KKU Res.j. 2015; 20(1) : 1-11
http://resjournal.kku.ac.th

Supervised Self Organizing Maps for Exploratory Data Analysis of Running Waters Based on Physicochemical Parameters: a Case Study in Chiang Mai, Thailand

Sila Kittiwachana[1*] *and Kate Grudpan*[1, 2]

[1] *Department of Chemistry, Faculty of Science, Chiang Mai University, Chiang Mai, 50200 Thailand*
[2] *Center of Excellence for Innovation in Analytical Science and Technology, Chiang Mai University, Chiang Mai, 50200 Thailand*
* *Corresponding Author: silacmu@gmail.com*

Abstract

This report demonstrated the use of a supervised self organizing map (SOM) for exploratory analysis of running waters based on their chemical criteria. Water samples from 10 different sites, representing 4 different water types – streams, a river, an irrigation canal and a sewage canal – were collected from some areas in Chiang Mai, Thailand, during 8-month period from May to December and analyzed for 16 physicochemical parameters. The samples were categorised into 8 classes (the 8 months from May to December) and 10 classes (the 10 sampling sites). This information was incorporated into the modeling using a supervised SOM methodology. The results were visualized using supervised colour shading and a unified distance matrix (U-matrix). The supervised SOM improved the correlation among the samples within group. It was possible to reveal the water sample clusters, either when organized according to the sampling times or sites. Moreover, all of the variation could be used for the analysis, eliminating the need to choose the specific dimensions or the number of principal components (PCs).

Keywords: *exploratory data analysis, supervised self organizing map (SOM), artificial neural networks (ANNs), principal component analysis (PCA), water analysis.*

1. Introduction

Exploratory data analysis (EDA) is an unsupervised modelling approach that can be used to analyse datasets to reveal and visualize their main characteristics in a way that is easy to understand (1). EDA looks for patterns in data and provide information about the relationship between samples and/or variables. EDA also helps answer whether any groups or clusters exist in the dataset and, if not, may indicate data of different origins. EDA differs from supervised modelling approaches, such as classification or pattern recognition, where the aim is to assign samples into some predefined groups. Principal component analysis (PCA) - a multivariate statistical technique that can be used to reduce dimensionality of a dataset, while retaining as much as possible of the variability of the original data - is one of the most common EDA methods (2). Although, PCA is simple and not an intensive computational model, it requires that the data was raised from a multivariate normal distribution. In addition, the optimum number of PCs should be carefully defined to ensure that most of the systematic variation in the data is taken into account in the analysis.

Self organizing maps (SOMs) are neural networks that offer an alternative to PCA (3). Like other neural network methodologies, SOMs employ adaptive learning algorithms, making them well suited for real systems. Training samples are used to train SOMs and then the results often represent by a two dimensional map. The samples are assigned to the map with the aim of preserving the relative distance between the training samples in the original data space as much as possible. By displaying as a map, the relationship between different types of samples in the training data can be reviewed. SOMs do not require the data to follow a multivariate normal distribution, a major advantage. SOMs have been applied as EDA methods in process monitoring (4), analytical chemistry (5), environmental analysis (6) and physicochemical information (7). SOMs could be an important alternative to PCA, particularly when the dataset includes several classes, because it allows full use of the map space. In contrast, only part of the space is used with PCA showing the main variation, so the samples are clustered into a smaller area.

Traditionally, SOMs are used for unsupervised exploratory data analysis. It is possible to employ these methods in a supervised mode. In some cases, SOM maps generated using unsupervised learning may not be well organised with respect to minor variation, but instead are strongly influenced by major variation in the dataset. Supervised SOMs are not restricted by this problem as long as groups or clusters of samples exist in the dataset and this additional information is provided during the training process (3). In this research, supervised SOMs were applied for exploratory analysis of water samples collected from running waters in some area of Chiang Mai, Thailand, based on their physicochemical criteria. For comparison, the results were compared to those using the unsupervised SOMs. Treatments of the supervised and unsupervised SOMs with results will be discussed in details. The SOM results will also be compared with the result using the traditional PCA analysis.

