CHAPTER 2

Materials and methods

2.1 Soil samples and soil analysis

In this study, soil samples were taken on the harvest day from the surface layer (0 - 20 cm.) of the paddy fields specifically used for the cultivation of Thai jasmine rice variety Khao Dawk Mali 105 or *Oryza sativa* L. cv. Khao Dawk Mali 105 during November-December of 2012 and 2013, from the north and northeast of Thailand totally 877 samples. All soil properties were analyzed at Ubon Ratchathani Rice Research Center, Ubon Ratchathani, Thailand. In this work, pipette method which is popular for analyzing soil textures such as percentage of sand, silt and clay contents, Bray II method which is used for determining extractable P and Walkley-Black method were applied for organic matter (OM) measurement [14]. The analyzed soil properties could be seen below.

2.1.1 Physical properties

Soil physical properties used in this study was soil textures, normally including sand (more than 0.05 mm.), silt (0.05 – 0.002mm.) and clay (less than 0.002 mm.). The contents of each particle type were analyzed using the pipette method, probably the most commonly method used for particle size determination by weight of each particle size class of soil samples.

The soil sample was sieved to separate sand fractions from the silt and clay and calculated as %Sand of weight by weight after removing organic partitions. The rest sample which are including silt and clay was analyzed based on suspension and dispersing of soil particles in solution to measure density of the suspension at a specific depth after the critical particle-size fraction has settled to that depth [8]. %Sand, %Silt and %Clay of all soil samples was shown in Table 2.1.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Table 2.1 Averages and standard deviation of %Sand, %Silt and %Clay of soil samples

Provinces (No. of samples)	%Sand	%Silt	%Clay
1. Ubon Ratchathani (81)	3.51 ± 4.41	94.56 ± 7.35	11.93 ± 5.13
2. Amnat Charoen (58)	4.74 ± 2.91	83.68 ± 8.39	11.58 ± 6.54
3. Yasothon (38)	5.18 ± 2.03	76.63 ± 10.91	18.19 ± 9.74
4. Surin (18)	6.81 ± 3.31	70.39 ± 9.98	22.80 ± 7.32
5. Nakhon Ratchasima (30)	25.28 ± 11.63	53.61 ± 16.80	21.11 ± 7.29
6. Maha Sarakham (19)	12.45 ± 6.86	72.74 ± 12.06	14.81 ± 7.24
7. Sakon Nakhon (21)	14.98 ± 8.10	49.96 ± 16.79	35.06 ± 10.47
8. Nong Khai (15)	14.36 ± 8.11	62.09 ± 4.21	23.55 ± 8.27
9. Chiang Rai (28)	22.87 ± 7.77	58.07 ± 8.90	19.07 ± 5.39
10. Phayao (22)	25.65 ± 9.46	56.45 ± 9.63	17.90 ± 5.85

2.1.2 Chemical properties

There are nine extractable elements and two chemical properties were analyzed by suitable procedures. Soil sample and water in the ratio 1:1 was measured pH of samples [8]. Extractable P was determined using the Bray II method, the air-dried soils were extracted with 0.03 M NH_4F and 0.1 M HCl in a 1:20 soil: solution ratio and were shake for two hours. Then the solution was filtered through 0.45mm filter papers and was determined the phosphorus as Orto-phosphate [14]. Soil samples was extracted with ammonium acetate solution and NH_4EDTA solution and was used to determine the contents of extractable Ca, Mg, K and Na and extractable Fe, Cu, Mn and Zn using atomic absorption spectroscopy (AAS) which is the technique that excite atoms of elements in sample solutions to the excited stage and measure the absorbed wavelength [8]. The organic matter (OM) level in the soil was determined using the Walkley-Black method which soil organic C was oxidation of $Cr_2O_7^{2-}$ [14]. The excess dichromate is then titrated with a standard $FeSO_4$ solution and the amount of organic C oxidized is calculated from the amount of dichromate reduced. A colorimetric determination of the amount of Cr^{3+} produced can also be used to quantify organic C [8]. All chemical properties of samples in each province were shown in Table 2.2.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Table 2.2 Average and standard deviation of soil pH, extractable element concentrations and %organic matter (%OM) of samples from 10 provinces

Provinces No.	Soil pH	Avail.P (ppm)	Exch.K (ppm)	OM (%)
1	4.92 ± 0.56	16.37 ± 20.15	28.19 ± 48.56	0.86 ± 0.35
2	5.31 ± 0.67	6.36 ± 5.70	18.27 ± 16.27	0.55 ± 0.28
3	4.92 ± 0.31	14.44 ± 13.85	43.10 ± 49.81	1.01 ± 0.51
4	4.87 ± 0.40	10.48 ± 7.73	30.73 ± 23.46	0.75 ± 0.48
5	6.60 ± 0.80	13.99 ± 16.01	89.27 ± 44.15	1.08 ± 0.30
6	5.54 ± 1.32	12.99 ± 11.72	77.07 ± 50.59	0.98 ± 0.26
7	4.96 ± 0.42	2.70 ± 1.65	56.26 ± 36.75	1.25 ± 0.62
8	4.63 ± 0.28	3.34 ± 2.17	29.79 ± 16.94	1.22 ± 0.45
9	4.97 ± 0.57	21.45 ± 28.99	62.74 ± 41.44	1.98 ± 0.77
10	5.38 ± 0.59	10.41 ± 9.76	74.89 ± 41.36	1.85 ± 0.68

Table 2.2 (Continued)

Provinces No.	Na (ppm)	Cu (ppm)	Fe (ppm)	Zn (ppm)
1	16.18 ± 18.89	0.32 ± 0.31	95.77 ± 49.94	0.98 ± 1.12
2	48.31 ± 80.74	0.28 ± 0.21	72.02 ± 53.84	8.01 ± 33.1
3	58.56 ± 50.47	0.42 ± 0.19	115.9 ± 45.80	1.03± 1.01
4	24.69 ± 24.79	0.48 ± 0.15	77.32 ± 32.13	1.09 ± 1.56
5	302.1 ± 161.1	1.31 ± 0.63	28.93 ± 22.61	0.79 ± 1.08
6	276.0 ± 332.7	1.37 ± 0.77	90.50 ± 46.18	1.51 ± 2.48
7	56.72 ± 73.58	0.75 ± 0.38	105.3 ± 48.98	0.54 ± 0.32
8	21.04 ± 17.59	0.83 ± 0.67	103.7 ± 39.86	0.73 ± 0.72
9	21.71 ± 16.34	2.73 ± 1.97	185.5 ± 108.7	2.02 ± 1.72
10	45.35 ± 38.58	1.98 ± 1.02	103.9 ± 54.06	0.99 ± 0.78

Table 2.2 (Continued)

