The value of ET_0 can be measure from Lysimeter. This method is very expensive and can only be utilized by well-trained person, therefore there are many approaches introduced to indirectly estimate the value of ET_0 namely Penman-Monteith equation [2], Hargreaves equation [3], etc. Penman-Monteith equation is one of the most globally used method for ET_0 estimation [2]. It can achieve the highest accuracy among the other methods when there is enough data.

Artificial intelligence techniques were recently applied to estimate the value of ET_0 e.g. artificial neural network [4], [5], [6], [7]. Artificial neural networks (ANNs) are trained from data collected by sensors which are installed in the crop fields. The data consists of pairs of input objects (e.g. temperature, humidity, solar radiation, wind speed) and desired outputs (i.e, ET_0). Once the model has trained, it can predict the value of ET_0 from the input data collected from the sensors. Evapotranspiration is very complex and nonlinear model. However, only a single hidden layer ANN is enough to mimic the model together with six inputs which are minimum and maximum temperature, minimum and maximum relative humidity, wind speed, and solar radiation [4]. ANN was also successfully applied to estimate evaporation rate with air temperature, humidity, wind velocity and solar radiation [5].

Problems arise when a size of the crop field is very large. The cost of the sensor installations will be increased according to the size and some area may not be able to install the devices. Therefore, many researchers aim to analyze and report of which input features to be used in the prediction. Hence the cost and the number of devices could be reduced. Moreover, the computational time is reduced. There is an evidence that using inputs of air temperature, wind speed, humidity and solar radiation in Malaysia gives the highest accuracy of the prediction [5]. However, the selection of variables depends on the area where the experiment is conducted such as USA [4] Malaysia [5], Brazil [6], and Burkina Faso [7]. To the best of the authors' knowledge, there is no report about this matter in Thailand.

In real-world implementation, data loss is inevitable. Missing data can occur for many reasons: storage or sensor mechanisms are malfunction, system fails to respond when the data is transmitted. Therefore, it is not able to estimate the current reference evapotranspiration. We can use classic approaches to solve this problem. This can be done by mean substitution or list-wise deletion. However, these approaches can reduce the prediction accuracy. One of the popular statistical methods used for estimating missing values is Expectation-Maximization (EM) algorithm. It aims to compute the maximum likelihood estimation in the presence of missing data [8].

In this paper, we aim to predict the reference evapotranspiration by artificial intelligence techniques i.e. ANNs, linear regression (LR). We also study the relevance of the features using LR. Then we select sets of features and use them to train models. Moreover, we also apply EM algorithm to solve the missing data problem. The data used in the analysis is collected from two different provinces in Thailand.

The paper is organized as follows. Section II outlines the Penman-Monteith method, ANNs and EM algorithm. Section III explains experimental framework which includes the data collection, experimental results and discussion.

II. Methodologies

A. Penman-Monteith Method

It is a standard method which has been used for modeling evapotranspiration at Food and Agriculture Organization (FAO) of the United Nations [2]. The Penman-Monteith equation can be defined as,

$$ET_0 = \frac{0.408\Delta(r_n - G) + \gamma\frac{900}{T+273}U_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34U_2)} \quad (1)$$

where ET_0 is the reference evapotranspiration (mm/day), r_n is the net radiation at the crop surface (MJ·m^2/day), G is the soil heat flux density (MJ·m^2/day). T and U_2 are the mean daily air temperature (°C) and the wind speed (m/s) at 2 meters height, respectively. e_s and e_a are the saturation vapor pressure (kPa) and actual vapor pressure (kPa), respectively. Δ is the slope vapor pressure curve (kPa/°C), and γ is the psychrometric constant (kPa/ °C).

B. Artificial Neural Network

Artificial Neural Network (ANN) is one of the nonlinear statistical data modeling technique. They aim to mimic the human brain functions and consist of weighted artificial neurons (nodes) in a layer. They can determine complex relationships between inputs and outputs of the data. A typical ANNs might have a hundred neurons and many layers. A basic structure of neural networks is shown in Fig. 1.

