On Applying Regression and Neural Network to Predict Rainfall Using Satellite Based Index

Ratiporn Chanklan*, Keerachart Suksut, Kedkard Chaiyakhan, Nuntawut Kaoungku, Kittisak Kerdprasop and Nittaya Kerdprasop

***Abstract*— In this paper, we adopt a statistical method using the linear regression analysis to study relationship between the satellite based vegetation index and the ground based rainfall data, and then apply the data mining method using the neural network to induce a model to predict the amount of annual rainfall. The model is intended to be useful for drought monitoring. Remote sensing data used in our study is the Normalized Difference Vegetation Index (NDVI) obtained from the NOAA STAR. The ground station rainfall data during the years 2005 to 2014 in Nakhon Ratchasima province, Thailand, are obtained from the Meteorological Department. The study of NDVI and ground-based rainfall relationship has been done through the correlation coefficient analysis. The preliminary study results show that vegetation index and rainfall positively correlate with 1-month lagged time. The studied period from June to September shows the strong correlation (r = 0.715, on average). We then induce the rainfall predictive model using neural network with the NDVI and rainfall as the input parameters. The performances of using only a rainfall parameter and a combination of NDVI and rainfall parameters are also compared. The experimental results show that using the remote sensing NDVI data together with the ground based rainfall data can improve accuracy of the neural network model to predict the future annual rainfall.**

***Index Terms*— Remote sensing, Normalized Difference Vegetation Index, Annual rainfall prediction, Neural network**

I. Introduction

A drought is water shortage in an area occurring from precipitation deficiency or the unseasonally lacking of rainfall. The effects of drought are numerous: a lack of water for humans and other living animals, insufficient water for the crops, food inadequate due to crop damage. The agriculture in Thailand depends on natural water, mostly rainfall. Therefore, lacking of enough rainfall for a long period of time has serious effect to agricultural yields. Traditionally, drought monitoring has been done based on meteorological information from ground stations, which does not cover all areas in the country. Due to the limited numbers of ground stations, it results in less, discontinuing, and incomplete data for timely drought warning. Therefore, data analyses are inaccurate and sometime out of date because it takes long time to collect data.

Satellite data can be used to help monitoring drought situation. Environmental observation satellites have been launched to record a continuous, spatial patterns. The satellite data cover the whole global surface area and the data are available for almost real-time access [1]. The remote sensing is thus suitable to track the changes and to monitor the impact of drought on crops [2],[3],[4]. Thus, the intuitive idea of this work is that if we can find the relationship between the remote sensing data and the ground station data from the Meteorological Department, then in the process of model inductive we can use the remotely sensed data incorporating with the ground-based data for predicting the rainfall in the upcoming year. Such model has the advantage of timely monitoring of drought.

In this work, we firstly analyze the relationship between Normalized Difference Vegetation Index obtained from the NOAA satellite and the monthly rainfall data from the Meteorological Department in Nakhon Ratchasima province, Thailand. After confirming its positive correlation, we then build a model to predict the rainfall using artificial neural network. The inputs for our neural network model are the lagged 1-month rainfall and lagged 1-moth NDVI data.

II. Background Theories

A. Normalized Difference Vegetation Index

In this study, we use the remotely sensed data, which is the Global Vegetation Index (GVI)in the area of Nakhon Ratchasima province in the northeast of Thailand. GVI is the basic index for measuring the greenness of the earth surface through the monitoring for density and healthy of land surface vegetation. One specific product of GVI is the Normalized Difference Vegetation Index (NDVI), which is the computation of signals sensed by the channels 1 and 2 of the satellite that aggregates the 4 square km global area coverage daily.

Manuscript received September 26, 2016; revised January 10, 2017. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

R Chanklan is a doctoral student with the School of Computer Engineering, Suranaree University of Technology, 111 University Avenue, Muang, Nakhon Ratchasima 30000, Thailand. (corresponding author: +66994696164; e-mail: arc_angle@hotmail.com).