Provinces No.	Mg (ppm)	Mn (ppm)	Ca (ppm)
1	20.28 ± 44.82	10.45 ± 9.92	120.2 ± 107.6
2	16.16 ± 13.36	11.62 ± 13.60	142.0 ± 136.9
3	17.76 ± 11.37	16.47 ± 11.04	125.4 ± 99.25
4	28.07 ± 15.68	35.48 ± 16.62	299.6 ± 135.6
5	140.2 ± 75.61	32.65 ± 18.66	1661± 1487
6	153.6 ± 143.1	42.73 ± 32.24	459.8 ± 303.4
7	57.08 ± 30.34	37.77 ± 29.78	406.4 ± 199.9
8	30.29 ± 30.15	21.06 ± 20.57	151.4 ± 138.6
9	120.6 ± 82.90	63.81± 71.70	784.4 ± 636.1
10	148.8 ± 100.6	70.82 ± 72.44	1079 ± 722.5

After that analyzed data should be treated for ensuring that there is no outlier in the dataset and the interested information will be extracted without influence by

any noise or residual. This protocol is usually called data pretreatment or data preprocessing.

2.2 Data pretreatments

The results, were collected from the chemical experiment or experimental instrument such as concentration of chemical compounds, NIR spectra or GC-MS chromatograms, called data. Due to only the data could not answer the scientific questions such as which factor effect to the quality of aromatic rice or what is the correlation between salt strass and N concentration effect on yield and aromatic quality of fragrant rice. When the answers of scientific equation were called information such as concentration of N present direct effect to yield of rice or the interaction between N concentration and salt strass present negative effect on yield and aromatic quality of fragrant rice [26]. Therefore, the information was extracted from the data by data analysis methods. On the other hands, the data should be ensuring that the data was not covered by other interferences. For example, Figure 2.1 is a chromatogram that demonstrates retention times of compound No. 1-6. According to very low signal to noise ratio of compound No. 1, it is possible to say that this compound may be shielding by other compounds which present higher signal. As this reason, data should be prepared for the analysis process based on data pretreatment protocols like data screening or outlier detection and data preprocessing to confirm that the signal of compound No.1 is not only a noise of experimental system. If that signal is an actually signal of important compound, it should not be shielded by other signals. Otherwise, if that signal is only an outlier, is should be ignored.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyri iversity
All v e d

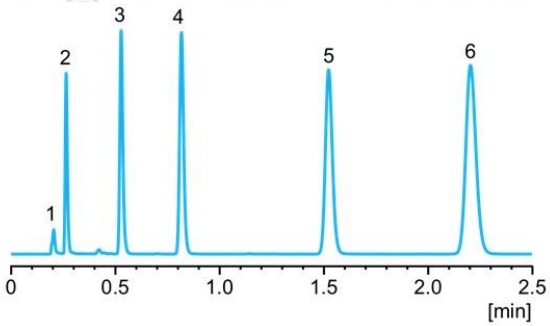


Figure 2.1 A chromatogram demonstrates retention times of compound No. 1-6

2.2.1 Data screening or outlier detection

The aim of data screening is to remove unacceptable event unusual or incorrect data such as a sample with pH value 27 or other nonsystematic mistakes in terms of data collection. The typical way for outlier detection usually based on statistic methods like chi-square test and z-score [12]. In this work, outlier detection based on majority vote such as percentage predictive ability (%PA) [16] when 2/3 of data from each class membership were randomly selected and used as training samples while the rest of the samples were used as test samples [27]. This algorithm was repeated for several times, called iterations. The %PA is a percentage of times that a sample is correctly classified and can be calculate as follow:

$$\%\mathrm{PA} = \frac{\mathrm{F}_{corrected}}{\mathrm{F}_{picked}} \mathrm{x}100 \qquad (1)$$

When $\mathrm{F}_{corrected}$ is number of time that a sample was correctly classified and F_{picked} is number of time that the sample was picked as a test sample. For example, using 50 iterations, if a sample is picked 30 times to be used as a test sample and from these 30 iterations if there are 21 times that this sample is correctly classified. This means that %PA for this sample is 70%. Therefore, it is possible to define that samples with too low %PA value should be incorrect or unusual data. In this work, sample with %PA lower than 30% were assigned as outliers due to very low stability of samples that are classified correctly.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

2.2.2 Data preprocessing

After outlier detection, the next step of data pretreatment effect to the extracted information significantly cause this process could prevent interested data from interruption of noise or interferences. For example, an interested data is the component No.1, in Figure 2.1. The signal may be shielding by other signals. As the result, the data should be reduced or remove some effects of other components.

In this work, the data were standardized by dividing each factor by its standard deviation after mean centering using the following equation:

$$s_{ij} = \frac{(x_{ij} - \bar{x}_j)}{\sqrt{\sum_{i=1}^{I}(x_{ij} - \bar{x}_j)^2 / (I-1)}} \quad (2)$$

where $\bar{x}_j = \sum_{i=1}^{I} x_{ij} / I$ is the mean for variable *j* calculated over all *I* samples. After standardization, all the variables were adjusted on the same scale as the following figure. This was to ensure that all variables had equal influence on the model [12].

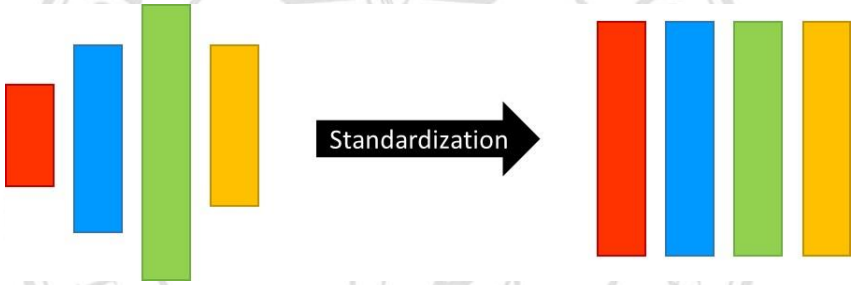


Figure 2.2 Standardization diagram

2.3 Chemometrics

Chemometrics is an application of mathematic, statistic and computer for analyzing complicated multivariate chemical data [12]. This method achieves more advantage over univariate analysis that more than one parameter could be simultaneously used for providing information [13]. For example, the criteria for some nice rice. If you focus only percentage of amylose, you may get rice with very good texture but have no aromatic quality or take a long time for cooking. Therefore, it should be better if we concentrate on many properties like amylose content, aromatic quality and time for cooking. Moreover, the corrected data which is ensured that there is the least interference remain in the data, could be analyzed by suitable methods. The methods in chemometric areas could be categorized into five groups per the purposes of the analysis such as design of experiments, multivariate calibration, process monitoring, data exploratory or exploratory

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

data analysis (EDA) and classification [12]. Therefore, in this study will focus on the method known pattern recognition or classification.