Consider the vector of m-dimensional inputs $\mathbf{x}=[x_1, x_2, ..., x_m]$ which have weight w_i associated with each input in a neuron. An output of the neuron, u, is a linear combination of inputs and weights:

$$u = \sum_{i=1}^{m} w_i x_i \quad (2)$$

The output of the neuron is fed into the activation function:

$$\hat{y} = \phi(u) \quad (3)$$

where $\phi(\cdot)$ is a sigmoid function:

$$\phi(u) = \frac{1}{1 + e^{-u}} \quad (4)$$

The feed forward ANNs can be adjusted in order to improve the performances i.e. number of nodes and activation function. Moreover, back propagation algorithm was applied in ANNs in order to reduce the error and increase robustness of the algorithm [9]. After the feed forward process, the error signals will be propagated backward through the network in order to adjust the weights in each node.

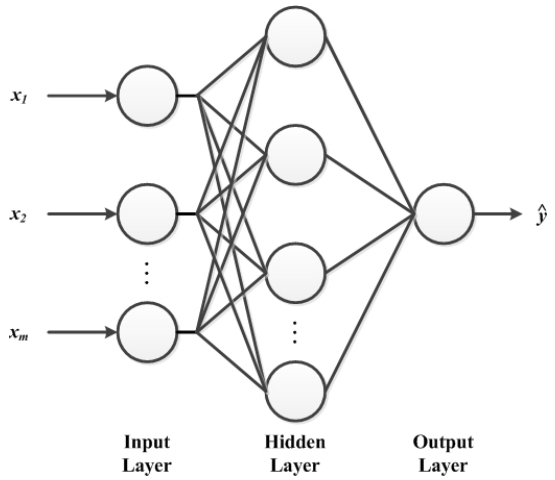


Fig. 1. Neural networks with single hidden layer.

C. Expectation-Maximization Algorithm

Expectation-Maximization (EM) algorithm is a widely used statistical technique to handle the incomplete data problem. Assume that $\theta^{(t)}$ is the t^{th} step in an iterative procedure. $\mathbf{X}$ is a set of observed data. $\mathbf{Z}$ is treated as missing values. The EM algorithm consists of two iterative steps which are as follows [8]:

- The expectation-step (E-step): In this step, it aims to calculate the expected value of the complete log-likelihood which can be calculated by,

$$Q(\theta|\theta^{(t-1)}) = E[\log p(\mathbf{X}, \mathbf{Z}|\theta)|\mathbf{X}, \theta^{(t-1)}]. \quad (5)$$

- The maximization-step (M-step): In this step, a new estimated is given by,

$$\theta^{(t)} = \arg\max Q(\theta|\theta^{(t-1)}). \quad (6)$$

III. Experiments

The data is collected from two automatic weather stations which are in Chiang Mai province and Ubon Ratchathani province. Both provinces are located to the northern and north eastern region of the country, respectively, as shown in Fig. 2.

The recorded climatic data is from 2006 to 2011. It was sampled and stored every hour. The features are listed in Table I. There is 2.2% of missing data. Hence, there is 11.18% of the samples which are not able to compute ET_0. The missing data was simply excised from the samples, therefore ET_0 can be calculated for every sample.

Next, we evaluate three different scenarios in this paper: prediction of reference evapotranspiration when (i) there is no missing data by different learning algorithms, (ii) there is different amount of missing data in real-world implementation, and (iii) the feature selection process is performed.

Fig. 2. Locations of the investigation areas in Thailand, namely Chiang Mai province (A) and Ubon Ratchathani province (B).

TABLE I
LIST OF THE FEATURES USED IN THE EXPERIMENT.