2. Experimental data

The data used in this study were literally from the report (8). Water was sampled from 10 sites in Chiang Mai, Thailand, representing 4 water types - streams, a river, an irrigation canal and a sewage canal. Sixteen physicochemical parameters were analysed: temperature, conductivity, velocity, pH, acidity, alkalinity, total phosphate, dissolved oxygen, biochemical oxygen demand, nitrate, ammonia, total phosphate, iron, copper, manganese and zinc. After eliminating missing or incomplete recording of data, the data was reduced to 10 parameters: temperature, conductivity, pH, alkalinity, total hardness, dissolved oxygen, nitrate, ammonia, iron and manganese. The names and locations of the study areas can be seen in Figure 1. The measurements were performed once a month from May to

December; therefore, generating 80 samples for analysis. It should be noted here that the missing data in the 6 incomplete recording parameters could be interpolated using PCA (9). However, using all of the parameters, the analysis results were not significantly different from those using only the complete parameters, implying that the missing parameters were not important for the analyses. Therefore, only the 10 complete recording parameters were used in this research.

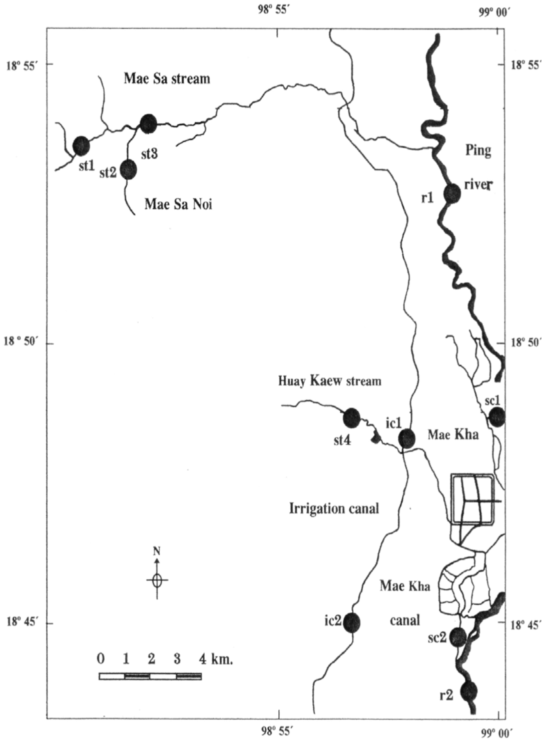

Figure 1. Map of the study areas (st: stream, r: river, ic: irrigation canal and sc: sewage canal). Image scanned from the report (8).

3. Methods

3.1 Self Organizing Maps (SOMs)

Self Organizing Maps (SOMs) are artificial neural networks that can be used to present the characteristic structure of a dataset in a low-dimensional display (10). SOMs usually involve a map, which is often represented by a two dimensional grid consisting of a certain number of units. The map is trained with training samples with the aim of locating the positions of the training samples on the map such that the relative distance between them in the original data space is preserved as much as possible. As a matter of fact, the training is to minimize the quantization error of the representation map (11). The unsupervised SOM algorithm has several stages and parameters that need to be set. These have been described in the literature (12); therefore, only the essential steps are described here.

Figure 2 shows a schematic diagram of how to generate an unsupervised SOM map. A trained map consists of a total of K ($=P \times Q$) map units, where P and Q are the number of rows and columns of the map, respectively. Each map unit k is characterised by a weight for each variable, resulting in a $1 \times J$ weight vector $\boldsymbol{w}_k$, where J corresponds to the number of variables. Therefore, each variable j has K weight units. To avoid ambiguity, it is important to note that the variables used in this context refer to the physicochemical parameters recorded from the water samples. In this research, the initial weight values were generated from randomly selected values from a uniform distribution within the measured range of variable j. This map will be trained, which means that each map unit weight will be interactively updated to become more similar to the vector representing the training samples. In each iteration t, a sample $\boldsymbol{x}_z$ is randomly selected from the data matrix $\boldsymbol{X}$, where z is a randomly selected integer from a uniform distribution in the interval 1 to I (where I is the number of samples) and is newly generated for each iteration *t*. The sample $\boldsymbol{x}_z$ is then compared to the weight vectors of each map unit $\boldsymbol{w}$. The map unit whose weight vector is most similar to the response vector of the currently selected sample is designated the 'winner' or 'best matching unit (BMU)' of the selected sample and becomes the centre of learning for that iteration. After that, the weight vectors in thedneighborhoods around the BMU are trained so that they are adapted to be more similar to the input sample $\boldsymbol{x}_z$. Although several indices can be used for determining the similarity, this study used the Euclidean distance index (13). Once the weight vectors have been updated, the entire process is repeated for $t = 1, 2, \ldots T$, using a randomly selected sample $\boldsymbol{x}_z$ for each value of *t*. A flow diagram showing the stages of SOM training can be seen in Figure 3. After the training is complete, the samples can then be mapped onto the units according to how similar they are to the corresponding units' weight vectors.