Classification is among the most important techniques in chemometrics [28]. The established models based on known samples were applied to identify unknown samples into the class memberships that they should belong to. This group of techniques were often divided into two major groups; one class classification or class modeling and multiple class classification or classification, respectively. The difference between class modeling and classification could be seen in the classification results such as an unknown sample might be classified into none, only one or more than one classes using class modeling method whereas the unknown would be defined into only one group based on classification model. On the other world, established classification model is more rigid and also known as hard modeling when compare to class modeling's which can be called soft modelling [12]. Moreover, the way for visualization is also different as you can see in Figure 2.3. In the case of class modelling was shown in Figure 2.3 (a) each model of known samples was established separately so each model belongs to its own space. Therefore, each model in class modelling could not comparable. Whereas model based on classification method, shows in Figure 2.3 (b) that every classes are modeled into the same space, are comparable. Moreover, the boundary of soft models depends on confident limit so the fitness of the model is changeable. As the reason, the result of the both classification approaches might not be the same.

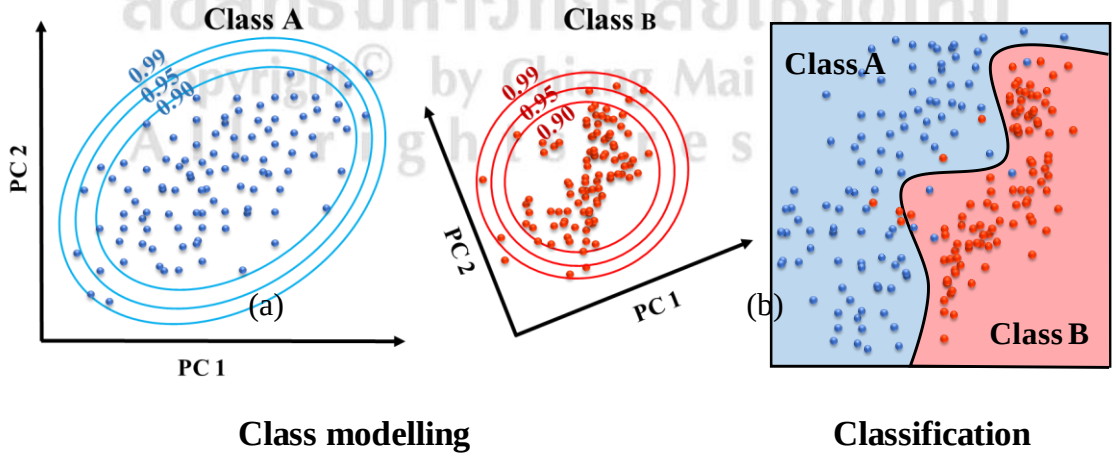


Figure 2.3 Visualization for class modelling method (a) and classification (b)

In the mean times, both methods are suitable for different data behavior. For example, if the only one answer of the classification results are required, the classification methodologies could be applied but when the flexible answers are acceptable, class modeling methods should be required. Therefore, the most suitable method should be considered. In this work, principal component analysis (PCA) and self-organizing maps (SOM) algorithm would be focused.

2.3.1 Principal component analysis (PCA)

Principal component analysis (PCA) is among the most common EDA methods [12]. PCA defines some new data dimension called principal components (PCs) which could be used to present the variation of the data where as much as possible the important information is retained. By investigating the first few PCs, it is possible to determine the relationship among the samples since the samples sharing the similar characteristics would be projected nearby each other.

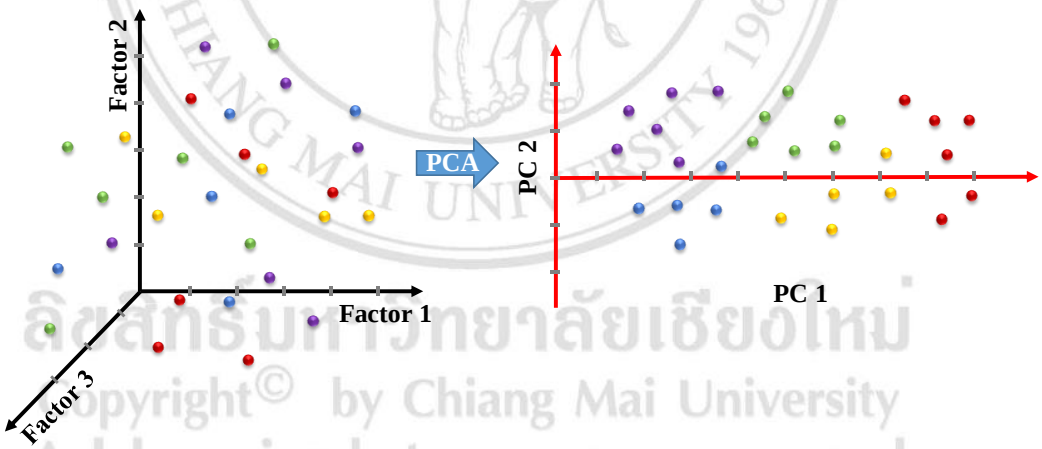


Figure 2.4 Example of principal component analysis (PCA)

As shows in Figure 2.4, there are five group of samples with three measurable factors. The data was presented in three dimensional axes which seems very complicate. When the data was analyzed using PCA and demonstrated in the first two principal components (PCs). Each group of samples can be observed easily in the plot of PCA result. Owing to the main variations of the data were extracted

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

so most of the data variation could be presented by only first few PCs. In addition to some of residue would be removed as the following diagram.

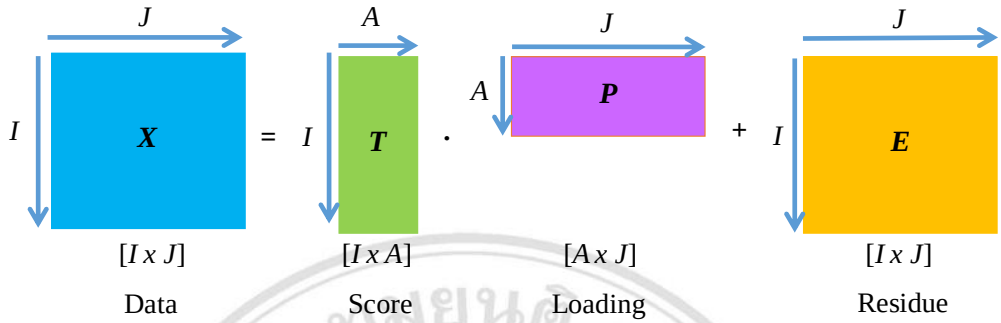


Figure 2.5 PCA diagram

According to Figure 2.5, ***X*** is a data matrix with *I* samples and *J* parameters. ***T*** is a matrix that contains score values of all *I* samples in PC 1 to PC A^{th}. The score value could be used to distinguish character of each sample. The samples with similar characteristics would have similar value of PCA score. ***P*** is a matrix of loading with *A* rows and *J* columns which presents behavior of each parameter, could be seen in loading plot. Moreover, the rest of non-systematic data would be leaved as the matrix of residue (***E***) with the same number of row and column as ***X***. The residue, also called noise or non-systematic variations, would not be included in systematic data which are score (***T***) and loading (***P***). Therefore, PCA result should be better than the raw data [12].