K. Suksut is a doctoral student with the School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand. (e-mail: mikaiterng@gmail.com).

K. Chaiyakhan is a lecturer with the Computer Engineering Department, Rajamangala University of Technology Isan, Nakhon Ratchasima, Thailand. (e-mail: kedkarnc@hotmail.com).

N. Kaoungku is a lecturer with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand. (e-mail: nuntawut@sut.ac.th).

K. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: kerdpras@sut.ac.th).

N. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand. (e-mail: nittaya@sut.ac.th)

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The basic concept of NDVI is based on the fact that internal mesophyll structure of healthy green leaves reflects near-infrared (NIR) radiation, whereas the leaf chlorophyll and other pigments absorb a large proportion of the red visible (VIS) radiation. This function of internal leaf structure becomes reversed in case of unhealthy or water stressed vegetation [5]. The calculation of the NDVI value is thus performed with the Equation1.

$$NDVI = (NIR-VIS) / (NIR+VIS) \quad (1)$$

where, NIR is near infrared and VIS is visible red band of electromagnetic spectrum. The value of NDVI ranges between 1 and +1. It is found below 0.1 in the areas with barren rock, sand and snow cover, whereas it may range from 0.6 to 0.8 in temperate and tropical rainforests. NDVI is suitable for monitoring drought, estimating healthy status of vegetation, crop growth conditions and crop yields [6],[7].

B. *Rainfall Data*

Rainfall is very important in meteorology because water is a major factor related to the living of people, living creatures, and agriculture. Healthy vegetation and plentiful crop lands are all depending on rainfall. A measure of rainfall is normally done by a rain gauge placed in open space for 24 hours. Observations of daily rainfall are nominally made at 7:00 am to 7:00 pm local clock time each day. The automatic rain gauges can measure rainfall continuously 6, 12, 24 hours or weekly.

C. *Correlation Coefficient*

The correlation is a statistical measure used to explore relationship between the variables. The degree of correlation is interpreted from the correlation coefficient (R). The correlation coefficient is a numerical value between -1 and 1. It expresses the strength of the linear relationship between two variables (x and y). The direction of the relationship between the two variables can be shown by scatter plot. There are three possible types of relationship: positive correlations, negative correlations, and zero correlations.

Positive correlation is the kind of relationship such that the increase or decrease the value of one variable will cause a corresponding increase or decrease in value of the other variable.

Negative correlation is reverse relationship in which the increase or decrease on a variable's value will cause the other variables change their values in opposite direction.

Zero correlation is the situation in which the two variables have no relationship. The correlation coefficient can be calculated using Equation 2.

$$R = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (2)$$

where n is the total number of samples, $x_i = (x_1, x_2, \ldots, x_n)$ are the values of variables x, and y_i is the value of variable y. If the correlation coefficient is closer to 1 or -1, it represents the relationship between the variables at a high level. If the value is close to 0, it represents the relationship between the variables at a low level or no relationship.

D. *Artificial Neural Networks*

The Artificial Neural Network (ANN) is the most widely used form of neural networks. An ANN is a computational approach inspired by studies of the brain and nervous systems in biological organisms. The powerful functionality of a biological neural system has been attributed to the parallel-distributed processing nature of the biological neurons [8]. Conventionally, the network (fig.1) has three main levels: input layer, hidden layer, and output layer. Nodes in the input layer called the input nodes. The number of nodes in the input layer is equal to the number of features (attributes, independent variables). Nodes in the hidden layer are called the hidden nodes. Number of nodes in the hidden layer is defined by a user. Node in the output layer is called an output node. The number of output nodes is equal to the number of data groups (or target, dependent variable). Nodes in the network are connected with lines; from input nodes to hidden nodes, and from hidden nodes to output nodes. Each line connecting one node to another has weight (w).