Index	Features
1	Day-of-year
2	Height above mean sea level (m)
3	Latitude (Radian)
4	Wind speed (m/s at 2 meters height)
5	Sunshine duration (hour/day)
6	Maximum temperature ($^\circ$C)
7	Minimum temperature ($^\circ$C)
8	Mean temperature ($^\circ$C)
9	Relative humidity (%)

A. Scenario 1: No Missing Data

The experiment was run based on five-fold cross validation. The data is divided into five unique folds of roughly equal size. Then the algorithm is trained five times, leaving out one of the folds from training each time the machine is trained. The omitted fold will be used to compute the performance criterion, here, R^2 is used. R^2 is the coefficient of determination which measure how well output is predicted by the model. It can be calculated as follow:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{7}$$

where $\hat{y}$ is a prediction and $\bar{y}$ is the mean value of the target output.

We compared ANNs, and LR together. However, ANNs are required to tune their parameters to get the optimal performance, for example, the number of hidden layers and the number of nodes in the hidden layer. We again employed five-fold cross validation to tune the parameters in each training set. According to [4], the single hidden layer ANNs is sufficient for evapotranspiration model, therefore, we used the single hidden layer ANNs in this experiment. Hence, there is only one parameter to be tuned which is the number of node in the hidden layer. The optimal model of ANNs is with 38 nodes in the hidden layer. The experimental results are shown in Table II which reports R^2 of each algorithm. The results clearly show that ANNs yield the best performance. This is because the evapotranspiration is nonlinear model but LR is linear case.

TABLE II
PERFORMANCE COMPARISON BETWEEN ANN AND LR IN SCENARIO 1.

Algorithms	R^2
ANN (9-38-1)	**0.9999**
LR	0.9365

TABLE III
PERFORMANCE COMPARISON BETWEEN ANN AND LR WHEN MISSING DATA IS TAKEN INTO ACCOUNT.

Algorithms	Percentage of missing values	R^2 of re-estimated features	R^2 of the prediction
ANN	5%	**0.9771**	**0.9834**
LR			0.9204
ANN	10%	0.9472	0.9675
LR			0.9017
ANN	15%	0.9137	0.9443
LR			0.8759

In Thailand, there are three seasons which are as follows: (i) cool season (mid-November–mid-February), (ii) hot season (mid-February–mid-May), and (iii) rainy season (mid-May–mid-October) [10]. Fig. 3(a) and 3(b) shows the reference evapotranspiration computed by Penman-Monteith method and predicted by ANNs in Chiang Mai province and Ubon Ratchathani province, respectively. Clearly, they show that ANNs is competitive to the Penman-Monteith method. The reference evapotranspiration of both stations were at the highest point in April which is in the hot season. Then it reduced in rainy and cool seasons.

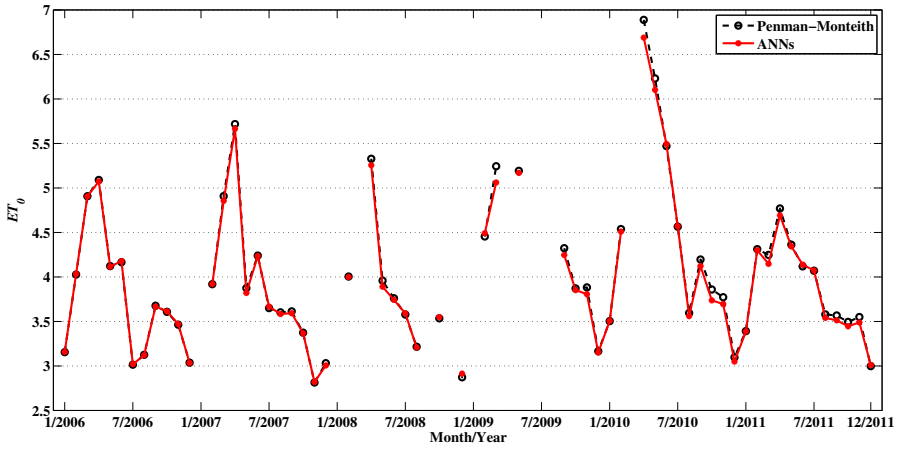
B. Scenario 2: Missing Data in Real-world Implementation