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

PCA algorithm is that, firstly, define a vector with maximum sum of square as an initial score ($^{initial}\hat{\boldsymbol{t}}$). Then, calculate loading vector from the following equation.

$$^{unnorm}\hat{\boldsymbol{p}} = \frac{^{initial}\hat{\boldsymbol{t}}'.\boldsymbol{X}}{\sum \hat{\boldsymbol{t}}^2} \qquad (3)$$

$$\hat{\boldsymbol{p}} = \frac{^{unnorm}\hat{\boldsymbol{p}}}{\sqrt{\sum {}^{unnorm}\hat{\boldsymbol{p}}^2}} \qquad (4)$$

After get the normalized loading, it is possible to calculate new scores as follow.

$$^{new}\hat{\boldsymbol{t}} = \boldsymbol{X}.\hat{\boldsymbol{p}}' \quad (5)$$

The convergence of the $^{new}\hat{\boldsymbol{t}}$ and $^{initial}\hat{\boldsymbol{t}}$ should be confirmed based sum of square difference between both score ($\sum\left(^{initial}\hat{\boldsymbol{t}} - ^{new}\hat{\boldsymbol{t}}\right)^2$). If it is small the score could be used for calculating loading (***p***) otherwise replace the $^{initial}\hat{\boldsymbol{t}}$ by the value of $^{new}\hat{\boldsymbol{t}}$ and $^{initial}\hat{\boldsymbol{t}}$ fine the using the equation (3) - (5). After fining the suitable sore (***t***) and loading (***p***) the residual can be calculated as follow.

$$^{resid}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{t}.\boldsymbol{p} \quad (6)$$

The further PCs could be calculate using $^{resid}\boldsymbol{X}$ of the previous PC [12].

2.3.2 Self-organizing map (SOM) or Kohonen network

A self-organizing map (SOM) or Kohonen network is among the most well-known artificial neural networks (ANN) in chemometric area [13]. This learning network generates low-dimensional data visualization map which is usually a two-dimensional for expressing relationship between samples in the original high-dimensional space [17]. SOM is an adaptive learning model so the map can be constructed without assuming any mathematical functions or in another word it is a nonparametric method. Therefore, SOM will not expect the underlying data distribution [18]. At the first time, SOM was particularly developed for visualizing of complicate data which could be called EDA method [17]. However, SOMs, later, were adapted to be used for calibration and classification tasks [18-20]. In addition, SOMs have an advantage over the other nonlinear methods in that insight complex relationships in the dataset could be revealed using the composite of the map vectors allowing the possibility to examine the importance and behavior of the parameters during classification [23].

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

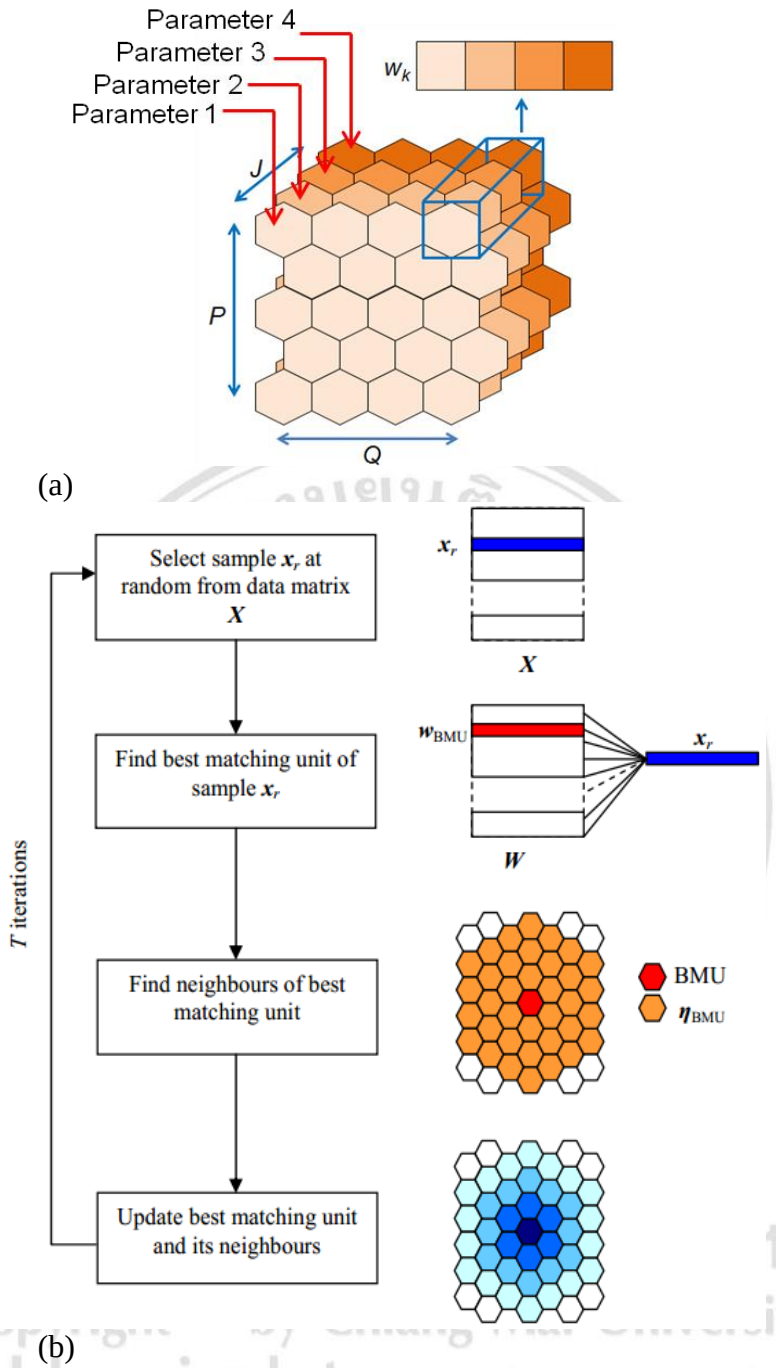


Figure 2.6 Initial SOM map with 4 parameters where *P* and *Q* = 4 (a) and training process (b) [17]

The algorithm of SOM begins with generating an initial map. The size of the map (*P* x *Q*) relates to number of the map units and the number of map layers is equal to the number of parameters (*J*) [13], shown in Figure 2.6 (a) Weight vector ($\boldsymbol{w}_k$) is an array of value in every layer, named component plane, of a map unit. Before training process, an algorithm step for map learning, the random values were

generated in each unit in the initial map. The values in each layer could be within range of each parameter from the training samples.

