CHAPTER 3

Results and Discussion

3.1 Data pretreatment

Data pretreatment, normally, including outlier detection and data preprocessing. In this study, outlier detection based on statistic index called percentage of predictive ability (%PA), descripted in CHAPTER 2, was applied. Random 2/3 of samples were used as training samples and the rest as test samples and performed classification models based on multiple self-organizing maps. After that all samples were predicted based on 20 maps for 50 titrations. The samples with %PA less than 30%were removed because they present the too high dissimilarity between themselves and the rest samples in the same class. Moreover, the class with number of samples less than 10 should be combined with the class with bigger number of samples and present the similar properties, assume that should be the nearest sampling location. Therefore, after this procedural, there are 330 samples left from 10 provinces instead 877 samples and 20 provinces.

According to each soil property got vary rang as shown in Table 2.1 and Table 2.2, the screened data was standardized by average and standard deviation of each parameter as descript in CHAPTER 2, equation (2). The average and standard deviation values of only training samples was used with both training set and test set.

3.2 Model optimization for classification models

According to different algorithms of pattern recognition approaches, different algorithms for optimization the most suitable condition or controlled parameters were used. For example, Kohonen network based methods should optimize number of map units and number of iterations when the optimized factor of *k*-nearest neighbors was the minimum number of samples from the same class that would be used for defending group of samples thus the suitable optimization approaches for each method were applied.

3.2.1 Single and multiple self-organizing maps (SSOM and MSOMs)

To determine the optimum size and arrangement of the map for the SSOM and MSOMs, the algorithm named growing self-organizing map (GSOM) was applied. According to SOM algorithm, the value of map units in the initial maps, the map before training process, were random. The %MQE was fixed as a threshold that the model can represent the behavior of training samples but not over fitting. Therefore, the most suitable %MQE should be based on percentage of correctly classified (%CC) as shows in the following figure.

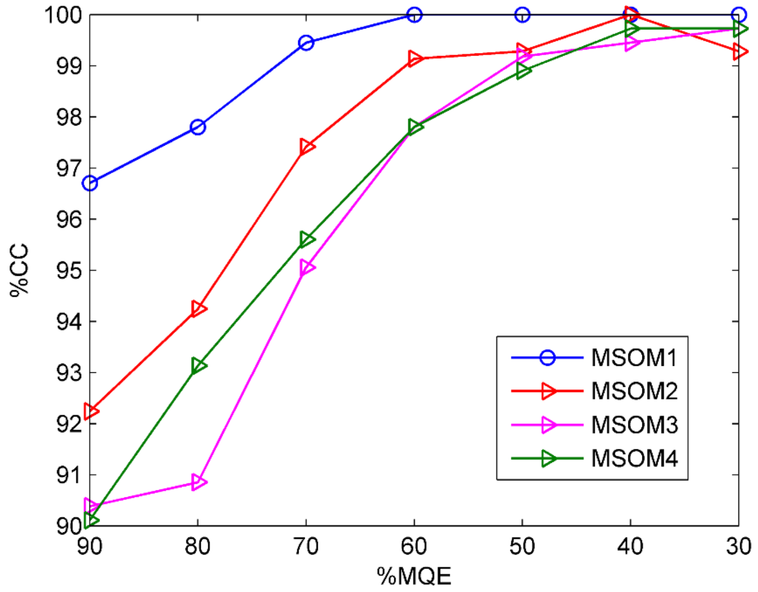


Figure 3.1 The %CCs of the training sets of the MSOM models using different %MQEs as thresholds for the GSOM

Figure 3.1 shows the development of percentage of correctly classified (%CC) of the growing self-organizing map (GSOM) models using the different values of the percentage of mean quantization error (%MQE) as threshold for the multiple self-organizing maps (MSOMs). All different MSOM algorithms, developed in this study, were applied for comparison. The training sets %CCs showed an increase trend with the decreasing %MQE. Using the GSOM, the map was grown in such a way to reduce the QE of the map units so that the topology of the data could be better preserved and the training samples were distributed across the trained map. This improved the predictive ability of the model, and therefore the %CCs showed the increasing trend. In most cases, the %CC values stopped improvement after the %MQE had reached approximately 50%. This suggested that the predictive abilities of the models were struggling to improve after this point. Even though, the %MQE, the threshold for stopping growing process, was fixed as the same value but the SOM model in each replicated always different in terms of size and location of projected samples. The size of SOM maps often vary as you can see below.

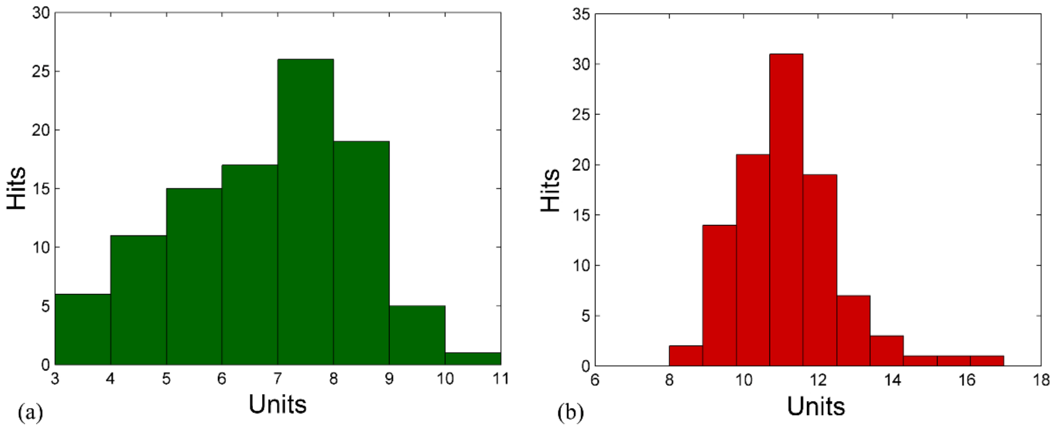


Figure 3.2 The histograms where the distribution of the smaller dimension, (a), and the bigger dimension, (b), of the trained maps after the development of GSOM

According to Figure 3.2, the single self-organizing map with SSOM1 was applied based on soil samples from the Amnat Charoen province. The threshold, percentage of mean quantization error (%MQE) was set as 50%. The calculation was repeated for 100 times. It can be seen that the optimum size of the trained maps from each calculation were not exactly the same. However, the favorite size of the trained map for this dataset was 7×11. In fact, this variation was expected and the trained samples could be located differently on each trained map. This could be because the initial maps were randomly generated and the selection order of the input samples during the training process could be different for the calculation. Therefore, the concern about dissimilar of each model, the model cannot represent the same characteristic of the training samples well, would be proved by suitable number of iteration in training process so the behavior of samples always the same.

