Clustering the Rainfall Stations in Thailand Using Self-Organizing Map

Natita Wangsoh and Wachirapond Permpoonsinsup
Faculty of Science and Technology, Pathumwan Institute of Technology, Bangkok, Thailand.
natita.pit@gmail.com

Abstract. The purpose of this study is to classify the pattern of rainfall stations in Thailand over the period from the year 2000 to 2009 using Self-Organizing Map (SOM). The number of stations is collected based on the complete rainfall data during rainy season. The optimal cases of learning rate function and different two SOM array sizes are investigated. The Gaussian function is applied as the neighborhood function. The results show that linear function and the 2×2 SOM array size are most suitable for classifying the pattern of rainfall. The similar characteristic of daily rainfall in each station is clustered into four patterns. Thus, SOM can be classified the similar rainfall stations, efficiently.

1. Introduction
Rainy season in Thailand has occurred between mid-May to mid-October under the influence of the monsoon seasonal winds called southwest monsoon. Rainfall plays an important factor in water management directly impacted the human activities, especially in agriculture and transportation. The characteristic of rainfall such as the amount, the intensity, the duration, the frequency and the seasonal distribution in each region can provide the useful information to help people in planning of the land use. Some researches study the characteristic of rainfall. Reference [1] analyze the character, pattern and distribution of rainfall. They said that the knowledge of the distribution and the amount of rainfall is very much essential for successful management of agriculture.

To classify the characteristic of rainfall, clustering technique is one of the most commonly used. Self-Organizing Map (SOM) is first introduced in 1990 by Kohonen as a special class of neural network. It has proven to be a very powerful tool for data analysis in the aspect of cluster visualization [2]. SOM is effectively applied in many fields of data compression, pattern recognition, optimization, medical application, industrial instrumentation, etc. Additionally, SOM is usually applied in the field of meteorology to study the weather pattern and to classify the similarity of the meteorological data. Reference [3] applies SOM in wavelet coefficients to identify the similar zones of precipitation behavior over the Iberian Peninsula. Reference [4] uses SOM to classify stations with similar precipitation characteristics. Reference [5] establishes SOM to specify the pattern of specific humidity over southern of Thailand. All studies prove that SOM has a reliable performance for classifying the meteorological pattern. This study aims to investigate the similar pattern of rainfall stations by using SOM during rainy season in Thailand between the years 2000 to 2009. The characteristic of rainfall in each rainfall station will therefore be classified.
2. Data Description
Daily rainfall data used in this study are from the Thai Meteorological Department (TMD). The dataset covers the 10 years from 2000 to 2009 starting from every 15th of May to 15th of October. Thirty rainfall stations are collected based on the completed data set. The spatial distribution of 30 stations is shown in Figure 1.

![Figure 1. The thirty rainfall stations in Thailand.](image)

3. Self-Organizing Map (SOM)
The principle of SOM is in the concept of a transformation a complex, high-dimensional input space into a simple low-dimensional discrete output space [2]. The spatial locations, coordinates, of the nodes in the output space are indicative of inherent statistical features contained in the input space. SOM is trained by unsupervised learning algorithm included three essential phases: competition, cooperation and adaptation. Before training, the initial values of learning rate, radius of the neighbourhood, number of iterations and the number of patterns (SOM array size) are required [6].

At the start of learning, weight vector, \( \mathbf{w}_i \), is generated by a random numbers and input vector, \( \mathbf{x} \) is a random distribution which corresponding to the column index. The set of weight vector is formed as:

\[
\mathbf{w}_{ij} = \mathbf{w}_i, i = 1,2,\ldots,k, j = 1,2,\ldots,k_y,
\]

where \( k_x \) is the number of rows and \( k_y \) is the number of columns. The algorithm then goes to three phases for calculating.

In competition, the Euclidian distance between the input vector and the neuron with weight vector of the given neuron, \( \mathbf{w}_c \), is computed as

\[
d(x,w) = \left\| \mathbf{x}(t) - \mathbf{w}_c(t) \right\| \tag{1}
\]

The neuron with the most similar weight vector to the input will search for the winner neuron (best matching unit, BMU). BMU is calculated as

\[
BMU = \text{argmin} \left\| \mathbf{x}(t) - \mathbf{w}_c(t) \right\| \tag{2}
\]

In cooperation, the collected neighborhood function used in this study is the Gaussian function computed as

\[
h^C_{ij} = \alpha(t) \times e^{-\frac{1}{2} \sum_{i,j} (x-y)^2} \tag{3}
\]