We again evaluated the results on five-fold cross validation and used the optimal models from the previous scenario. We randomly marked the test data as missing with 5%, 10%, and 15% of the data. The missing values were re-estimated by EM algorithm, then the data was trained by learning algorithms. Table III shows the experimental results. Using ANNs still gives the highest R^2 of the prediction in every cases. As expected, the performances of the re-estimated feature and the prediction of ANN and LR dropped when more missing data arises. It should be noted that only feature 4–9 can be missed as the others do not require sensors to collect the data.

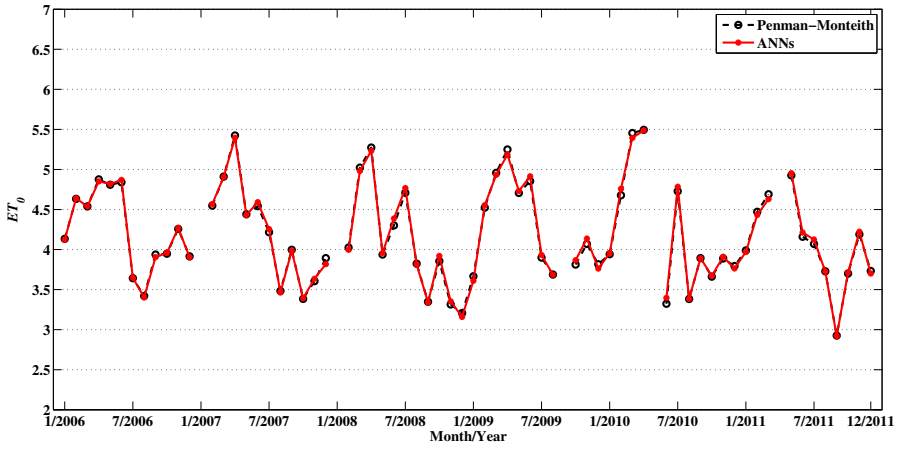
Fig. 4 shows comparisons of ET_0 computed by Penman-Monteith equation with ANNs (9-38-1) model (for the first random split) when there is no missing value, and 5%–15% of missing value. In the case which has no missing value, it has less scattered estimate than the other cases. When the amount of missing values increase, the more scattered estimate are.

C. Scenario 3: Selection of Features

The importance of the features used in the prediction is studied. LR seeks a linear combination of the variable to predict the outcome. The weight of the LR can roughly indicate the relevance of the features used in the prediction. According to the scenario 1, we have five linear regression



(a)



(b)

Fig. 3. Monthly reference evapotraspiration from 2006 to 2011 in (a) Ubon Ratchathani province and (b) Chiang Mai province by Penman-Monteith equation and ANNs.

models. Then we calculated the average of the absolute values of each element in w_i across five vectors. According to Fig. 5, it can be seen that sunshine duration is the most important feature. Moreover, we ranked the values of absolute weight in descending order which gives the following order:

$$5 > 7 > 8 > 4 > 9 > 1 > 6 > 3 \sim 2$$

In this scenario, we trained models by adding new features one-by-one according to how important the features are. Five-fold cross validation is used to evaluate the performances and search for the optimal parameters. Missing value cases were also considered too.

Table IV shows the performances of ANN and LR when different set of features are used. The classifiers with 8 features (i.e. feature 5, 7, 8, 4, 9, 1, 6, and 3) and 9 features gave the highest accuracy for both ANN and LR cases. Fig. 6 shows the relative improvement or worsening in the accuracy for ANN and LR when we considered new features in the process. In ANN, adding feature 7 (minimum temperature) to feature 5 (sunshine duration) could improve the performance by 40.99% of using only feature 5. In addition, when we considered mean temperature and wind speed in the process, the performances were improved 5.7% and 7.2%, respectively. Again, the performance were improved by 1.6% and 4.2% after adding relative humidity and day-of-year, respectively. Unfortunately, considering maximum temperature, latitude, and high above mean sea level in addition, did not improve much. Overall picture is much the same for LR.