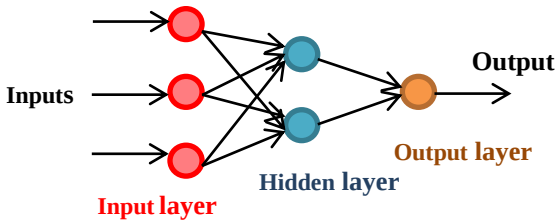


Fig.1. Architecture of the Artificial Neural Network.

Input data is vector of elements: $p = [p_1, p_2, \ldots, p_R]$, R is number of elements (or dimension) in input data. Each line in the network is annotated with weight: $W = [w_1, w_2, \ldots, w_R]$. The network works by multiplying weight on each edge to the input data, summing results from each incoming edge of the node, and finally summing a bias (b) of that node. The summation result of each node is denoted by n. The result (n) has to be transformed through a transfer function to obtain the final computation result of each node, denoted as a. The calculation of the output value from a neuron as shown in fig.2 is summarized in Equation 3.

$$a = f(n) = f(Wp + b) \quad (3)$$

where $n = w_{1,1}p_1 + w_{1,2}p_2 + \cdots + w_{1,R}p_R + b$

$n = Wp + b$

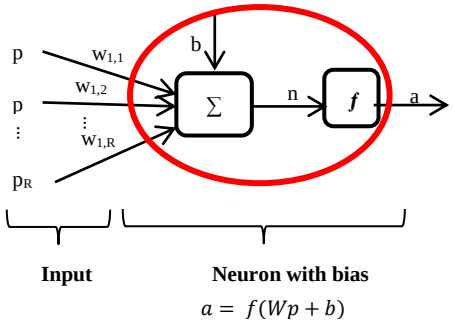


Fig.2. Neural unit with incoming input edges and a bias.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

For the weight and bias, it is can be adjusted from the learned data. If the output value from neural network is false as compared to the true value in the training data, the weight will be update using the error as a guidance. The process continues until the predictive error of the neural network model is less than the acceptable threshold.

III. MATERIALS AND METHODS

In this work, the study area is located in Nakhon Ratchasima, a largest province in the northeast of Thailand (fig.3). We used remote sensing data, which is the Normalized Difference Vegetation Index or NDVI, obtained from the global vegetation health products of NOAA STAR (http://www.star.nesdis.noaa.gov/smcd/emb/vci/VH/vh_browseByCountry_province.php?country=THA&provinceID=28&year1=1981&year2=2015X), and monthly rainfall data in Nakhon Ratchasima area during the years 2005 to 2014. The data are obtained from the Meteorological Department (http://www.dnp.go.th/statistics/dnpstatmain.asp).

Our main objective of study is to find the relationship between Normalized Difference Vegetation Index and rainfall data through the analysis of correlation. The focus of our study is the correlation of the two variables during the southwest monsoon season. To explore the relationship during monsoon period, we use data during the months of June to September. This choice is because the southwest monsoon prevails over Thailand between mid-May to mid-October.

We then use linear regression to determine the relationship between two variables; the independent variable (X) is the Normalized Difference Vegetation Index and the dependent variable (Y) is the rainfall. The NDVI and rainfall data have been lagged from 1 to 6 months. We find the linear correlation between the independent variables and the dependent variable. The preliminary result shows that the lagged 1-month is the discriminative factor, as shown in Equation 4.

$$f(Rainfall_t) = f(Rainfall_{t-1}, NDVI_{t-1}) \qquad (4)$$

Fig.3. The study area: Nakhon Ratchasima, Thailand. (www.maphill.com)

In the subsequent step of rainfall prediction model induction with neural network, we thus use rainfall lagged 1-month and NDVI lagged 1-month as input parameters of the neural network. The output node of the network is a node to predict rainfall in current time (t).

IV. EXPERIMENTAL RESULTS

We plot the linear regression to find the relationship between two variables (NDVI and rainfall data). The results are shown in fig 4.