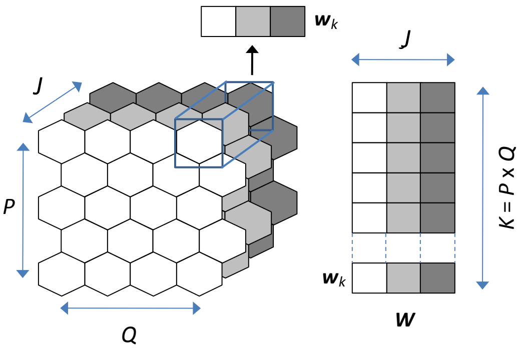

Figure 2. $P \times Q$ map with J weights containing a total of K map units and the corresponding weight matrix $\boldsymbol{W}$.

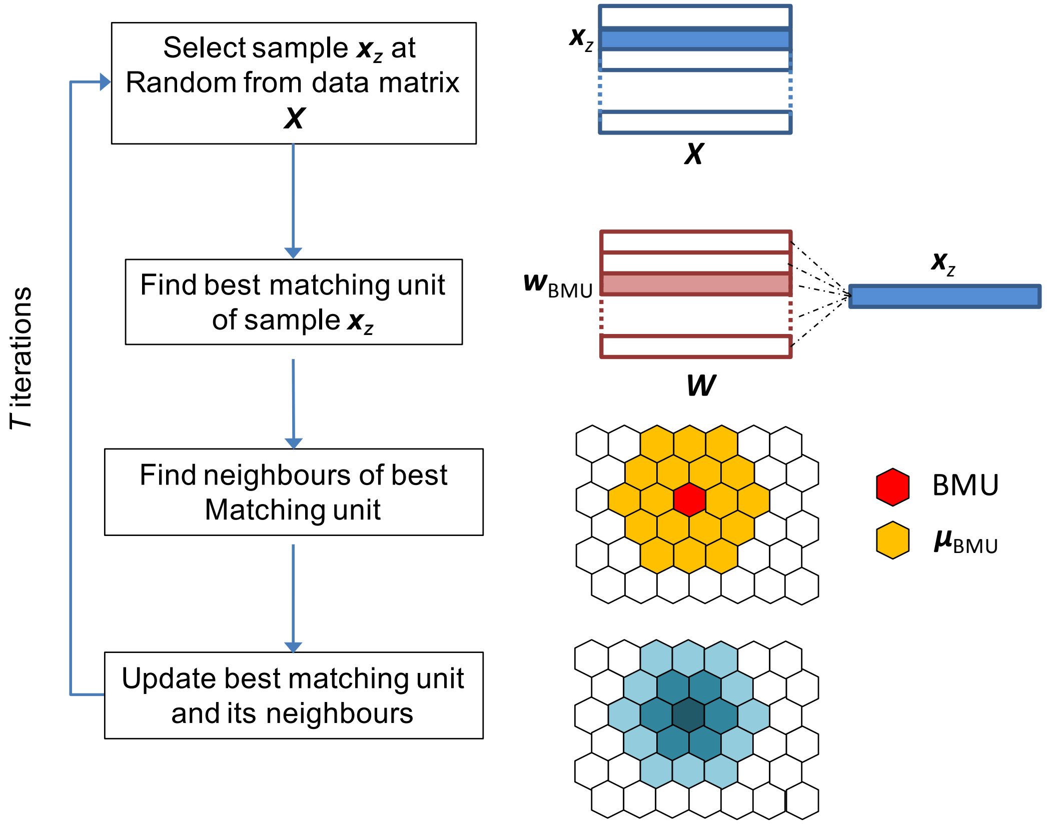
Select sample x_z at Random from data matrix X
Find best matching unit of sample x_z
Find neighbours of best Matching unit
Update best matching unit and its neighbours
T iterations
x_z
X
w_{BMU}
x_z
W
BMU
μ_{BMU}

Figure 3. Schematic of the SOM training process.