Moreover, we also examined the feature selection process together with missing values case by ANN as shown in Table V. The overall picture is still much the same for the case with no missing values. The performances of the algorithms drop when the amount of missing values increases. Here, ANNs (8-49-1) with 5% missing values yield the best performance.

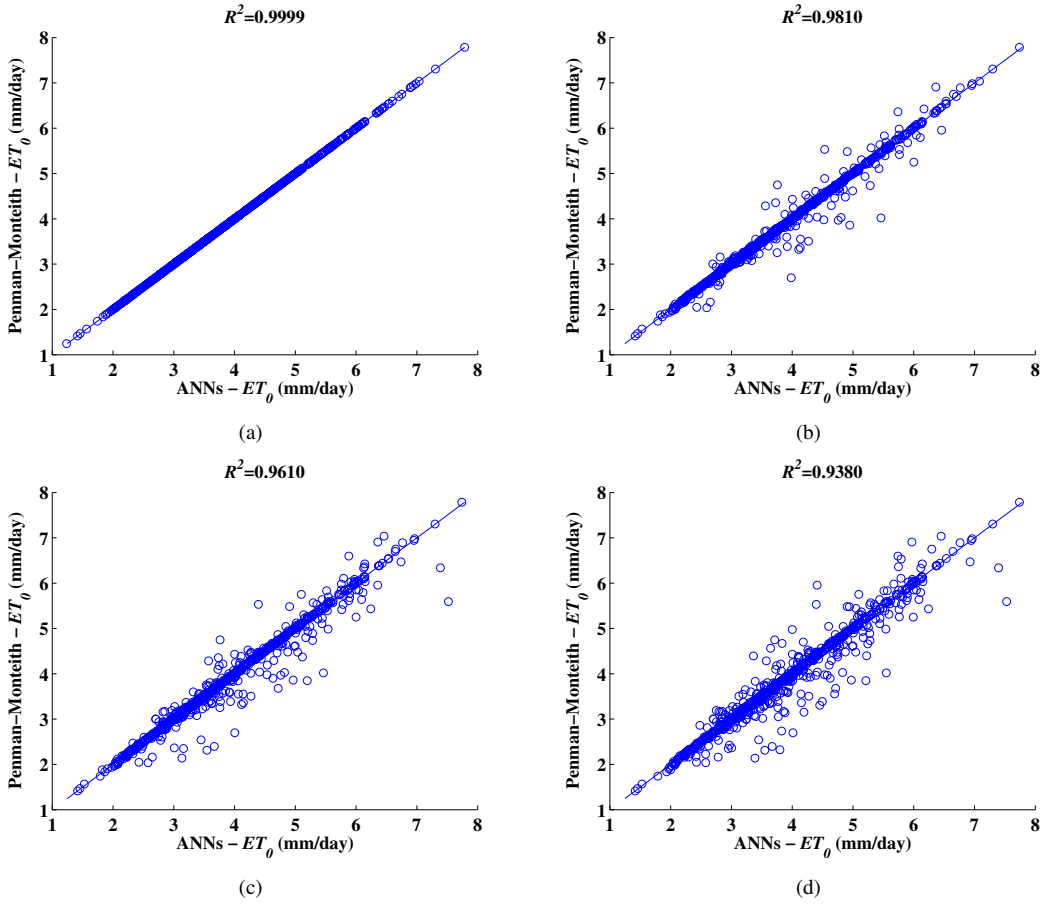


Fig. 4. Comparison of ET_0 computed by Penman-Monteith equation with ANNs (9-38-1) model when there is (a) no missing value, (b) 5% of missing value, (c) 10% of missing value and (d) 15% of missing value.