3.2.2 Other linear pattern recognition methods

In this study, there were several linearity methods were applied such as linear discriminant analysis (LDA), partial least squares-discriminant analysis (PLS-DA) and soft independent modelling of class analogy (SIMCA). Firstly, the most suitable condition of LDA would not be optimized because it was classical discriminant method which the calculation of the model based on the Euclidean distance between centroid of each class and samples. Secondly, the parameter that affect classification performance of PLS-DA like number of principal components for the model should be optimized. The best number of PCs was optimized based on root mean square error of cross validation (RMSE-CV). The PLS-DA models based on many number of PCs were performed. The model with minimum value of RMSE-CV would be defined as the most suitable condition for the PLD-DA model. Finally, there are few parameters for SIMCA model that would be found the best condition including optimum number of PCs for SIMCA model and percent confident limited which relate to the flexible of the model. In this study, percent confident limited was fixed at 95% and the optimum number of PCs was determined based on cross validation algorithm. The dataset that applied with all methods above should have number of samples more than number of parameters.

3.2.3 Other non-liner pattern recognition methods

According to the developed algorithms, MSOMs, are neuron network based methods so some none-liner classification techniques both discriminant based and neuron network based were applied like quadratic discriminant analysis (QDA) and *k*-nearest neighbors (*k*-NN), discriminant methods, counter propagation network (CPN) and supervised Kohonen network (SKN), neuron network models. Firstly, the optimized condition of QDA which is a discriminant method, would not be optimized as same as LDA. Secondly, *k*-NN do not require complicate statistic computations as well as other discriminant methods. Only one parameter that should optimize is *k* which is the number of samples in the same group with the shortest distant based on Euclidean distance between a test sample then the test sample will be defined as that group. Finally, there were several parameters for CPN and SKN which are Kohonen based method such as the number of iteration and the number of map units. According to the script functions of CPN and SKN were created by others group researches, is was too complicated to apply GSOM for optimized the suitable condition of CPN and SKN. Therefore, the most effective condition would be fined by vary one parameter at a time.

3.3 Chemometrics

The methods which were applied in this study could be classified into data exploratory and classification propose. PCA was applied for obvious pattern of the data as same as visualization map, supervised color sheading, of single self-organizing. Some pattern recognition methods were performed to compare in terms of classification performance with MSOMs such as LDA, QDA, PLS-DA, SIMCA, CPN, SKN and *k*-NN. It is possible to classify these methods into two groups including linear and non-linear methods. LDA is a classical linear pattern recognition as well as PLS-DA and SIMCA which are well known linear methods. In addition, some non-linear techniques not only QDA and *k*-NN, classical pattern recognitions, but also CPN and SKN, neural network were applier.

3.3.1 Exploratory data analysis (EDA)

After data treatment processes, PCA with common and iterative algorithm, NIPALS, was performed to exploratory the behavior of the treaded data using in-house Matlab scripts. There are two criterial for labeling the soil samples. The first one is dividing data into two groups based on regions of soil sample collections. The samples, were sampling from northern region, were labeled by red whereas sampling from northern region were labeled by blue color. Another criterion is defining data into ten classes based on provinces of soil sampling areas. All soil samples were labeled with ten different symbols and colors. The results of PCA were presented as PCA score plots with first few principal components (PCs).

3.3.1.1 Principal component analysis (PCA)

Firstly, the soil data that the writer received from Ubon Ratchathani Rice Research Center, Ubon Ratchathani, Thailand totally 877 samples with 15 parameters. There are several samples that at least one missing data of some parameters. Therefore, the samples with missing value were removed. There were 704 samples left and the pattern of the data could be seen in PCA score plot as below.

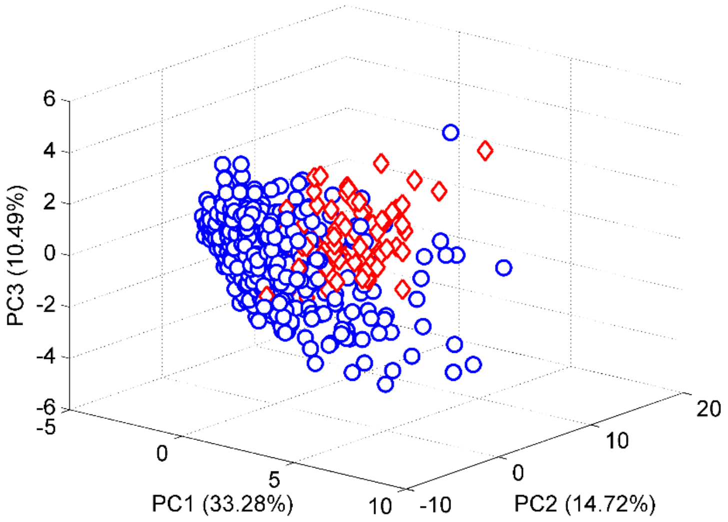
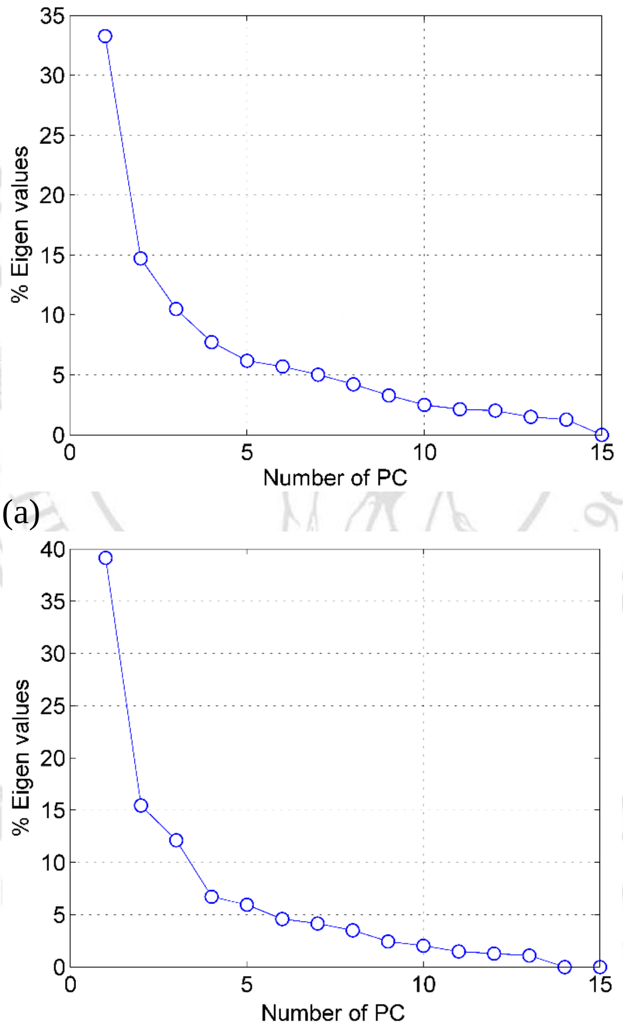