3.2 Supervised SOMs for water monitoring

SOMs can be categorised into two different groups according to how the models are trained: unsupervised and supervised. For unsupervised SOMs, as described in the previous section, only the information from measurements is used, whereas for supervised SOMs, an addition of weighting factors is applied to each sample as a constrainer prior to the training process. There are several ways to attach the weighing factors. In this work, the weighing factors are attached by additional class sample vectors containing the weighting factors corresponding to the sample memberships to the data matrix, resulting in a matrix called a supervised sample matrix. This variable sample matrix is augmented by these additional columns (Figure 4). For example, when the weighting factor used is 1, if a sample is in the 3rd (B) of 5 classes (A to E), then the additional class sample vector is [0, 0, 1, 0, 0] for that sample. This supervised weight matrices can then be trained in the same manner as an unsupervized SOM. Prior to the training process, each of the parameter vectors was standardised to its standard deviation so that all of the parameters were adjusted on the same scale to avoid the ones with greater elements dominating the analyses (14). The weighting factor used was the mean of all values in the variable sample matrix. In this research, although the class membership was included in the same way as for supervised SOMs, it was not used when locating the BMUs of the training samples. This was because that the aim of the supervised SOMs was to exploratorily analyze the water data. In a different way, some other related techniques such as learning vector quantization (LVQ) tend to use the class membership when defining the BMU and updates the representation map accordingly so that the boundaries could be drown between each of the sample classes. However, the purpose of such methods is for classification (15).

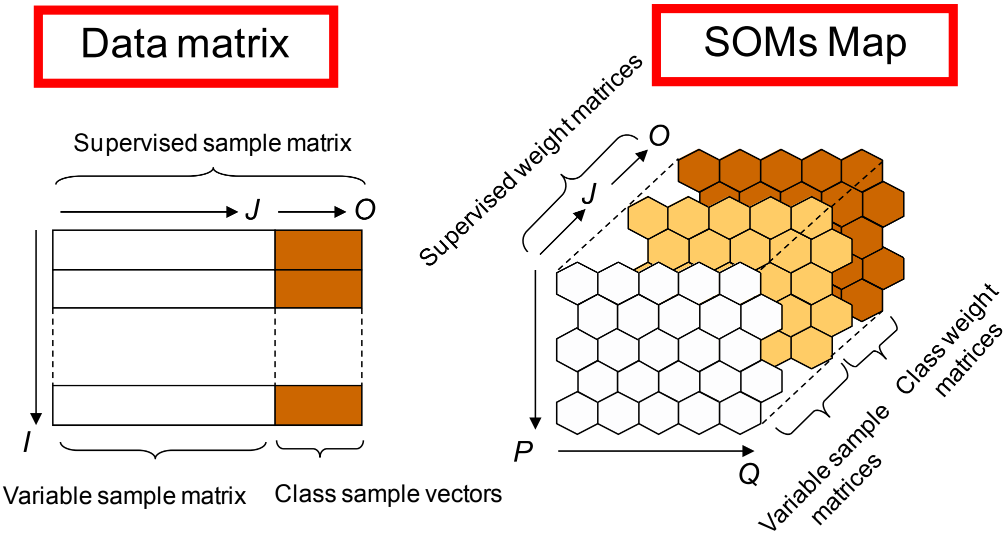

Figure 4. Supervised SOM training from I samples and J variables with O classes using P×Q trained map.

Each of the SOM maps was trained for 10,000 iterations to ensure that the map had a sufficient number of chances to learn about each sample. Some other parameters, such as the initial learning rate and initial neighbourhood width, were set following the recommended methodology (12). If the map space is two dimensional, there are a number of ways of visualizing the trained map and the relationship between samples, such as supervised colour shading (3) and unified distance matrix (U-matrix (12), which will be discussed in detail in the Results and discussion section.

4. Results and discussion

4.1 Exploratory analysis using unsupervised and supervised SOMs

The data used in this research were organised based on the sampling times (8 classes - May to December) and sampling areas (10 classes - 10 sampling areas). Four different SOMs were constructed and their supervised colour shading maps are shown in Figure 5 and 6 in order to visualize and explore the different structures within the data. Using supervised colour shading, each map unit is shaded according to the nearest class, which means that every unit is coloured according to their nearest BMUs. In this work, each map consists of a total of 15 × 20 map units, where the BMU for each sample are labelled accordingly. The number of units determined the resolution of the map. With more interpolation units (units that are not the BMU of any training sample and represent the transition between adjacent units), higher resolutions can describe the data in more detail, but training takes longer.