The training process was started by random a sample ($\boldsymbol{x}_r$) from training set ($\boldsymbol{X}$). In this study, seed random was fixed. The distances between the random sample and each of map units were calculated based on the Euclidean distance [17].

$$\boldsymbol{D}_{(\boldsymbol{x}_r,\boldsymbol{w}_k)} = \sqrt{\sum_{j=1}^{J}(\boldsymbol{x}_{rj} - \boldsymbol{w}_{kj})^2} \qquad (7)$$

When $\boldsymbol{w}_{kj}$ is a value of the parameter *j* in the unit *k*. A map unit with minimum distance or dissimilarity was assigned as the best matching unit (BMU) of the random sample ($\boldsymbol{x}_r$) and surrounding units of the BMU were defined as neighborhood units. The number of the neighborhood units depends on the neighborhood width (σ) for iteration *t*, could calculate using following equation:

$$\sigma_t = \sigma_0 exp\left(-\frac{t\ln(\sigma_0)}{T}\right) \qquad (8)$$

Where *T* is the number of iterations and the initial value of σ (σ_0) was set as the smaller dimension of the initial map. According to equation (8), the σ_t slightly decrease in every more *t* iteration. The transformation was performed by adaptive learning of each of the unit or neural of the weight matrixes so that they became more similar to the target training samples ($\boldsymbol{x}_r$). The learning rate (α) could be calculated using the equation below.

$$\alpha_t = \alpha_0 exp\left(-\frac{t\ln(\sigma_0)}{T}\right) \qquad (9)$$

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Although each of the map units processed their own discrete identity, the transformation was performed in the way the topological order of the training data could be preserved. This preservation allowed the relationship between the neighborhood map units to be existed. For the reason, groups of map units that processed the similar properties could be observed [23]. The map units that were close tended to share some similarity whereas the greater the distance, the greater the characteristic difference or they were from the different class. After the initial map is trained, several data visualization methods could be used such as unified distant matrix or u-matrix, supervised color shading, hit histogram and component planes [17].

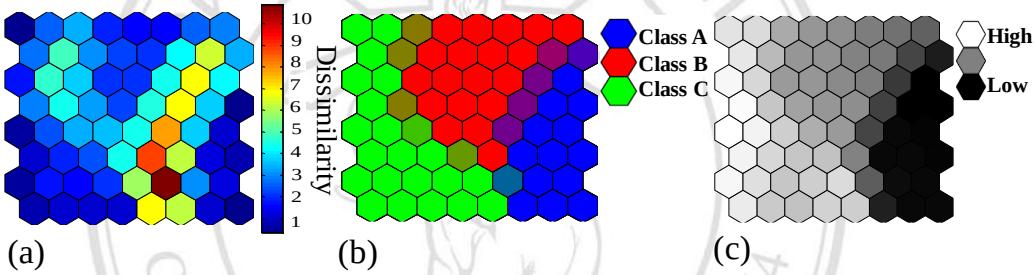


Figure 2.7 Map visualization of SOM (a) unified distant matrix or u-matrix (b) supervised color sheading and (c) component plane

Figure 2.7 demonstrates an example of three basic map visualization of SOM. The dissimilarity of map unit was presented in different colors by u-matrix, Figure 2.7 (a). High value of dissimilarity will be showed in red whereas blue represents for high similarity. The red tone map unites obvious as a boundary between groups of samples. Therefore, there are three groups of samples in Figure 2.7 (a). Another popular map visualization is supervised color sheading, Figure 2.7 (b), which each map unit will be colored based on class that each unit belong to. For example, the unit at right bottom corner has similar characteristic with samples in class A so that unit will be presented in blue color that belong to class A. Moreover, the pattern of parameter that effect on cluster or behavior of samples can be seen in layers of trained map which are called component plane. The units' colors represent the value of each parameter such as (c) dark torn means low value of the parameter which dominants for samples in class A in Figure 2.7 (b) and the pattern

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

of component accords to clustering of samples in Figure 2.7 (a) and (b). Therefore, the value of the parameter in component plane, Figure 2.7 (c), could be used as a criterial for classification.

Normally, SOMs could be categorized into two different variants due to how the models are trained: unsupervised and supervised SOMs. For the unsupervised SOMs, only the information about the predictors (or measurements such as soil parameters) is used, whereas, for the supervised SOMs, the information of the responses is also included during the training process. In general, both of the unsupervised and supervised SOMs could be employed for classification problems. [17, 21] Various approaches could be employed to utilize SOMs for classification problems [17, 21 and 28]. In this work, the classification rule was established based on the similarity between each unit of the trained map and the representative point of the training samples or the centroid point for each class. The similarities between each map unit and the centroids were evaluated using the Euclidean distance as follows:

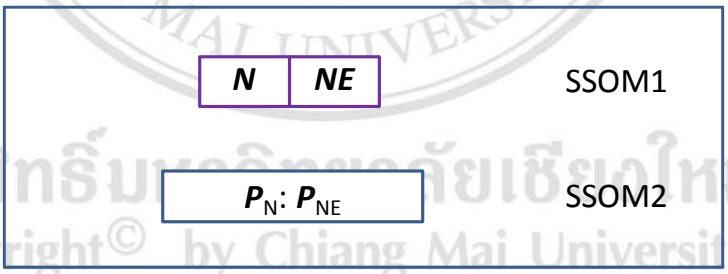
$$s_{(\mathbf{w}_k,\mathbf{c}_g)} = \left\| \mathbf{w}_k - \mathbf{c}_g \right\| \qquad (10)$$

where $s_{(\mathbf{w}_k,\mathbf{c}_g)}$ is dissimilarity between a centroid (c) of a class membership g and a map unit w_k. In this way, the class membership for a trained map unit could be labeled according to the class membership of the centroid having the smallest dissimilarity. After that, the class prediction of an unknown could be done by identifying the class membership of its corresponding BMU. A slightly similar classification algorithm has been reported by Cervera et al. [29] where the map units were labeled according to the majority of the training samples of a given class hitting on a particular map unit. However, the classification performance could be distorted if the numbers of samples in each class were unbalanced. In fact, there are other possible variants but this approach was used for simplicity and provided a demonstrable and simple protocol.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

A variety of extensions and modifications of SOMs was proposed for classification tasks such as counter propagation artificial neural network (CP-ANN) [30], supervised Kohonen network (SKN) [28], *XY*-fused network (XYF) [31] and bi-directional Kohonen network (BDK) [28]. These nonlinear modeling techniques have been successfully used in a wide range of applications, for example, chemistry [32, 33], agriculture [22, 30], process monitoring [16, 24] and medical and pharmaceutical analysis [16, 32]. Such applications were based on the mapping of samples on a single layer of SOM. Still, it is possible to represent the data using several maps at the same time.