Figure 3.3 Score plot of the first three PCs where the samples were labeled according to the regions they were collected, blue represent northeast region and red represent north region.

Although some missing data were ignored, some soil samples still contain some gremlins values of some parameters such as soil samples with pH 28 or summation of %Sand, %Silt and %Clay of some soil samples are higher than 100. Therefore, the first three components contain less than 60% of the data variation and the others contain less than 10% of variation as show in Figure 3.4 (a).



(a)

(b)

Figure 3.4 %Eigen values of the 704 (a) and 330 (b) soil samples which are contained by each principal component (PC)

As the result, %PA which is the statistic index based on majority vote, was applied for outlier detection. The samples with %PA less than 30% were removed. Moreover, the class with number of samples less than 10 were combined with the classes with bigger number of samples and present the similar properties, assume that should be the nearest sampling location. Therefore, after this procedural, there are 330 samples left for 10 provinces instant 877 samples of 20 provinces. According to PCA scores of 330 soil samples, the first three components contain 66.74% variation of the data, Figure 3.4 (b), and the rest PCs contained approximately 6% and lower %variation.

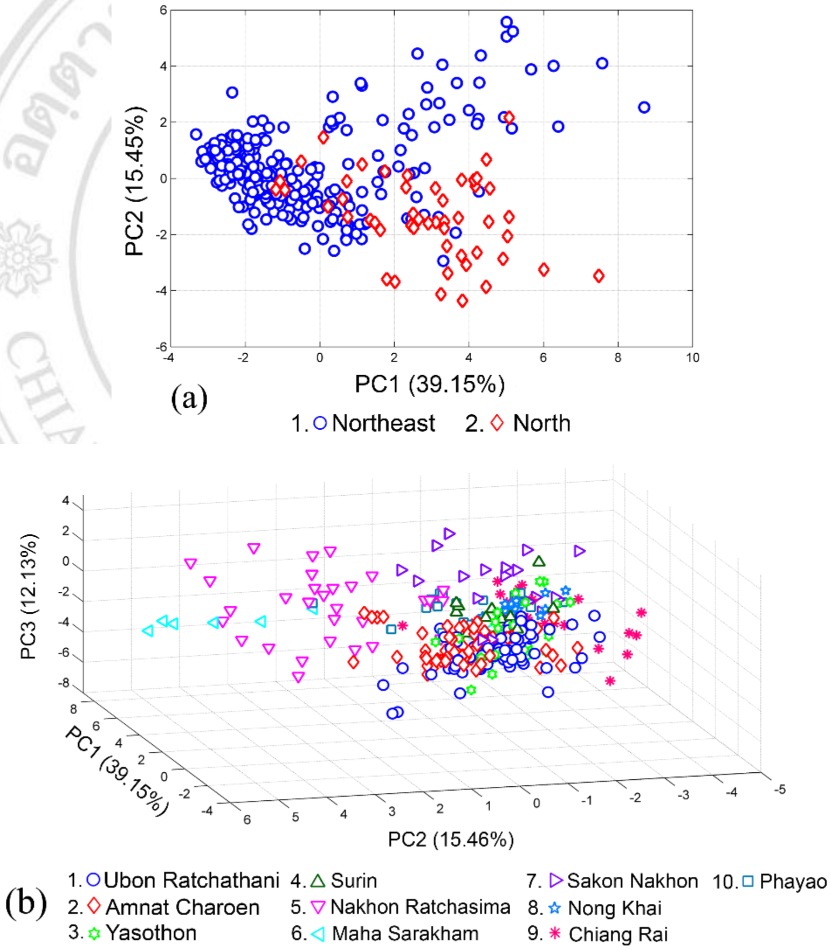


Figure 3.5 A score plot of the first two PCs where the samples were labeled based on the regions they were collected (a) and a score plot of the first three PCs where the samples were labeled according to the provincial origins (b)

The PCA score plots shown in Figure 3.5. The data labeled based on regions of collected areas demonstrated in Figure 3.5 (a), red color represent samples from north region (N) whereas blue color belong to samples from northeast region (NE). Although clear clusters among the samples in the first two PCs could not be identified, it is possible to differentiate between the soils samples from the northern (N) and northeastern (NE) regions. When the samples were labeled according to the provincial sources and shown using the first three PCs in Figure 3.5 (b), the three PCs model still could not adequately describe this dataset. Even though, it is possible to see that which parameter effects on the different position of soil samples in loading plot of PCA as follow.