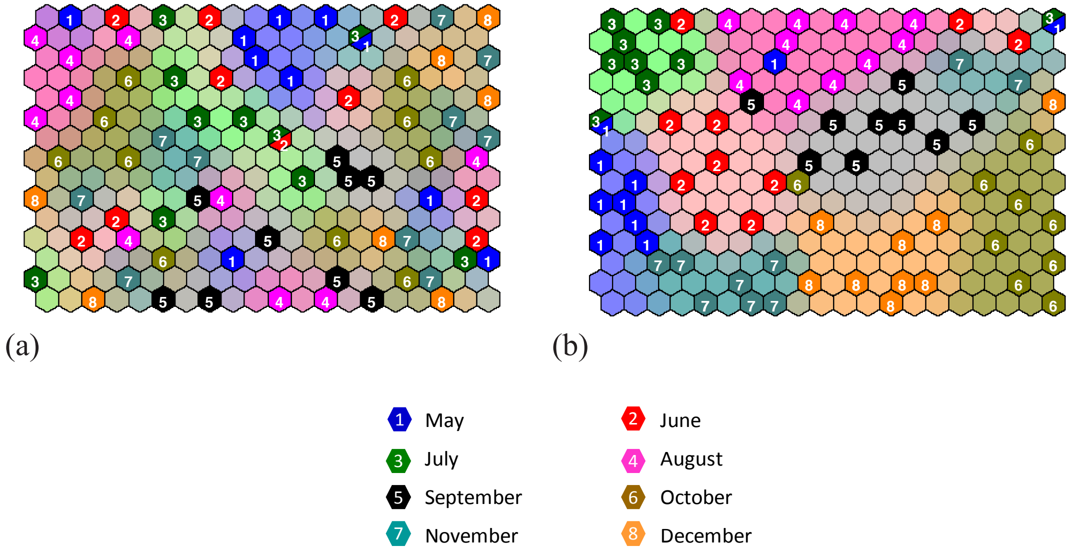

Figure 5. Colour shading visualization for the trained maps when each sample was labelled according to its sampling time. Each unit shaded according to the nearest class and the numbers represent the BMU of each sample. (a) unsupervised and (b) supervised SOMs.

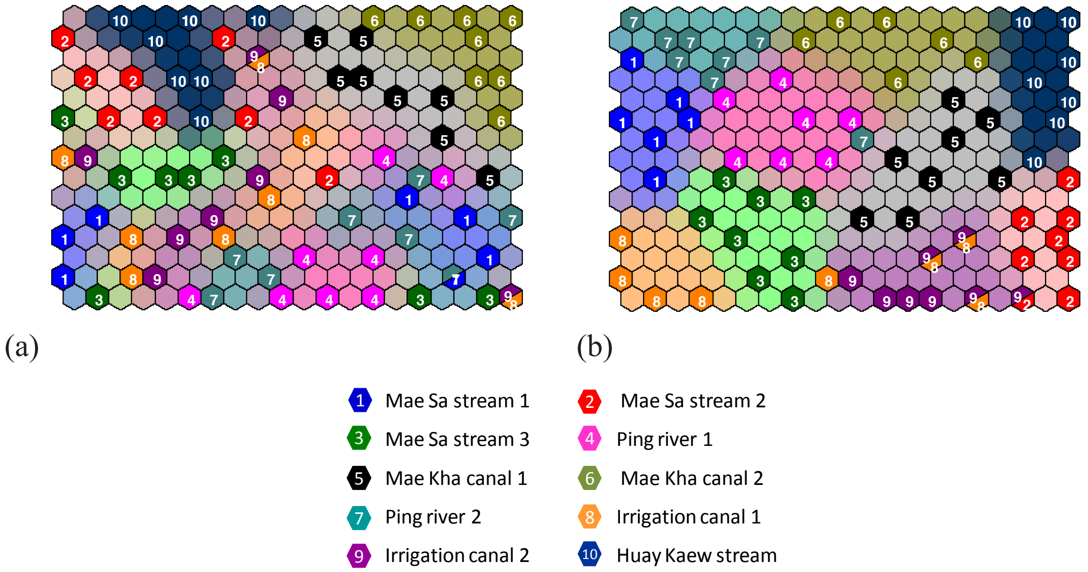

Figure 6. Colour shading visualization for the trained maps when each sample was labelled according to its sampling sites. Each unit shaded according to the nearest class and the numbers represent the BMU of each sample. (a) unsupervised and (b) supervised SOMs.