When we analyzed yearly relationship between NDVI and annual rainfall from 2005 to 2014, the results show the correlation coefficient as both positive and negative values. This means that the relationships of the NDVI and annual rainfall during the past ten years are not in the same direction each year. The correlation coefficient range is -0.134 to 0.438. This implies that NDVI and rainfall give quite poor correlation. By merging the ten-year data as one group, we find the correlation coefficient to be 0.159. A single-group NDVI and rainfall relationship is show in fig 5.

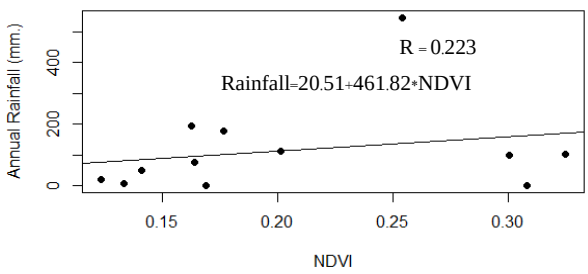
To further analyze NDVI-rainfall relationship, we plot NDVI and rainfall values to see a fine-grain relationship in the monthly period, and the result is shown in fig 6. In fig.6, the monthly NDVI and rainfall during the years 2005 to 2014 are plotted as separate graphs.

From fig. 6, we can now notice the trend of NDVI to increase from June to September, every year from 2005-2014. There are two increase trend in the rainfall plot; that is from June to August, and another peak from August to September. June to September is actually the raining season that has some effect from the monsoon. Thailand is under the influence of two monsoon types: southwest monsoon and northeast monsoon. Southwest monsoon prevails over in mid-May to mid-October. During these months, weather are cloudy and rainy. Northeast monsoon prevails over in mid-October to mid-February, with the clear, cold, and dry weather. Fig 6 shows rainfall in June to September with increasing amount of rainfall. The increasing trend also appears in the NDVI plot. We therefore find the correlation coefficient of NDVI with rainfall during June to September, and the results are shown in Table 1.

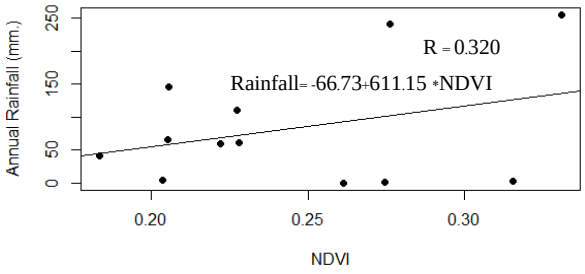
TABLE I
THE CORRELATION COEFFICIENT RESULTS

Year	January to December	June to September
2005	0.223	0.917
2006	0.320	0.940
2007	-0.098	0.726
2008	0.114	0.983
2009	-0.076	0.709
2010	0.438	0.880
2011	-0.134	-0.114
2012	0.239	0.600
2013	0.552	0.690
2014	0.225	0.817
2005 to 2014	0.159	0.518

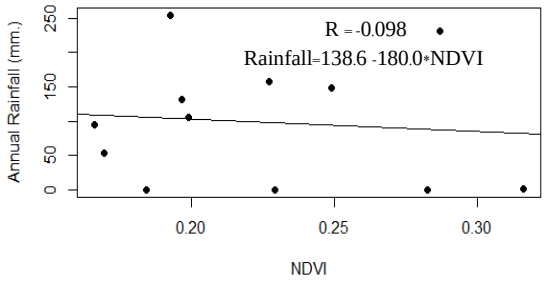
ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)