According to this study, several pattern recognition methods were performed to compare the performance of classification. Both SSOM and MSOMs models were established by in-house scripts. Two different criterial of classification were used. First, a single self-organizing maps (SSOM) model was constructed to classify whether the samples were from the north (N) or northeast (NE) region of Thailand, will be called SSOM1. Second, it was also possible to train a single map to classify the soil samples collected from the 10 different provinces which is known as SSOM2 as the following diagram.



ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Figure 2.8 Overview of the SSOM for the soil classification

2.3.3 Multiple self-organizing maps (MSOMs)

MSOMs are the extension of a conventional SOM where some extra maps were optionally included in the model to create additional layers for representing different features of data [24]. There are several ways to identify the number of

the additional layers. For example, an extra layer of map was included if there were too many samples mapped on a unit or region of the parent map. In this case, the additional layer of the map could hierarchically process specific characteristics in the data and sometime called hierarchical self-organizing map (HSOM) [29]. Alternatively, the number of layers in MSOMs could be defined according to the class information of data [24]. Although the initial maps will be trained by an unsupervised algorithm, it was possible to extract the classification rules from each of the map units based on the similarity to an information of the modeling samples. The classification rule, as descripted in 2.2.3, was established based on the similarities between each unit of the trained map and the modeling samples. The similarities between each map unit and the centroids of the samples for each class were evaluated using the Euclidean distance [25]. In this way, each of the map units could be used to represent the class membership of which it had the smallest dissimilarity to.

The prediction of an unknown sample could be done by identifying the class membership of its belonging BMU. The prediction of this supervised SOM was similar to the idea presented in a report [29], however, in this work, all of the input parameters were used to establish the classification criteria instead of using the potential input parameters observed on the SOM component planes. In fact, there are other possible variants. Recently, several researches have employed the MSOMs for classification problems. For example, the application of multiple self-organizing maps (MSOMs) was present by Sim and Sági-Kiss for simultaneous classification and prediction the aged and group of apple samples based on volatile profiles [24]. However, it was not clear whether the extension the extra layers improved the predictive performance of the models. In addition, the MSOMs introduced by Sim were also constructed in the supervised mode.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Therefore, the difference between traditional SOM or SSOM and MSOMs is the number of total map, only one for SSOM but the number of map for MSOMs would equal to the number of clusters in the dataset [24, 25] so MSOMs should be classified as supervised method. For example, Figure 2.9 (a), there are two groups of samples in this dataset as you can see in the supervised color sheading

of SSOM and pattern of parameter j^{th}, Figure 2.9 (b), effect on the position of samples in the shading map. The samples with low value of parameter j^{th}, dark color in component plane, are belong to class A but the behavior of this parameter in class A cannot be seen. On the other hand, the pattern of parameter j^{th} in class A and B can be observed easily in component plane of MSOMs, Figure 2.9 (c). As the result, the advantage of MSOMs was the behavior of parameters in each class of the data could be seen whereas the different between patterns of the parameters among classes of samples could be presented in SSOM.

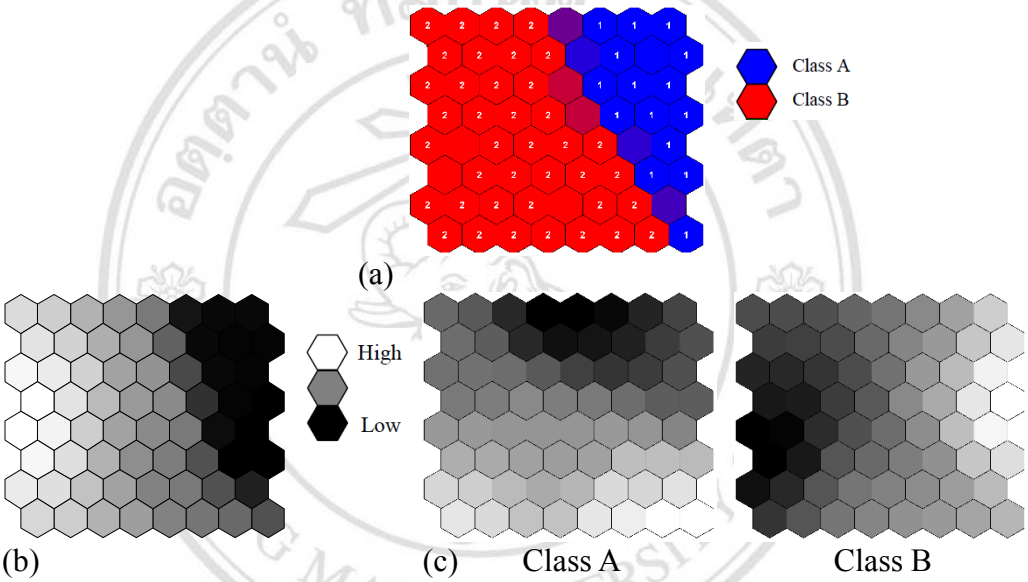


Figure 2.9 Supervised color shading of SSOM (a), component plane of parameter j^{th} based on SSOM (b) and MSOMs (c)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Moreover in this work, multiple self-organizing maps (MSOMs) could be applied with four different algorithms based on two different criterial as same as SSOM. Firstly, two layers of MSOMs were constructed to classify whether the samples were from the north or northeast of Thailand which is called MSOM1. MSOM1 model would be a comparative study of SSOM1 because both algorithms followed the same classification criteria. Secondly, MSOM2, two maps of collected areas which were classified based on region the north (N) and northeast (NE), were used for categorization the provincial origin of each soil samples. On the other worlds, MSOM1 were applied to classify for the sampling provinces by the next layer of a SSOM.

Therefore, MSOM2 would be semi-MSOMs and this would be a comparative case to SSOM2. Thirdly, the samples were corrected by MSOM1 algorithm then only the correctly classified samples were predicted their sampling provincial areas. In particularly, MSOM3 is the algorithm which classified by multiple-multiple SOM based on region and provincial respectively. Lastly, the training samples from each of the provinces were separately used to construct the MSOMs. This classification model consisted of 10 different maps and was named as MSOM4. Schematic diagrams of the SSOM and MSOMs for classifications of the soil samples are presented in Figure 2.10.