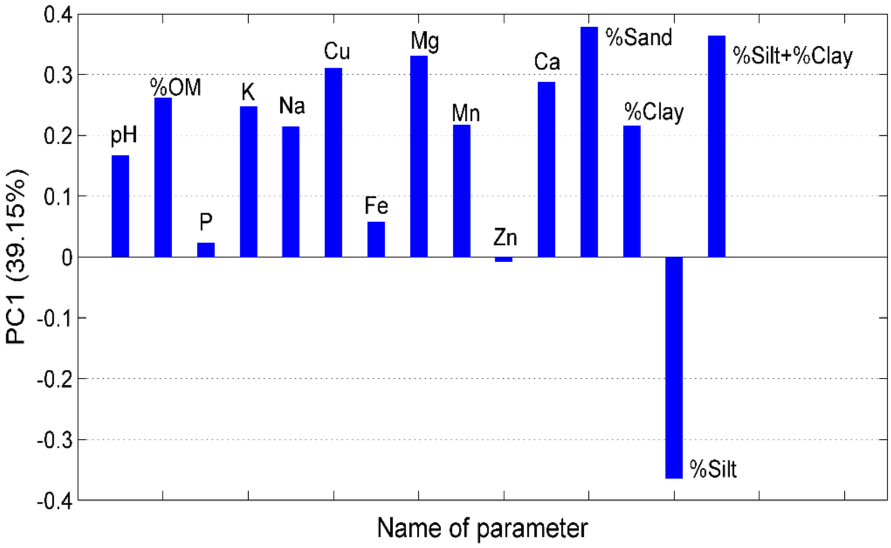


Figure 3.6 Loading plot of PCA model based on 330 soil samples in PC1

According to the loading of each parameter of the first principal component (PC1), the most powerful parameter that effects on the score values of the soil samples is %Silt, followed by %Sand (the second priority). Therefore, it seems that the physical properties, %Silt and %Sand, play more important role than other chemical properties such as pH and major and minor element concentrations which are represented in lower absolute loading values in the first principal component (PC1). On the other hand, Fe content, pH, Na content and %OM affecting on samples characteristic based on higher value of loading in the second principal component (PC2). It means that the different between samples in the first PC were effected by some physical properties (%Silt and %Sand) when the behavior of the samples were based on some chemical properties such as pH, %OM and some chemical concentration (Fe and Na).

From the literatures, soil could possess different characteristics if they are in different geographic areas. The unique could be due to the managements and the types of vegetation that covers. Other natural factors such as their parent rock materials and variations in the climates could also contribute to the unique soil properties. Therefore, it is possible to classify soil samples based on physical properties and chemical properties. Based on the loading of PCA, %Silt and %Sand are the two highest affecting parameters on PCA score in PC1 agree with some studies that soils in the northeastern region of Thailand, are usually sandy with a substantial amount of salt deposit whereas darker clay soils are expected to be found in the northern region [2].

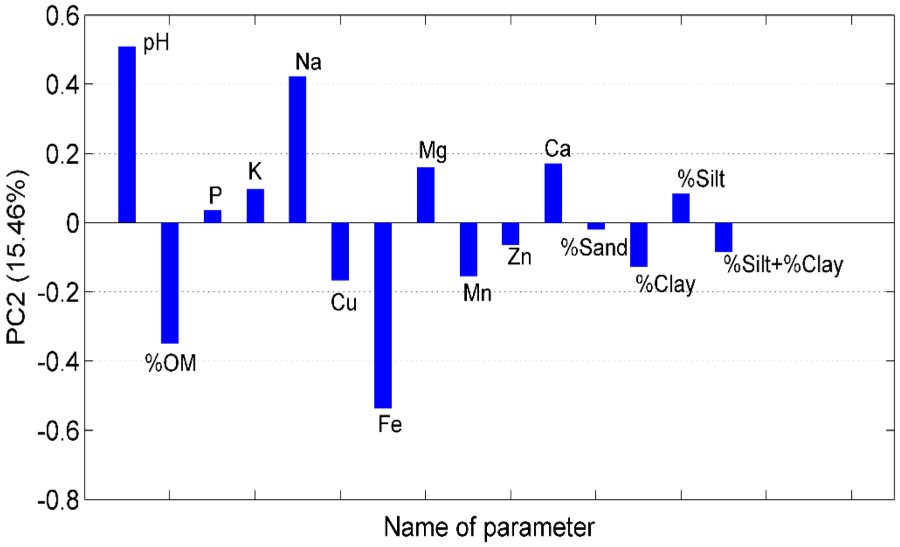


Figure 3.7 Loading plot of PCA model based on 330 soil samples in PC2

Moreover, pH, %OM and some element concentrations with high loading value in PC2 made the soil samples different, agree with literature reports, northern areas are typically rich of iron (Fe) whereas northeastern areas are normally rich of salt (Na) [2]. Accordingly, several methods and criteria could be adopted to establish a soil classification system [4].

Although the behavior between pattern of soil samples from two regions can be seen, it is not obvious that there were some provincial unique properties of soil samples, could be used for soil classification based on provincial sampling areas. The unclear characteristic may relate to PCA algorithm that some data was assigned as noise or residual (Figure 2.5) and PCA protocol usually require for normal distribution. Moreover, the relationship or correlation of the variation in the data are not linearity. For that reason, other pattern recognition methods were performed to fine the most suitable method for this soil dataset.

3.3.1.2 Single self-organizing map (SSOM)

Based on quite complicated data and non-normal distribution data, the neuron network method called self-organizing map may suitable. This technique usually requires for optimization parameters for the most suitable condition for modelling. Using the GSOM with 50% MQE threshold, as descripted in CHAPTER 2 issue 2.4.1, the SOM visualization of SSOM1 and SSOM2 with the BMUs indicated are shown in Figure 3.8 (a) and (b), respectively. Note that the class information was included where all units were shaded using the colors according to the nearest class in Figure 3.5. The labeled No. 1 and 2 on SSOM1 (Figure 3.8 (a)) represent for soil samples in the northern and northeastern region when No. 1-10 represent to soil samples in each province as presented in Table 2.1. SSOM1, a separation could be observed although there is a small isolated group of some soil from the north. The SSOM2 was a more complicated task than the SSOM1. Here, more samples were misclassified. The area of each class on the maps was roughly proportional to class size since there was an equal probability of each sample being trained.

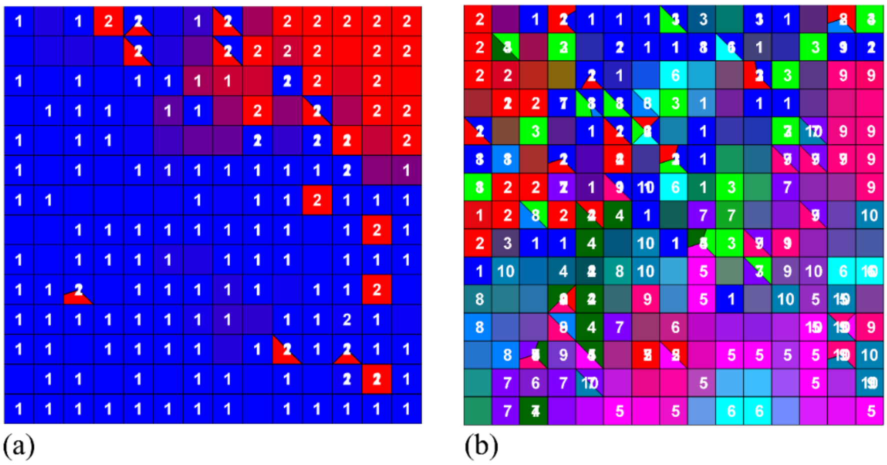
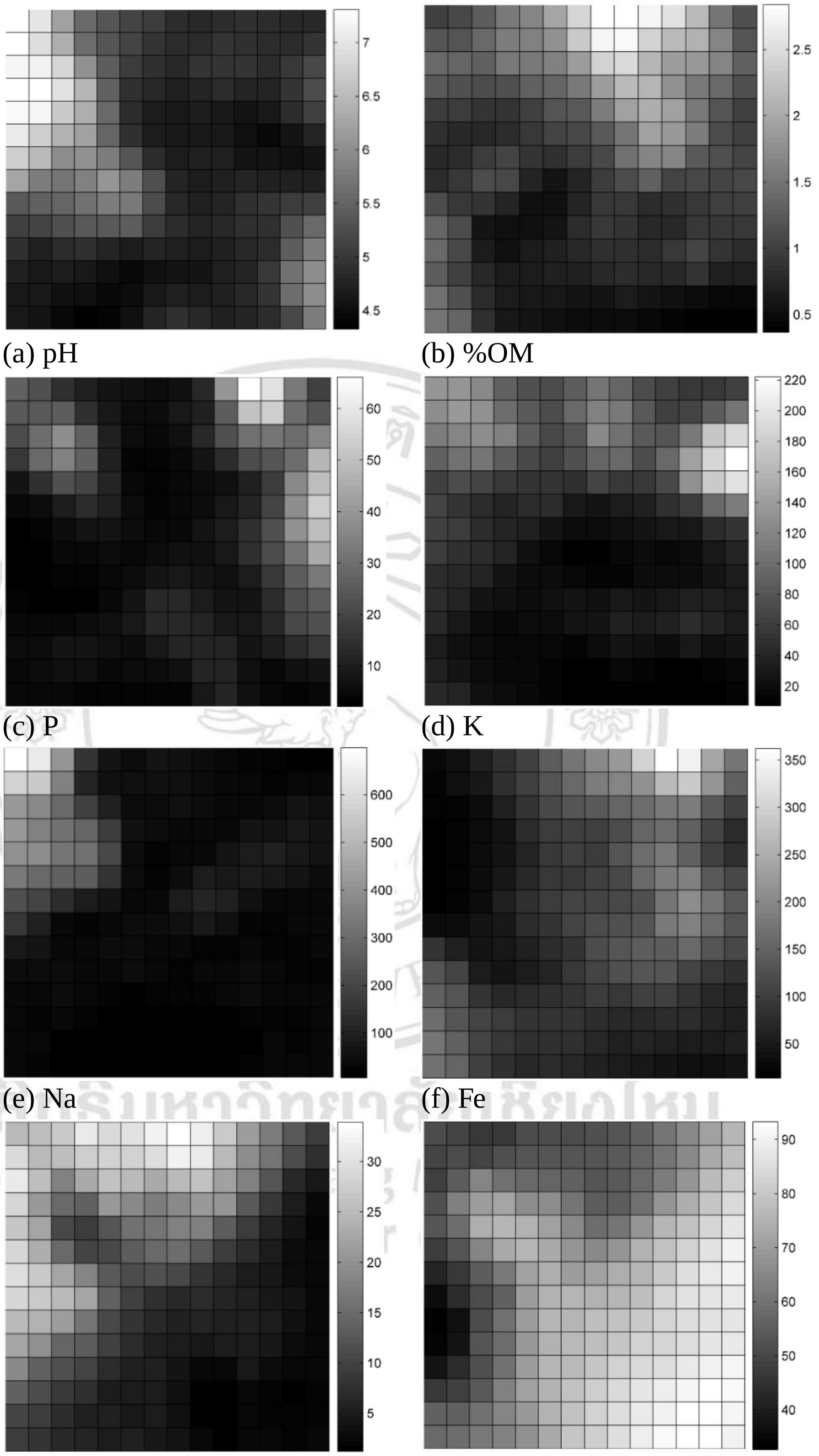

(a) (b)

Figure 3.8 SOM visualization of class structure with the BMUs indicated and all units shaded according to the nearest class for (a) SSOM1 and (b) SSOM2 using the GSOM with the 50% MQE as stopping criterion. The shading colors were the same as for the symbols in Figure 3.5

According to self-organizing map approach, the pattern of each parameter could be seen based on component planes which are the layers of trained map. Therefore, the component planes should have the same size with visualization map. Moreover, the parameters which are affecting on the location of soil samples should present similar pattern of samples projection. The following are the component planes of some parameters based on SSOM1.

The component planes based on SSOM1, Figure 3.9 (a), could be used for presenting the differences of some soil samples from the northeast (NE, blue, 1) and north (N, red, 2) of Thailand. Based on pH component plane, Figure 3.9 (a), shows that some soil samples from northeastern areas are slightly acidic soil than northern soil.



(a) pH

(b) %OM

(c) P

(d) K

(e) Na

(f) Fe

(g) %Sand

(h) %Silt

Figure 3.9 Component planes of pH (a), %OM (b), P (c), K (d), Na (e), Fe (f), %Sand (g) and %Silt (h) from SSOM1 trained map (Continued)

In terms of major soil nutrition seems some soil samples from northern areas have higher content of %OM, available P and K, Figure 3.9 (b), (c) and (d) respectively. Moreover, the pattern of some soil elements such as Na and Fe content, Figure 3.9 (e) and (f) in order, agree with literatures that some northeastern soil have got higher Na (salinity soil) than northern areas whereas some northern samples are luxuriant of Fe when compare with northeastern soil. On the other hand, the correlation between parameters could be seen form the pattern of the component planes as well. For example, the component plane of %Sand and %Silt, Figure 3.9 (g) and (h), presenting inverse relation among each other.

Although, some component planes of SSOM1 could present the different of soil samples from the two regions (NE and N), some samples have disagreement properties with the samples in the same class or region. For example, pH of some northeastern soil samples, Figure 3.9 (a), have slightly higher pH, shaded by dark color, than other soil samples from the same region. Therefore, it should be seen easier in the component planes of northeastern soil map. As the reason, the MSOMs were also applied for visualization as well.

3.3.1.3 Multiple self-organizing maps (MSOMs)

MSOMs algorithm that were developed in this study based on supervised approach. Each map was established separately. Therefore, supervised color sheading of MSOM1, labeled based on sampling regions (R1 and R2), cannot be used for data exploratory but the behavior of each map could be seen in unified distant matrix (U-matrix) map visualization and the pattern of each parameter was seen in the component planes. Moreover, the supervised color sheading maps of MSOM2 which shaded based on collecting provinces (P1-P10).

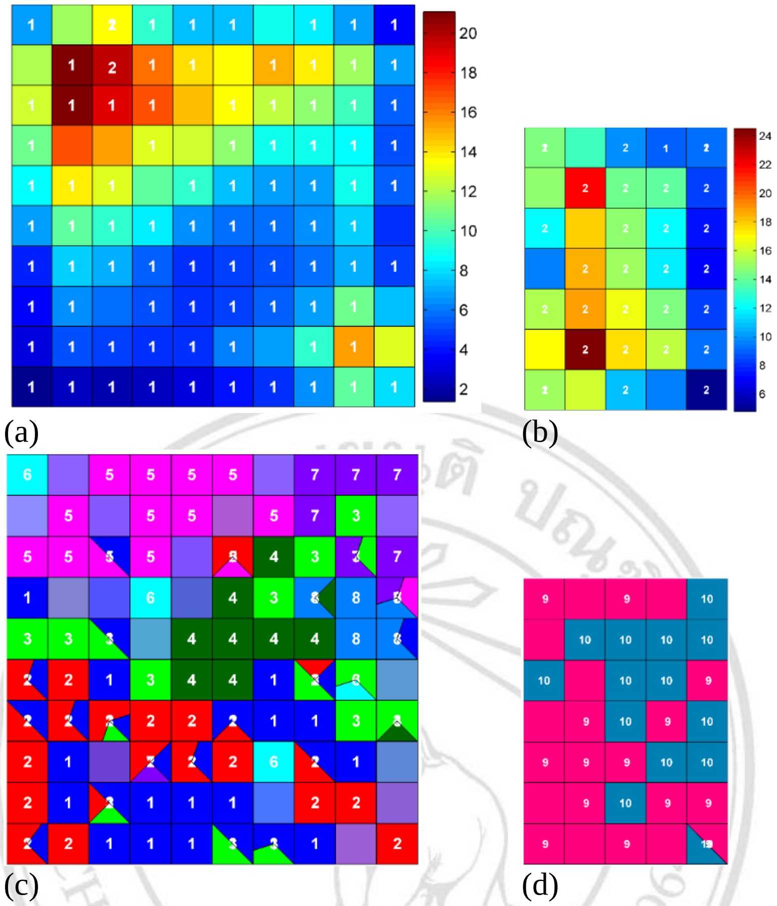


Figure 3.10 Maps visualization of MSOM1 and MSOM2 class structure with the BMUs indicated and all units shaded based on dissimilarity, (a) and (b), and provincial sampling areas, (c) and (d), of northeast (left column) and north (right column) map using the GSOM with the 50% MQE as stopping criterion.

As presenting in Figure 3.10, there are two maps of MSOM1 U-matrix map visualization and MSOM2 supervised color shading maps, northeastern Figure 3.10 (a) and (c) and northern Figure 3.10 (b) and (d) respectively. U-matrix maps of MSOM1, Figure 3.10 (a) and (b), there are some different groups of soil samples with high (red shading) and low dissimilarity (blue shading). It means among northeastern soil samples could be separated into smaller groups. On the other hand, in supervised color shading maps of MSOM2, Figure 3.10 (c) and (d), seems most soil samples in each map were projected close to the samples from that same province. For example, No.4 with dark green, No.5 with magenta and No.7 with purple sheading which represent for soil samples from Surin, Nakhon Ratchasima and Sakon Nakhon province in order whereas the map of SSOM2 could not be seen. Therefore, the developed algorithm give a better data exploratory results that traditional approaches such as PCA and single self-organizing map. Moreover, the divided maps could present more detail of parameters correlation as show in the following figure.