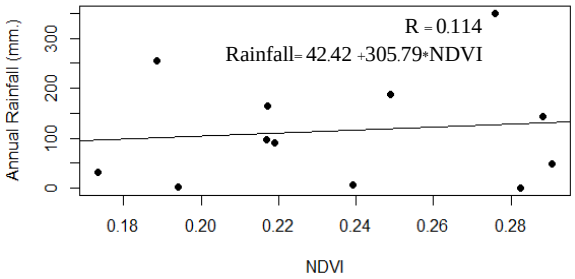
(a) 2005



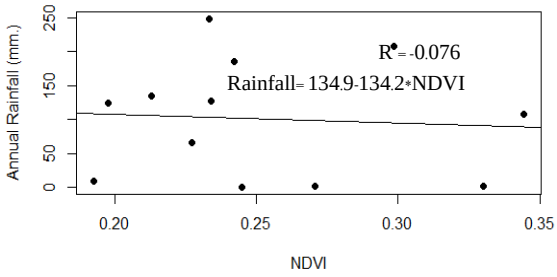
(b) 2006



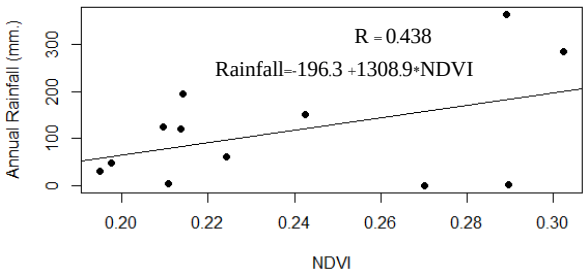
(c) 2007



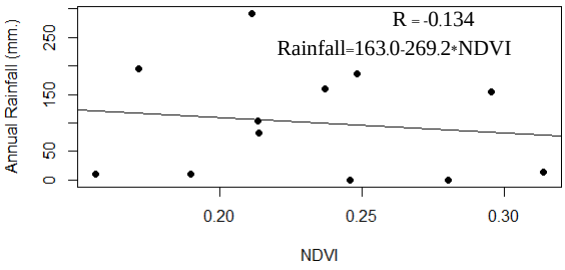
(d) 2008



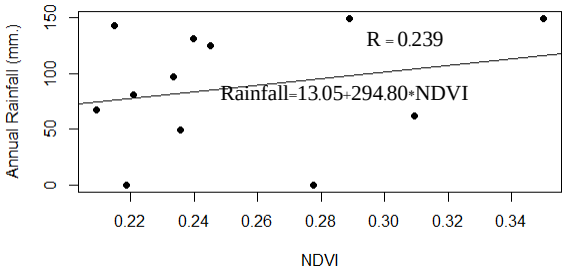
(e) 2009



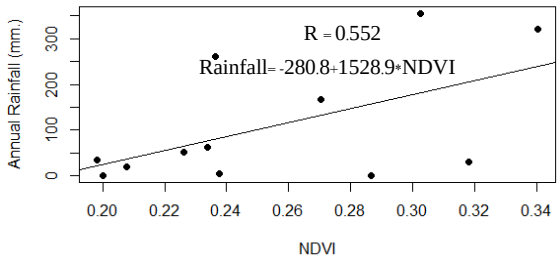
(f) 2010



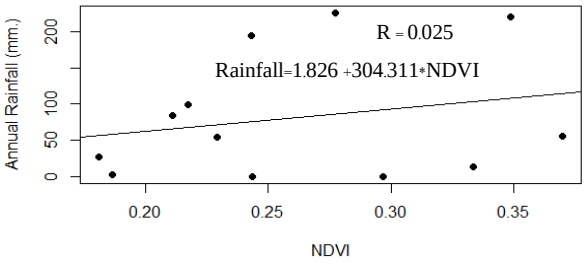
(g) 2011



(h) 2012



(i) 2013



(j) 2014

Fig.4. The correlation coefficient (R) of NDVI and rainfall.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