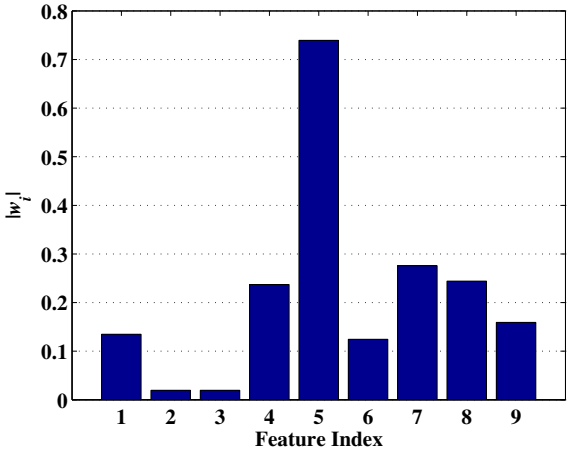


Fig. 5. An illustration of feature importance. It shows the absolute value of the linear regression weight vectors w_i average across all the models.

TABLE IV
PERFORMANCE COMPARISON BETWEEN ANNS AND LR WHEN FEATURE SELECTION IS CONSIDERED.

Features used	Algorithms	R^2
5	ANN (1-6-1)	0.5889
	LR	0.5709
5,7	ANN (2-10-1)	0.8303
	LR	0.8010
5,7,8	ANN (3-12-1)	0.8776
	LR	0.8511
5,7,8,4	ANN (4-32-1)	0.9410
	LR	0.9122
5,7,8,4,9	ANN(5-21-1)	0.9561
	LR	0.9260
5,7,8,4,9,1	ANN(6-29-1)	0.9963
	LR	0.9355
5,7,8,4,9,1,6	ANN(7-40-1)	0.9960
	LR	0.9355
5,7,8,4,9,1,6,3	ANN(8-49-1)	**0.9999**
	LR	0.9365
5,7,8,4,9,1,6,3,2	ANN(9-38-1)	**0.9999**
	LR	0.9365

IV. CONCLUSIONS

In this paper, we used ANN and LR to predict the reference evapotranspiration from the climatic data in Thailand. We have shown that ANN is competitive to the most widely used method the so-called "Penman-Monteith equation". The performance of LR is generally worse than ANN, however, this could be improved by using kernel regression as the evapotranspiration is nonlinear model. We also applied EM

TABLE V
PERFORMANCE OF ANNS WHEN FEATURE SELECTION IS CONSIDERED TOGETHER WITH MISSING VALUE SITUATIONS.

Features used	Algorithms	Percentage of missing values	R^2 of re-estimated features	R^2 of the prediction
5	ANN (1-6-1)	5%	0.9619	0.5752
		10%	0.9619	0.5752
		15%	0.9619	0.5752
5,7	ANN (2-10-1)	5%	0.9586	0.8046
		10%	0.9149	0.7610
		15%	0.8576	0.7140
5,7,8	ANN (3-12-1)	5%	0.9590	0.8544
		10%	0.9186	0.8306
		15%	0.8665	0.8070
5,7,8,4	ANN (4-32-1)	5%	0.9543	0.9086
		10%	0.9090	0.8803
		15%	0.8598	0.8520
5,7,8,4,9	ANN (5-21-1)	5%	0.9619	0.9329
		10%	0.9231	0.9098
		15%	0.8860	0.8838
5,7,8,4,9,1	ANN (6-29-1)	5%	0.9619	0.9772
		10%	0.9231	0.9527
		15%	0.8809	0.9272
5,7,8,4,9,1,6	ANN (7-40-1)	5%	**0.9740**	0.9772
		10%	0.9456	0.9674
		15%	0.9112	0.9335
5,7,8,4,9,1,6,3	ANN (8-49-1)	5%	**0.9740**	**0.9818**
		10%	0.9456	0.9605
		15%	0.9112	0.9390
5,7,8,4,9,1,6,3,2	ANN (9-38-1)	5%	**0.9740**	0.9835
		10%	0.9456	0.9617
		15%	0.9112	0.9391