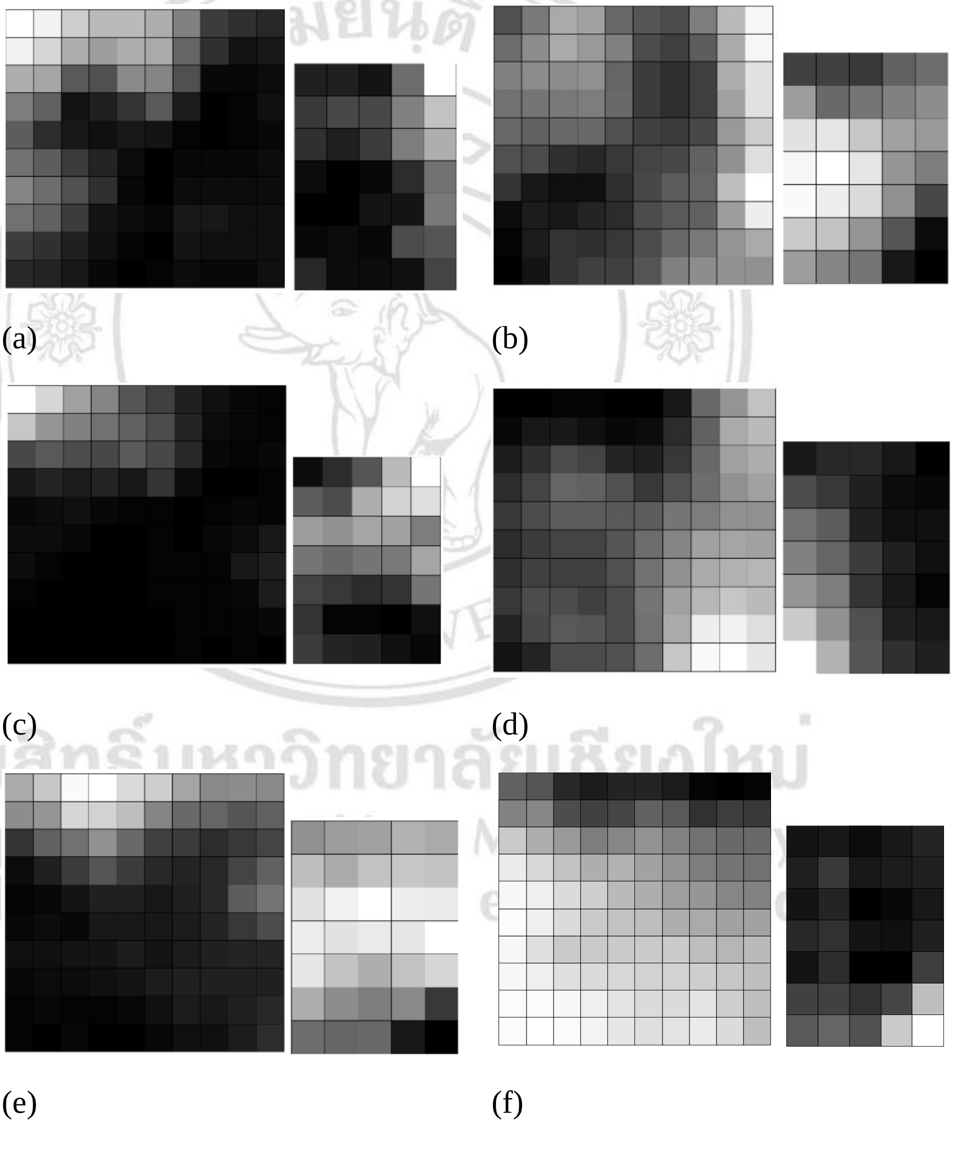


Figure 3.11 Component planes of pH (a), %OM (b), Na (c), Fe (d), %Sand (e) and %Silt (f) from MSOM2, northeastern (left) and northern (right) trained maps

Based on the MSOM2 component planes, there are some relationship between some parameters as presented Figure 3.11. For example, the inverts correlation between %Sand and %Silt can be seen Moreover, this correlation between Iron (Fe) content and pH could be obvious in the component of northeastern area. When the inversion correlation between parameters that can be seen in component planes of SSOM1 is only the correlation between %Sand and %Silt. Furthermore, an unclear inversion relation between pH and %OM was noticed because some high pH soil samples such as at the top left corner of Figure 3.11 (a) also have a bit low %OM as present in Figure 3.11 (b). According to these planes, it is possible to see the different values between parameters of samples in each class. For example, it was obvious that the content of sodium (Na) of samples at the top right corner of northern soil map, Figure 3.11 (c), which soil samples form Phayao province (No.10 in Figure 3.10 (d)) were projected, have higher concentration than the rest samples which almost are the samples from Chiang Rai province (No.9 in Figure 3.10 (d)). On the other hand, the samples from Phayao province (No.10 in Figure 3.10 (d)) in Figure 3.11 (d) have lower content of iron (Fe) than Chiang Rai samples (No.9 in Figure 3.10 (d)). In this study, there are four different algorithms of multiple self-organizing maps, were developed (MSOM1-4). Even though, there are only MSOM1 and MSOM2 which are more suitable for data visualization than the two rest algorithms. In addition to, too many maps component planes based on number of class samples would make the interpretation more complicate. Therefore, in this work, map visualization of MSOM3 and MSOM4 are not shown.

According to the results, it is possible to claim that the developed algorithms, some multiple self-organizing maps which are MSOM1 and MSOM2, give the better results than basic method such as PCA for exploratory data analysis. Moreover, for data visualization of self-organizing map approach, the MSOMs show more obvious classification gropability than traditional SSOM. As the reason, for the classification propose, it could be expected that MSOMs will give better pattern recognition performance than typical SSOM and other classical classification methods.

A SSOM model was constructed for both data exploratory and pattern recognition. Supervised color sheading maps were applied as the basic map visualization for exploratory data analysis. The colors of map units based on each criterion were followed the PCA. The distant between each map unit and the centroid of each class of samples based on Euclidean distant were calculated. The map units would be shaded by color of the class with minimum distant and the position of each sample was shown as number of the class that each sample belong to.

3.3.2 Classification models

In this study, soil samples could be classified using two categories such as regional class areas which there are two classes, northeastern (R1) and northern (R2), and provincial class areas, including 10 sampling locations of soil samples (P1-P10). Therefore, the classification results were separated into two sessions which make the interpretation and comparison easier as follow.

3.3.2.1 Classification based on regions (NE, R1 and N, R2)

According to Table 3.1, it is possible to see the performance indices of the proposed SOM algorithms. In all cases, as expected, %correctly classifieds (%CCs) of the auto-predictive training sets were overall higher than %CCs of the test sets. Based on regional classification categories including only two classes of samples, the different between classification performance of almost applied methods are not obvious, accept SIMCA and SSOM1 model which gave %CC less than 90% when the rest present above 90 %CC. Considering the situation when there were 2 classes in the dataset, the %CC from MSOMs model which is MSOM1, provided a greater %CC than that from a SSOM model which is SSOM1. For the 2 class membership situations (N and NE), although an acceptable %CC of the test set could be provided from the SSOM1 (84.28%), this classification accuracy was significantly worse than those from the classical Kohonen networks; counter propagation network (CPN; 94.84%) and supervised Kohonen network (SKN; 95.45%). Moreover, the classical linear and non-linear approaches such as LDA, QDA and PLS-DA give quite close number of %CCs, 93.33%, 91.52% and 94.24% respectively. It means that the classification, based on regions, has linear behavior because both of the linearity and non-linearity methods gave the similar classification performances. That is the reason why both linear and non-linear methods presented the similar classification performances. Furthermore, among the classical non-liner methods like QDA and *k*-NN which are based on distance between center of each class sample for assigning group of samples, CPN and SKN which are neuron network based method also present the comparable predictive performance.