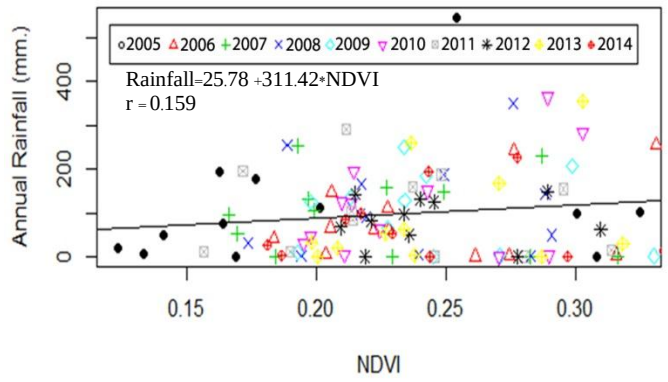
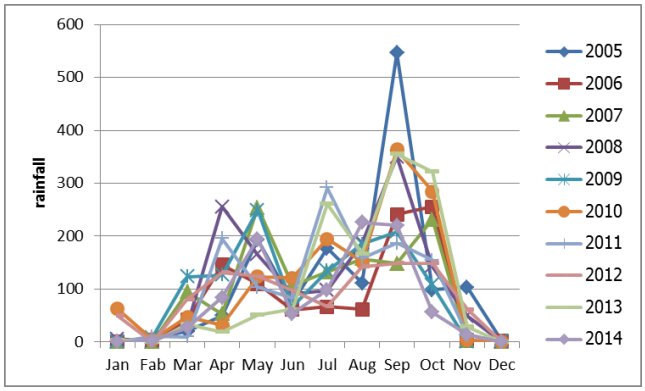
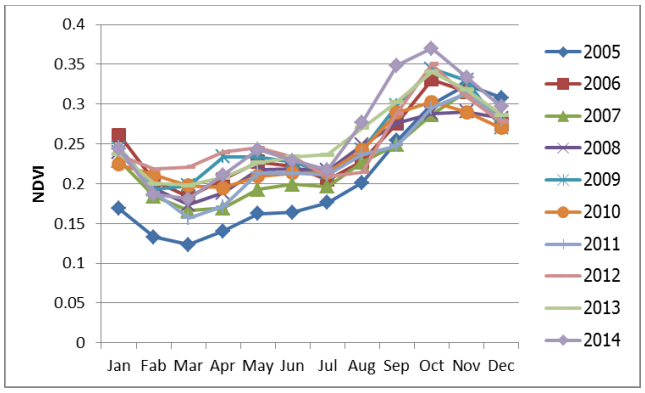


Fig.5. The correlation coefficient of NDVI and rainfall from 2005 to 2014.



(a) Rainfall



(b) NDVI

Fig.6. The graphs show monthly (a) rainfall and (b) NDVI.

From Table 1, it can be noticed that during the Southwest monsoon season, the correlation coefficient of NDVI and rainfall during the months of June to September shows positive direction with strong correlation in the range 0.6 to 0.9. There is only one exception in the year 2011; the NDVI-rainfall relationship shows negative sign. In the year 2011, the rainfall was very unusual affected by the Tropical Storm Nock-ten [9]. It caused big flood in Thailand. NDVI is the index for measuring the 'greenness' of the earth's surface. When flooding covered crops, the relationship between NDVI and rainfall showed negative direction, which is different from other years.

TABLE II
THE RESULTS IN PREDICT RAINFALL USING ARTIFICIAL NEURAL NETWORKS

Data	Input Parameter to ANN	R	RMSE
January to December	Rainfall	0.42	87.91
January to December	Rainfall, NDVI	0.62	74.02

After performing correlation analyses, we then apply artificial neural networks to build a model for predicting rainfall. For experimentation, we use the data from 2005 to 2010 as training data for a neural network, and the data from 2012 to 2014 are used as test data. We use data from January to December in both training and test data.

We also compare the model performance using two different set of input parameters: set one having only rainfall data as input parameter into ANN, and set two having rainfall and NDVI as input parameters for ANN. The results are shown in Table 2. The model that uses both rainfall and NDVI data as input parameters gives a less root-mean-square error (RMSE) than the model that has only rainfall as its input parameter. The correlation coefficient also shows good relationship. The RMSE metric used for model evaluation can be calculated as in equation (4).

$$\mathrm{RMSE} = \sqrt{\left(\frac{\sum(T_i - O_i)^2}{N}\right)} \qquad (4)$$

where T_i is the actual data, O_i is the value of predict, and N is the number of all data. The less value of RMSE means the more accurate prediction of the model.

V. CONCLUSION

In this study, we use remote sensing data of Normalized Difference Vegetation Index from the NOAA satellite and rainfall data from the Meteorological Department of Nakhon Ratchasima province in Thailand. The collected data are from the years 2005 to 2014. The analysis results of regression method reveal that a relationship during the months of June to September, which is the period of southwest monsoon, is the best relationships. NDVI and rainfall during this period have positive relationship. The regression model also shows that current rainfall can be estimated from the lagged 1-month rainfall and lagged 1-month NDVI.

To build a model for rainfall prediction with artificial neural network, we apply NDVI and rainfall as input parameters to the network. The model yields prediction result with lower RMSE as compared to the prediction with only rainfall data as input vector of ANN.

The results observed from our experiment show that rainfall data from the ground station and NDVI data from remote sensing are supplement each other for predicting rainfall and monitoring drought, which is the period with less rainfall than the usual. This emphasis the advantage of remote sensing data that can help timely prediction and can cover broad area of land surface.

REFERENCES

[1] Gu, Y., Brown, J. F., Verdin, J. P., & Wardlow, B. (2007). "A five-year analysis of MODIS NDVI and NDWI for grassland drought assessment over the central Great Plains of the United States," *Geophysical Research Letters*, 34(6).

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[2] Kogan, F., & Guo, W., "Early Detection and Monitoring Droughts From NOAA Environmental Satellites. In Use of Satellite and In-Situ Data to Improve Sustainability," *Springer Netherlands*, pp.11-18, 2011.

[3] Kuri, F., Murwira, A., Murwira, K. S., & Masocha, M, "Predicting maize yield in Zimbabwe using dry dekads derived from remotely sensed Vegetation Condition Index," *International Journal of Applied Earth Observation and Geoinformation*, vol 33, pp. 39-46, 2014.

[4] Dutta, D., Kundu, A., Patel, N. R., Saha, S. K., & Siddiqui, A. R., "Assessment of agricultural drought in Rajasthan (India) using remote sensing derived Vegetation Condition Index (VCI) and Standardized Precipitation Index (SPI)," The Egyptian Journal of Remote Sensing and Space Science. vol. 18, pp. 53-63, 2015.

[5] Dipanwita Dutta, Arnab Kundu, N.R. Patel, S.K. Saha and A.R. Siddiqui, "Assessment of agricultural drought in Rajasthan (India) using remote sensing derived Vegetation Condition Index (VCI) and Standardized Precipitation Index (SPI)," *The Egyptian Journal of Remote Sensing and Space Sciences*, 2015, vol.18, pp.53–63.

[6] Kogan, F.N., "Vegetation index for a real analysis of crop conditions," *In Proceedings of the 18th Conference on Agricultural and Forest Meteorology, AMS, W. Lafayette, Indiana, 15–18 September 1987*, Indiana, USA, pp. 103–106, 1987.

[7] Dabrowska-Zielinska, K., Kogan, F.N., Ciolkoszs, A., Gruszczynska, M., Kowalik, W., "Modelling of crop growth conditions and crop yield in Poland using AVHRR-based indices," *Int. J. Remote Sens*, vol. 23, pp. 1109–1123, 2002.

[8] K.C. Luk, J.E. Ball and A. Sharma, "A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting," *Journal of Hydrology*, 2000, vol 226, pp. 56–65.

[9] Watanasak Sornrunk, Tawatchai Kamoltam. "Ministry of Public Health:The operation Flood crisis year 2011," *Journal of Preventive Medicine Association of Thailand*, vol. 2, no. 2, 112-115, 2011.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)