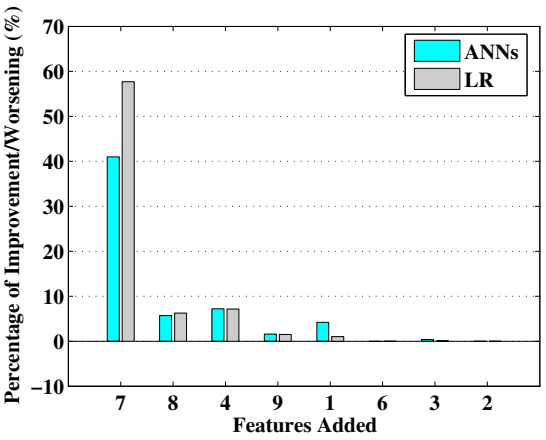


Fig. 6. Relative improvement/worsening in accuracy of ANN and LR when adding new features one-by-one according to how important the features are.

algorithm to solve the problem when missing data occur in real-world implementation. Therefore, we still can predict the reference evapotranspiration. When the amount of missing data increases, the performances of the re-estimated feature and the prediction of the learning algorithms will drop. However, this is still worthwhile. Moreover, the feature selection process was performed. We have shown the importance of the features used in the prediction. The most three importance features are sunshine duration, temperature and wind speed, respectively. In this study, we have only investigated on two automatic weather stations, therefore, we did not see much impact to the prediction with the height above mean sea level and latitude of the stations. However, if there is a number of automatic weather stations, these two features might have more effect to the prediction.

REFERENCES

[1] National Statistical Office, "Summary of the labor force survey in Thailand: March 2013," Ministry of Information and Communication Technology, Bangkok, Thailand, 2013.

[2] R. Allen, L. Pereira, D. Raes, and M. Smith, "Crop evapotranspiration-guidelines for computing crop water requirements," *FAO Irrigation and Drainage Paper*, vol. 56, pp. 15–78, 1998.

[3] G. Hargreaves and R. Allen, "History and evaluation of hargreaves evaportranspiration equation," *Journal of Irrigation and Drainage Engineering*, vol. 129, no. 1, pp. 53–63, 2003.

[4] M. Kumar, N. Raghuwanshi, R. Singh, W. Wallender, and W. Pruitt, "Estimating evapotranspiration using artificial neural network," *Journal of Irrigation and Drainage Engineering*, vol. 128, no. 4, pp. 224–233, 2002.

[5] M. Benzaghta, T. Mohammed, A. Ghazali, and M. Soom, "Prediction of evaporation in tropical climate using artificial neural network and climate based models," *Scientific Research and Essays*, vol. 7, no. 36, pp. 3133–3148, 2012.

[6] S. Zanetti, E. Sousa, V. Oliveira, F. Almeida, and S. Bernardo, "Estimating evapotranspiration using artificial neural network and minimum climatological data," *Journal of Irrigation and Drainage Engineering*, vol. 133, no. 2, pp. 88–89, 2012.

[7] T. Wang, S. Traore, and T. Kerh, "Neural network approach for estimation reference evapotranspiration from limited climatic data in Burkina Faso," *WSEAS Transactions on Computers*, vol. 7, no. 6, pp. 704–713, 2008.

[8] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[9] A. Dongare, R. Kharde, and A. Kachare, "Introduction to artificial neural network," *International Journal of Engineering and Innovative Technology*, vol. 2, no. 1, pp. 189–194, 2012.

[10] Royal Command of H.M. the King, "Climate," *Thai Junior Encyclopedia*, vol. 4, no. 6, 1978.