The organization of the data using unsupervised SOMs (Figure 5(a) and Figure 6(a)) is not very clear. The samples seem to be distributed across the maps and the sample clusters are not clearly distinguished. However, it is possible to notice some clusters in the map labelled according to the sampling areas (Figure 6(a)). For example, groups of samples collected from Mae Kha canal 1 and Mae Kha canal 2 can be observed on the top right corner, as well as a group of samples from the Huay Keaw stream on the top left side of the map. This indicates that the samples from these sites contain some physical or chemical properties that are uniquely different from the remaining sampling sites.

Clearer clusters can be observed in the supervised SOMs - 8 clusters in Figure 5(b), representing the sampling times from May to December; and 10 clusters in Figure 6(b), representing the 10 sampling sites. This reveals similarity among the sample groups. From the map visualizations, similar samples are located near each other and can probably be arranged into the same group. However, in Figure 5(b), some samples are mapped outside their clusters. For example, a May sample was incorrectly mapped into the August cluster. The November samples were divided into two groups, with the smaller group located on the top right side of the map. This group was also confused with some of the samples collected in May, June and July, indicating that some groups of water samples may not be well recognized by their sampling times. Although the characteristics of the water samples in the rainy and dry seasons may differ due to a dilution effect, the differing characteristics due to water sources contributed more to the model clustering. The clusters are more distinguishable in the map trained according to the sampling sites (Figure 6(b)). All of the 10 clusters from the 10 different sampling sites can be observed. Although some samples from more than one group fell onto the same cell, better separation could be obtained using a larger grid. In Figure 5(b), some of the samples collected from irrigation canal 1 were incorrectly placed into the cluster of samples collected from irrigation canal 2. This may reflect the fact that the samples came from the same water source - the irrigation canal - with the only difference their distance from the city. The water samples collected from irrigation canal 2 (further from the city) may have been less affected by the city and heavy traffic, but still shared common physio-chemical properties given the common water source and this may confuse the model (16).

Figure 7 shows the PCA scores plots of PC2 versus PC1 for the dataset used in this research. It is noted that the labelled colours in Figure 7(a) and 7(b) represent the sampling times and sites, respectively. Similar trends between the colour-shading unsupervised SOMs in Figures 5(a) and 6(a) and the scores plots can be observed - for example, the scattering groups (sampling times) and the proximity of groups in Figure 6(a) (Mae Kha canal 1 and Mae Kha canal 2). However, the SOM more clearly separates the groups than the scores plot, most likely because the information in the scores plot is limited to two or three dimensions at a time, whereas all the parameters were used to train the SOM, so there is no need to choose specific dimensions or the number of PCs. Also, the SOM allows the full use of the map space, whereas PCA uses only part of the space, so most of the variation in the data was clustered into a smaller region.

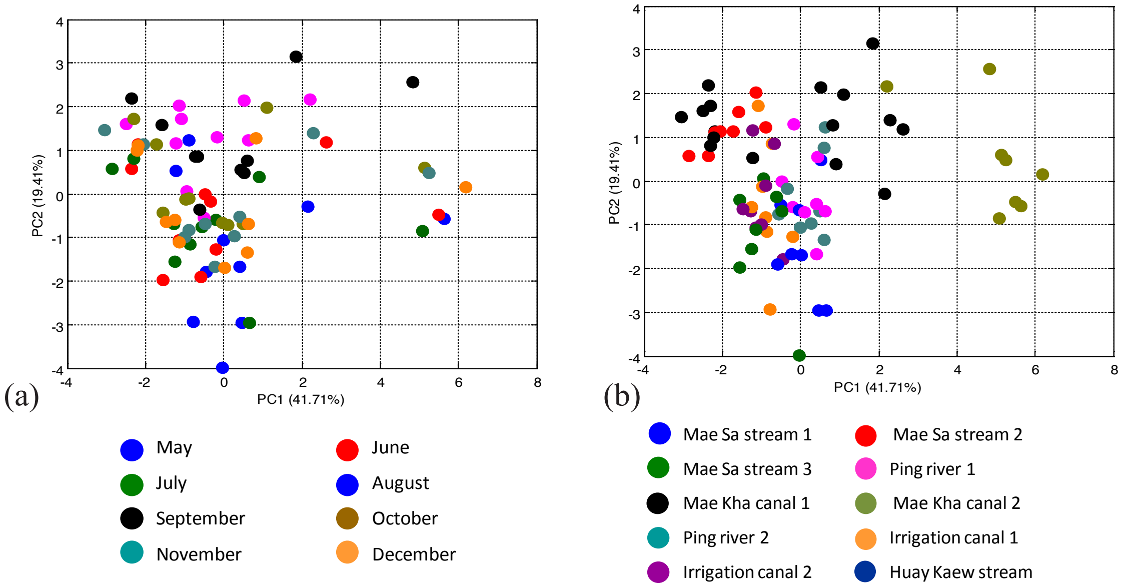

Figure 7. Score plots of PC2 against PC1 for the dataset in this paper, with each group coloured as in Figure 5. and 6. The samples were labelled according to their (a) sampling times and (b) sampling sites.

4.2 Interpretation of U-matrix

Besides the supervised colour shading maps discussed in the previous section, it is possible to visualize the trained map using unified distance matrix (U-matrix). The aim of U-matrix is to see the similarity of a unit to its neighbours and, therefore, reveal potential clusters presented in the map. The unified distance of each unit is calculated as the sum of the similarities between the weight vectors of a map, which in this case was the Euclidean distance. This generates map units whose neighbours are represented by similar weight vectors (e.g., in the middle of a cluster), a low value, and map units with very dissimilar neighbours (e.g., at the border of different clusters), a high value. If classes are present in the data, then the border between neighbouring clusters can be interpreted as a class border. In this research, the U-matrix values were converted to a colour scale (copper) and the SOM grids are plotted, shading each map unit with the scaled colour value.

From the U-matrix of the conventional SOM (Figure 8(a) and 8(c)), clusters of the samples were not clearly separated given no clear boundary among the groups could be observed. But for the supervised SOMs, the maps can be divided into regions that are similar to the results obtained from the colour shading. For both of the supervised trained maps, the boundary among the sample groups is slightly high and easier to observe. The number of regions correspond to the number of classes used for the supervised training. However, the clusters in Figure 8(c) are somewhat less noticeable that those in Figure 8(d), and so there may be some unsuspected samples, especially for the samples on the top right corner of the map. As mentioned previously, this may be due to the fact that the water conditions differ by too much when defined according to their sampling times. Considering both the supervised colour shading and U-matrix map, the map units can be assigned to their classes (the supervised colour shading Figure 6(b)), but the dissimilarity of the units to their neighbours are quite high, indicating that the groups were loosely organized (the U-matrix in Figure 8(c)).

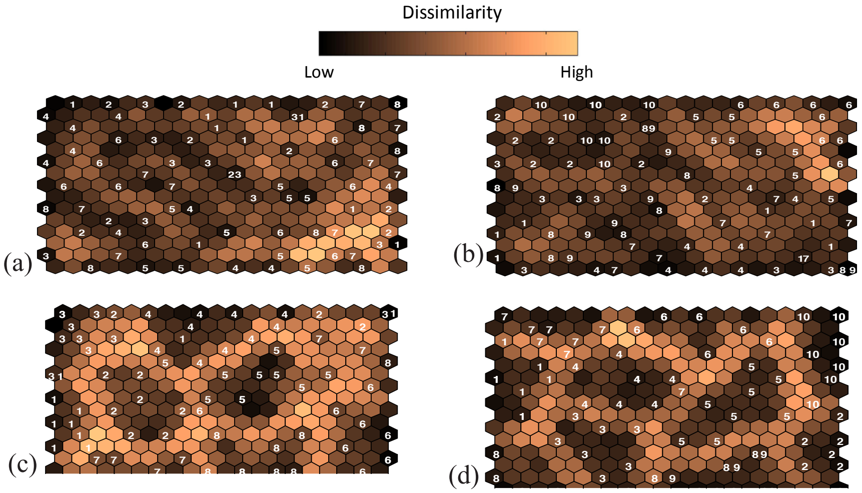

Figure 8. U-matrix visualization of the trained maps. The numbers represent the BMU of each sample where, for (a) and (c), the samples were labelled according to their sampling times and, for (b) and (d), the samples were labelled according to their sampling sites. (a) and (b) are unsupervised SOMs; and, (c) and (d) are supervised SOMs.