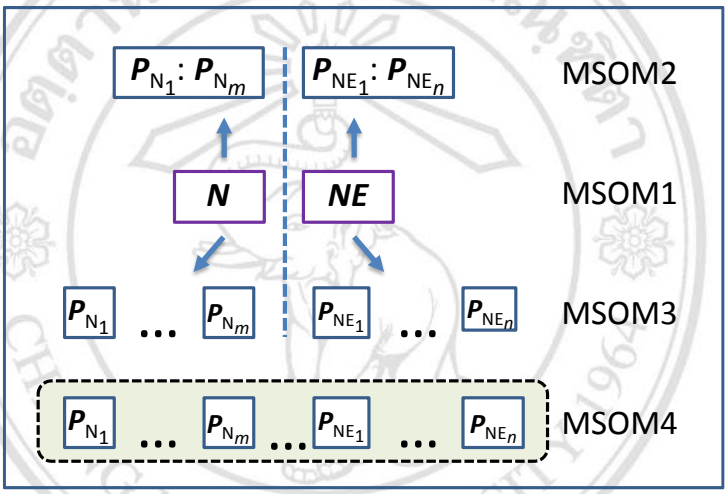


Figure 2.10 Overview of the MSOMs for the soil classification

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

2.3.4 Other classification methods

According to, some multivariate classification methods such as partial least squares discriminant analysis (PLS-DA), linear discriminant analysis (LDA), soft independent modelling of class analogy (SIMCA), *k*-nearest neighbours (*k*-NN), artificial neural network (ANN), counter propagation network (CPN) and radial basis function network (RBFN) were utilized to identify the geographic regions of soil samples [28, 30 and 31]. Although several research studies have employed multiple self-organizing maps for classification problems, it was not clear whether the extension of the extra layers actually improved the predictive performance of

the models. Therefore, in this study, the predictive results of multiple self-organizing maps were compared with some previously established Kohonen network methods such as counter propagation network (CPN) and supervised Kohonen network (SKN), as well as classical nonlinear classifiers such as *k*-nearest neighbours (*k*-NN) and quadratic discriminant analysis (QDA) and also some linear classifiers like linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA) and soft independent modelling of class analogy (SIMCA). Thus, it is possible to demonstrate the improvement in classification performance when MSOMs were applied to a classification task.

In this work, some neuron network based methods such as counter propagation network (CPN) and supervised Kohonen network (SKN) models were performed by algorithms which were presented in the study of Melssen et al. [28]. Several classical pattern recognition approaches were established by in-house script such as linear discriminant analysis (LDA), partial least squares-discriminant analysis (PLS-DA), quadratic discriminant analysis (QDA) and *k*-nearest neighbors (*k*-NN) when soft independent modelling of class analogy (SIMCA) were calculated by script functions, created by Milano Chemometrics and QSAR Research Group.

2.4 Model optimization

Traditionally, there are some factors that should be investigated before generate network model based on self-organizing map (SOM) algorithm such as number of map units or map size (height, *P*; width, *Q*) which should relate to number of samples and number of training iteration (*T*) which will affect to similarity between training sample and the trained map. As usual, many number of map units and iteration would be varied and chosen the best condition that the trained map can represent the most similar characteristic of training samples without over-fitting but this approach is time consuming. Therefore, in this study, the algorithm called growing self-organizing map (GSOM) was applied for fined the most suitable condition for data exploratory and classification models of both SSOM and MSOMs.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

2.4.1 Growing self-organizing map (GSOM)

The idea of GSOM was to increase the size of the map during the training process by inserting rows or columns of the map units. At the beginning, the starting map was small and, as the training progressed, it became larger. The insertions of additional units were done to reduce the map unit quantization error (QE) [12]. In this work, GSOM was trained in an unsupervised manner, and the insertion can be stopped when a set criterion was met. The steps of GSOM are described below:

(I) *Initial Setup and Training Process*: The initial map with a size of 2×2 units were generated. This newly created map was trained using an unsupervised SOM where the SOM parameters, such as the initial learning rate and initial neighborhood width, were set following the recommended methodology [17].

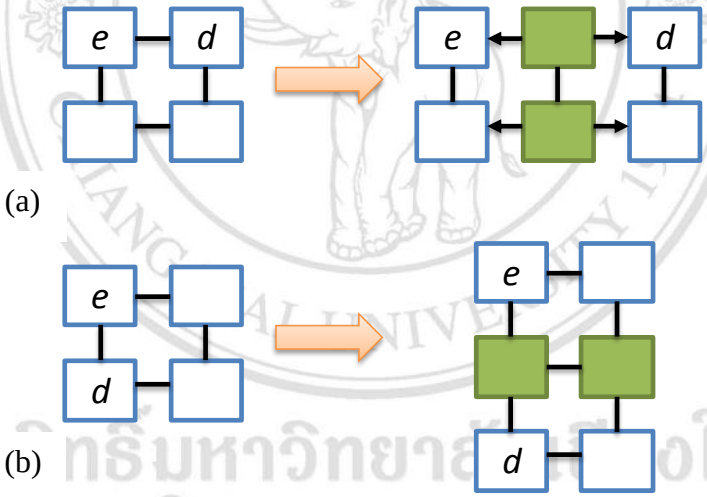


Figure 2.11 Insertion of (a) a column and (b) a row of map units (shaded green) in between the error unit *e* and the neighborhood unit *d*

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

(II) *Growth Process of the GSOM*: During the growth process, the map size was adapted by adding either a row or a column of units. The decision of where to insert the additional units was made by calculating the quantization error (QE) of each map unit *k* which can be calculated as:

$$\mathrm{QE}_k = \sum_{\mathbf{x}_i \in \mathbf{X}} \left\| \mathbf{w}_k - \mathbf{x}_i \right\| \qquad (11)$$

where $\boldsymbol{x}_i$ is a sample of the set of training set $\boldsymbol{X}$ which are mapped onto the map unit $\boldsymbol{w}_k$. If map units have high QE values, this suggests that there is dissimilarity or discontinuity on the map and implies that new units are needed to provide more space. To define location for the insertion of the map units, a unit with the highest QE was defined and denoted as *e*. This map unit *e* was compared to all its immediate neighboring units using the Euclidean distance, and the most dissimilar neighboring unit is determined and denoted as *d*. A row or column of units was inserted into to the current map between the map unit *e* and the most dissimilar neighbor *d* as shown in Figure 2.11. In this work, the inserted units were defined as the average of their corresponding neighbors. After adapting the map size, the learning rate and the neighborhood width were reset to their original values, and the map was again trained as per the previous training process.