Table 3.1 %PA, %MS and %CC for the soil data with 2 classes using CPN, SKN, *k*-NN, LDA, SIMCA, PLS- DA, SSOM and MSOMs

Methods	%PA[a]		%MS[a]		%CC	
	Training	Test	Training	Test	Training	Test
CPN	98.93	98.93	97.86	93.50	100	94.84
SKN	99.39	94.71	98.83	94.12	99.69	95.45
k-NN	96.11	89.76	96.12	94.00	95.95	90.29
LDA	87.15	85.95	82.64	82.93	93.94	93.33
QDA	88.33	85.01	83.99	82.21	93.64	91.52
SIMCA	88.40	85.61	89.45	87.13	89.40	87.88
PLS-DA	94.16	93.30	97.48	96.67	94.54	94.24
SSOM[b]						
SSOM1	84.48	70.80	98.58	97.05	85.75	84.28
MSOMs[b]						
MSOM1	100	99.68	99.85	99.50	94.88	94.29

[a] %PA and %MS are the means of all samples.

[b] 50% mean quantization error was set as a threshold for the growing self-organizing maps.

3.3.2.2 Classification based on provinces (P1-10)

Another criterion, when the soil samples were further separated into 10 classes, were shown in the following table. All %correctly classifieds (%CCs) of the auto-predictive training sets were overall higher than %CCs of the test sets same as the result from two regions in Table 3.1. Nevertheless, the different between the predictive ability of all methods were obvious. For example, LDA, QDA and traditional single self-organizing maps (SSOM2) presented poor classification performances due to less than 40 %CC although *k*-NN, SIMCA, PLS-DA and MSOM2, could be seen in visualization maps, Figure 3.10 (b), give slightly better %CC but still less than 60%.

Table 3.2 %PA, %MS and %CC for the soil data with 10 classes using CPN, SKN, *k*-NN, LDA, SIMCA, PLS- DA, SSOM and MSOMs

Methods	%PA[a]		%MS[a]		%CC	
	Training	Test	Training	Test	Training	Test
CPN	92.48	65.03	86.00	65.06	98.18	70.30
SKN	96.43	68.73	93.14	67.74	99.39	74.24
k-NN	78.49	50.41	79.41	76.12	78.35	50.84
LDA	33.76	31.27	87.55	87.28	37.88	33.94
QDA	35.86	31.84	89.34	85.64	40.00	34.85
SIMCA	63.49	44.49	76.67	62.53	69.52	49.37
PLS-DA	55.04	54.08	83.87	83.98	53.94	53.33
SSOM[b]						
SSOM2	38.28	30.17	87.96	92.47	35.42	32.66
MSOMs[b]						
MSOM2	59.67	51.84	76.32	71.05	54.95	44.24
MSOM3	96.27	68.78	92.89	66.57	95.88	73.03
MSOM4	97.15	70.52	95.04	70.03	97.87	76.36

[a] %PA and %MS are the means of all samples.

[b] 50% mean quantization error was set as a threshold for the growing self-organizing maps.

The expectable predictive performances were %CCs of CPN, SKN, MSOM3 and MSOM4 which are higher than 70% which are possible to see that the MSOMs model, MSOM4, again outperformed the SSOM model (SSOM2). The %CCs for both training and test sets appeared worse when the SSOM was extended to classify the soil samples according to their provincial areas (SSOM2). This implied that faithful classification results could be provided from the SSOM only if the classification task was not too complicated.

The test set %CC from MSOM2 was just slightly higher than that from the SSOM2. The results implied that the soil collected from the different provinces were quite different in their characteristics although they were from within the same regions. However, the significant increase of the %CC in MSOM3 where the MSOMs were consecutively employed to classify the provincial sources of the soils within the same region could be observed. In fact, the %CC of MSOM3 is relatively high, in this case, because the most of the samples were firstly correctly identified their regional origins by MSOM1 (97.52 %CC). Otherwise, the correct classification rate would be consequently lower in MSOM3. The %PA indicated how often a sample was correctly classified. The averages of the %PAs for the training and the test sets using SSOM2 and MSOM2 were quite low. This implied poor predictive ability of the models. However, the improvement could be expected by tuning the number of learning iterations and the %MQE threshold. Besides acting as performance indices, %PA and %MS for each sample could be used to investigate whether there are any strong outliers in the dataset (samples having a high %MS but low %PA).

When confronted by CPN, SKN and *k*-NN, the MSOMs (MSOM4) has proven to be able to provide the better classification results. The %CC from CPN was lower than that from SKN implying that SKN more often obtained better classification results. This result agreed with the report from Melssen et al. [28] that the supervised SKN network in general performed better than the classical unidirectional CPN network. *k*-NN is a simple method and it is much less complicated than the Kohonen network approaches. The *k*-NN classification is based on the estimation of the distances between the training samples in an unknown sample. The class prediction of the unknown is depended on the majority votes for the class memberships of its nearest samples. The underperformance of *k*-NN could be that the numbers of samples in each class of the training sets were not approximately the same. Thus, the votes could be biased towards the class with the greater number of members.

The significant improvement in the classification performance of the MSOMs could be due to that each of the maps was executively trained based on the samples from a specific group or class membership and, therefore, the maps could exclusively learn and independently form itself to represent the characteristic variation within the class data. On the other hand, if the samples from all the class memberships were organized into only a single map, some parts of the map should have to express the dissimilarity for the samples from different classes. These map units would need to represent the difference between samples from different clusters or they were boundaries. For example, in a data visualization (e.g. U-matrix), these map units could be used to distinguish the difference between samples from the other classes with high dissimilarity. They could not function as interpolation units and, therefore, were not useful for the classification.