It is important to note that the aims of supervised colour shading and U-matrix are similar, since they both can be used for SOM visualization. However, the difference between them is that when using the supervised colour shading, the class memberships of the map units can be obtained, while the U-matrix attempts to investigate the similarity of a map unit to its immediate neighbours and also determines if the closest groups or clusters of samples are similar and can be connected or not.

5. Conclusion

SOMs are artificial neuron networks that can be used for data exploration. Supervised SOMs provide an advantage over unsupervised SOMs - the class membership can be optionally used during the training process to improve the correlation of the samples within the group. The clusters of water samples in this research were revealed using these methods. This investigation agrees with the previous study (8), in which water samples could be classified according to sampling area based on the chemical parameters provided. However, SOMs offer an additional advantage – visualizations can be obtained using various display techniques such as supervised colour shading and U-matrix. SOMs are much more intensive computationally compared to traditional statistical methods, such as PCA. But with today's computing power, these calculations can be performed in a few minutes on a desktop computer.

6. Notation

I, number of samples; J, number of measurements; $\boldsymbol{X}$, $I\times J$ data matrix; O, number of classes; K, $P\times Q$ number of map units where P and Q numbers of rows and columns, respectively; $\boldsymbol{W}$, $K\times J$ weight matrix where $\boldsymbol{w}_k$, $1\times J$ weight vector; T, number of iterations; t, t^{th} iteration; z, a randomly selected integer in interval 1 to I; $\boldsymbol{x}_z$, z^{th} $1\times J$ sample vector of $\boldsymbol{X}$; $\boldsymbol{\mu}_{\mathrm{BMU}}$, $1\times J$ the weight vector in the neighbourhood around the BMU.

7. Acknowledgment

This work was supported by the Faculty of Science, Chiang Mai University.

8. References

(1) Brereton RG. Chemometrics : data analysis for the laboratory and chemical plant. Chichester, UK: John Wiley & Sons; 2003.

(2) Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr. Intell. Lab. 1987;2 (1-3):37-52.

(3) Brereton RG. Chemometrics for pattern recognition. Chichester, UK: John Wiley & Sons; 2009.

(4) Kittiwachana S, Ferreira DLS, Fido LA, Thompson DR, Escott REA, Brereton RG. Self-organizing map quality control index. Anal. Chem. 2010;82(14):5972-5982.

(5) Lloyd GR, Wongravee K, Silwood CJL, Grootveld M, Brereton RG. Self organising maps for variable selection: application to human saliva analysed by nuclear magnetic resonance spectroscopy to investigate the effect of an oral healthcare product. Chemometr. Intell. Lab. 2009;(2):149-161.

(6) Bae MJ, Park YS. Biological early warning system based on the responses of aquatic organisms to disturbances: a review. Sci. Total Environ.2014;466:635-649.

(7) Kittiwachana S, Wangkarn S, Grudpan K, Brereton RG. Prediction of liquid chromatographic retention behavior based on quantum chemical parameters using supervised self organizing maps. Talanta 2013;106:229-236.

(8) Napattalung M. Quality of running water in some areas of Chiang Mai based on chemical criteria [Thesis]. Chiang Mai: Chiang Mai University, 1997.

(9) Jackson JE. A user's Guide to Principal Components. New York, US: John Wiley & Sons; 2003.

(10) Kohonen T. Self-Organizing Maps. New York, US: Springer; 2001.

(11) Arnonkijpanich B, Hasenfuss A, Hammer B. Local matrix adaptation in topographic neural maps. Neurocomputing 2011;74(4): 522-539.

(12) Lloyd GR, Brereton RG, Duncan JC. Self organising maps for distinguishing polymer groups using thermal response curves obtained by dynamic mechanical analysis. Analyst 2008;133(8): 1046-1059.

(13) Maesschalck RD, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. Chemometr. Intell. Lab. 2000;50(1):1-18.

(14) Kittiwachana S, Ferreira DLS, Fido LA, Thompson DR, Escott REA, Brereton RG. Dynamic analysis of on-line high performance liquid chromatography for multivariate statistical process control. J. Chromatog. A2008;1213 (2):130-144.

(15) Lloyd GR, Brereton RG, Faria R, Duncan JC. Learning vector quantization for multiclass classification: application to characterization of plastics. J. Chem. Inf. Model. 2007;47(4): 1553-1563.

(16) Sridang P, Danteravanich S, Thananuphaphphaisarn S, Durand C. Assessment of water pollution in Khun Thale swamp. KKU Res. J. 2008;13(9):1037:1048.