The parameter \( \eta^C_{ij} \) represents the neighbor rank between nodes \( \mathbf{w}_c \) and \( \mathbf{w}_j \) (the radius of neighborhood). Two-dimensional vectors \( R_c \) and \( R_j \) include indices of \( \mathbf{w}_c \) and \( \mathbf{w}_j \) (numbers of row and column) [7].
In adaptation, weight vector is adjusted after obtaining the winning neuron to increase the similarity with the input vector. The rule for updating the weight vector is given by

$$w_i(t + 1) = w_i(t) + \alpha(t) h_{ij}(t) (x(t) - w_i(t))$$  \(4\)

Here \(h_{ij}(t)\) is a neighborhood function and \(t\) is the order number of a current iteration. Three learning rate functions, \(\alpha(t)\) are defined as follows:

Linear: \(\alpha(t) = \alpha(0) \cdot \frac{1}{t}\)  \(5\)

Inverse of Time: \(\alpha(t,T) = \alpha(0) \cdot \left(1 - \frac{t}{T}\right)\)  \(6\)

Power series: \(\alpha(t,T) = \alpha(0) \times e^\frac{t}{T}\)  \(7\)

Here \(T\) is the number of total iterations and \(t\) is the order number of a current iteration \([7]\). However, it is under the condition \(\eta_{ij} \leq \left[\alpha \max(k, k_i), 1\right]\) for all cases of analysis.

The percent of occurrence or frequency of each pattern is the number of occurrences divided by the total number of samples. The probability that specific humidity would map to any pattern is \(\frac{1}{n}\), where \(n\) is the number of patterns. The significance of the frequency can be determined by calculating the 95% confidence interval around the expected probability of \(\frac{1}{n}\). Assuming that the process is binomial, the 95% confidence limits are calculated by

$$p \pm 1.96 \left[\frac{p(1-p)}{N}\right]^{1/2}$$  \(8\)

where \(p\) is the probability that any sample maps to any pattern and \(N\) is the number of input vector used to train the map \([6]\).

To measure the quality of SOM, quantization error is used. It is computed from the average distance of input vectors \((x_i)\) to the weight vector of the winner node \((w'_j)\) of the BMU. A SOM with lower average error is more accurate than a SOM with higher average error \([8]\). Quantization error is calculated by

$$QE = \frac{\sum_{i=1}^{N} \| x_i - w'_j \|}{N}$$  \(9\)

4. Experimental Design

In the experimental case, daily rainfall data from TMD are used as the input vector for SOM. There are 30 rainfall stations in Thailand selected based on the complete rainfall data. The data include 154 days during rainy season in Thailand between 15th May and 15th October over the years 2000 to 2009. Thus, the total days consist, \(i\), of 1540 days represented as the number of rows. The total completed rainfall stations, \(j\), is 30 stations represented as the number of columns. The input vectors \(x_{ij}\) are in the matrix formed with the dimension of 1540x30 where \(i = 1,2, \ldots, 1540\) and \(j = 1,2, \ldots, 30\). In this study, three experiment cases of learning rate functions are investigated the optimal case with the Gaussian function. Two different cases of SOM array sizes are examined the number of rainfall station patterns. All initial parameters of SOM are set as in Table 1. In order to examine the pattern of rainfall stations, 10 percent trimmed mean is applied to remove the outlier data in each day over ten years. The average daily rainfall data is then calculated.
### Table 1. All initial parameters of SOM.

| Parameters                        | Value         |
|-----------------------------------|---------------|
| Initial learning rate             | 0.85          |
| Initial weight vector             | Random        |
| Maximum radius of neighborhood    | Size 10       |
| Maximum number of iteration       | 10,000        |
| SOM array size                    | 2×2, 3×3      |

### 5. Results

#### 5.1. The optimal learning rate function

In SOM algorithm, the types of learning rate functions impact the learning process in the aspect of controlling the size of the weight vector. Three learning rate functions: Linear, Inverse of time and power series are considered with two different cases of SOM array sizes. Table 2 shows the average quantization error for all experiment cases. The results proved that linear learning rate function establishes the lowest quantization error both of 2×2 and 3×3 SOM array sizes. Therefore, linear is most suitable learning rate function.

#### Table 2. The average quantization error for all experiment cases.

| Learning rate functions | SOM array size | 2×2 | 3×3 |
|-------------------------|----------------|-----|-----|
| Linear                  | 0.3466         | 0.3396 |
| Inverse of time         | 0.3896         | 0.3779 |
| Power series            | 0.4013         | 0.3840 |

#### 5.2. The optimal SOM array size

In order to investigate the suitability SOM array size, two different array sizes of 2×2 and 3×3 are examined. Frequency value outside the confidence interval is considered significant. The frequency of occurrence of rainfall station patterns is shown in Tables 3 and 4, respectively. The ordered pairs represent the node on SOM map.

#### Table 3. Node frequency for rainfall stations pattern with 2×2 SOM array sizes (%).

| Nodes | 1    | 2    |
|-------|------|------|
| 1     | 46.67| 13.33|
| 2     | 3.33 | 36.67|

For the 2×2 array size, four patterns, the probability of occurrence for each pattern is 1/4 or 25%. According to (6), the confidence interval is in the range of 9.50%-40.50%. From Table 3, the patterns of nodes (1, 1) located in the top left corners occur most frequently. Furthermore, a node (1, 1) is the only one node with the frequency values outside the confidence interval.

For the 3×3 array size, nine patterns, the probability of occurrence for each pattern is 1/9 or 11.11%. The confidence interval is in the range of 9.36%-12.86%. There are two patterns have no member. It means that the 3×3 array size cannot import all input vectors to the SOM algorithm. Therefore, the 2×2 array size is the optimal case in clustering the rainfall stations pattern for this study.

#### 5.3. Rainfall Pattern

The characteristic of rainfall stations during the rainy season of Thailand over the period from 2000 to 2009 is classified. From the previous section, the most suitable experiment case of SOM algorithm is a linear function with 2×2 SOM array size. Figure 2 illustrates the pattern of thirty rainfall stations from the year 2000 to 2009. The average rainfall amount in each day can be found. There are four patterns
classified by SOM algorithm. Usually the x-axis shows the number of days from 15th May to 15th October (154 days). The y-axis shows the average rainfall amount (mm). There is only one pattern significantly learned by 2x2 SOM array size, pattern (1, 1). However, this study considers the similarity of the rainfall stations when they are grouped by SOM. It means that they have similar daily rainfall amount in rainy season over ten years. As a result, fourteen stations in pattern (1, 1) consist of Suphanburi, Chumphon, Tak, Nakhonsawan, Lopburi, Prachinburi, Sakaew, Phuket, Trang, Phumipon Dam, Mae sot, Ubonratchathani, Chiang Mai and Nakhon Sri Thammarat. There are eleven rainfall stations classified to pattern (2, 2) including Khonkhan, Chiang Rai, Phare, Nan, Uttaradit, Nong Khai, Sakon Nakhon, Nakhon Phanom, Roi et, Surin and Phuket (centre). For pattern (1, 2), SOM provides the similar four rainfall stations of Lampang, Loei, Petchaburi and Koh Samui. On the other hand, it has only rainfall station maps to the pattern (2, 1). Thus, the rainfall amount at Kanchanaburi during rainy season not match to others.

![Graphs showing average rainfall amount for different patterns during rainy season.](image)

**Figure 2.** The average rainfall amount of rainfall station patterns during rainy season (mm).

6. Conclusion
The amount is the important characteristic of rainfall effected to the daily life activities. SOM is an unsupervised learning algorithm used in this study to classify the similarity of daily rainfall amount in the rainy season during the year 2000 to 2009 over Thailand. Thirty rainfall stations are selected for this study. Three learning rate functions and two different SOM array sizes are investigated. The quantization error proves the robustness of SOM algorithm. The experimental results found that linear function is most suitable learning rate function for this study. Besides, the 2x2 SOM array size establishes the best results in clustering the pattern of rainfall stations. Four patterns of rainfall stations are grouped, proficiently. Besides, the average daily rainfall amount of each pattern can map to any rainfall station in the same pattern. It is helpful for farmers to plan about their agricultural activities and also for water management planning.

Acknowledgments
Authors wishing to acknowledge faculty of science and technology, Pathumwan Institute of Technology for financial support.
References

[1] Lala I P R, Bora P K, Ram V, Singh A K, Singh R and Feroze S M 2015 *Indian Journal of Hill Farming* vol 28 no 1

[2] Umut A and Secil E 2012 *An Introduction to Self-Organizing Maps* (Atlantis Press) ed Kahraman C pp 299–319

[3] Morata A, Martin M L, Luna M Y and Valero F 2006 *Theor. Appl. Climatol.* vol 85

[4] Crane R G and Hewitson B C 2003 *Clim Res* vol 25

[5] Wangsoh N, Watthayu W and Sukawat D 2015 *The 11th IMT-GT International Conference on Mathematics, Statistics and Its Applications*

[6] Wangsoh N, Watthayu W and Sukawat D 2015 *International Journal of Modeling and Optimization* vol 6 no 1

[7] Stefanovič P and Kurasova O 2011 *Nonlinear Anal. Model. Control* vol 16

[8] Gabrielsson S and Gabrielsson S 2015 *The use of self-organizing maps in recommender systems* http://www.rslab.movsom.com/paper/somrs/somrs.pdf