(III) *Termination of the Growth Process*: As more units are inserted into the growing map, the map units are adapted and each of them represents a smaller number of training samples. Therefore, the mean quantization error (MQE) or the average of the QEs over the trained map should decrease. The MQE value used in this work can be calculated as follow:

$$\mathrm{MQE} = \frac{\sum_{k=1}^{K} \mathrm{QE}_k}{K} \tag{12}$$

where the training samples were mapped on a total number of *K* map units. The MQE value indicates whether the map units are well organized so that they can efficiently represent the training samples. Consequently, the MQE for the optimum size should be low. Thus, for the GSOM, the MQE value could be used as a stopping criterion for the growth of the map. If the calculated MQE was smaller than a defined threshold, the growing process could be stopped and this optimized map was then used for constructing the classification model. In this work, the MQE were expressed as a percentage of the MQE of the first trained map (%MQE) in step (I).

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

For a larger value of %MQE, some information in the training samples might be neglected and this could deteriorate the model performance, whereas a model with greater performance was expected when a smaller value of %MQE was set. However, the model was more likely to be overfitting and required a longer process time.

2.5 Model validation

Using SOMs, the models could be prone to over optimism because SOMs inductively learn to build models based on the training samples. As the training data, can very much characterize the models, the selection of the training samples is of importance. Several methods could be used to define the optimum selection of training sets [27]. Still, the choice of sample selections could be based on the intrinsic properties of the data provided. In practice, the selection of training samples could be different and therefore results in different optimal solutions. To evaluate the reliability of the SOM models, a bootstrap methodology was used [12]. After that, some statistic indices based on the majority vote including percentage predictive ability (%PA), percentage model stability (%MS) and percentage correctly classified (%CC) were computed to evaluate the model performance [16].

2.5.1 Bootstrap cross validation

Generally, cross validation is popular approach among the validation methods which the most well-known are leave one-out cross validation and bootstrap. Both approaches have different limitations. For example, leave one-out method is time consuming because the models will be established equal to number of samples and slightly over-fitting when deal with huge number of samples due to there is only one sample will be act as test sample and other rest are training samples. Nevertheless, limitation by the number of samples when duel with small data set. On the other hand, bootstrap is a bit under-fitting when applied with small number of samples since some parts of samples will be used for creating models, sometime the number of samples is too small. Therefore, the most suitable validation method is number of samples and the methods that were applied.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

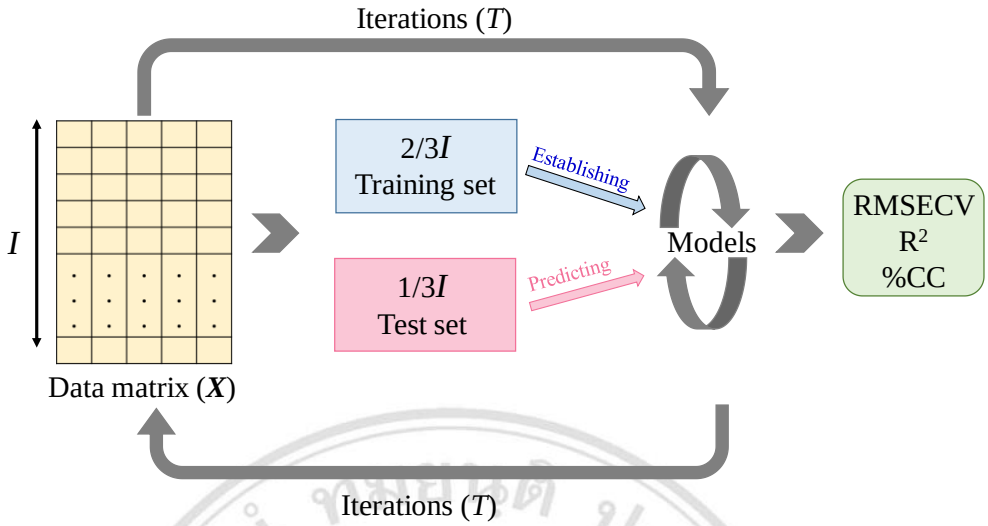


Figure 2.12 Diagram of bootstrap cross validation

Bootstrap cross validation, several predictive models were formed in an iterative fashion. During the iteration, each sample was used in either the training or the test sets. Generally, 66% or 2/3 of the all samples from each class membership were randomly selected and used as training samples while the rest of the samples were used as test samples. This algorithm was repeated for several times. The performance of established models will be presented in terms of statistic values such as root mean square error of cross validation (RMSECV), correlation coefficient (R^2) or percentage of correctly classified (%CC) as shows in Figure 2.12. In this study, the statistic indexes based on majority vote such as Percentage of predictive ability (%PA), model stability (%MS) and correctly classified (%CC) were applied to present classification performance along with Bootstrap cross validation.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

2.5.1.1 Percentage of predictive ability (%PA)

The percentage of predictive ability (%PA) is a percentage of times that a sample is correctly classified. The way to calculate %PA was explained in 2.2.1 equation (1). For example, using 100 iterations, if a sample is picked 60 times to be used as a test sample and from these 60 iterations if there are 45 times that this sample is correctly classified. This means that %PA for this sample is 75%. The high percentage of PA represent for high

frequency of each samples that was classified into the correct group which each sample should belong to. As the reason, the samples with very low %PA, high frequency of miss-classification, might be assigned as outlier so this algorithm was applied for data screening or outlier detection.

2.5.1.2 Percentage of model stability (%MS)

After %PA of the classification model went out. It is possible to measure the stability of the classification model based the calculated %PA. On the other words, the percentage of model stability (%MS) represents to how stable of the classification model is. The model with low %MS means lots of samples were usually defined into different classes in each iteration so the model was not very stable. When %PA is available, the %MS of two class classification can be calculated as follows:

$$\%MS = 2(|\%PA - 50|) \quad (13)$$

2.5.1.3 Percentage of correctly classified (%CC)

Finally, the percentage of correctly classified (%CC) measures how often a sample is correctly classified and determines a voting result. If a sample is classified more frequently (in this work, %PA >50%), it is then determined as correctly classified. For example, if there are 40 samples and 30 have %PA greater than 50%, the %CC is 75% which mean there 75% of samples were classified correctly. In this research, the calculations of the performance indices, the *k*-nearest neighbours (*k*-NN), linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), quadratic discriminant analysis (QDA), soft independent modelling of class analogy (SIMCA), counter propagation network (CPN), supervised Kohonen network (SKN), SSOM and MSOMs were implemented using in-house Matlab scripts (Matlab V7.0, The Mathworks Inc, Natick). To compare the classification performance, classification

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright © by Chiang Mai University
All rights reserved

models based on every pattern recognition methods were established several times and all statistic indexes, %PA, %MS and %CC, were calculated for representing the suitability of each classification algorithm for classifying group of soil samples, the dataset used in this study.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved