CHAPTER 1

Introduction

1.1 Research background

Soil could process different characteristics if they are from different geographical areas. The differences could be due to types of vegetable covers. Other natural factors such as their parent rock materials and variations in the climates could also contribute to the unique of soil properties [1]. In Thailand, soils in the northeastern region are usually sandy with substantial salt deposits whereas darker clay soils are expected to be found in the northern region [2]. Several methods and criteria could be adopted to establish soil classification system. A typical way to classify soil is based on soil physical and chemical properties such as particle size, color, pH, organic carbon content and element concentration. In fact, classification of soil is a part of soil quality assessment (SQA) that assigns types of soil samples depending on land used purposes like orchard, cropland or mining. On the other hands, classification of soil would classify based on the physical and chemical properties of the soil [3].

In 1900s, a simple way to classify types or characteristics of soil was to correct soil samples from field by burrowing or drilling and then analyzed soil morphology by comparing with the developed soil classification system [4] such as United States department of agriculture (USDA) soil taxonomy [5] and world reference based for soil resources (WRB) [6]. On the other hand, soil parameters such as soil organic matter, iron oxides, amorphous or short-range-order aluminosilicates and carbonates could be used to indicate the characteristic status of the soil [7, 8]. Recently, multivariate statistical methods such as cluster or factor analysis were introduced and applied in this area [9, 10]. This approach could provide more accurate results because many of the physical and chemical parameters could be incorporated into the modeling at the same time for the soil classification [11]. The group of techniques was later called chemometrics.

Chemometrics is a multivariate data analysis method [12]. It gains the advantage to univariate analysis in that more than one parameter could be simultaneously used for providing information [9-11]. The techniques in chemometrics could be categorized into groups according to the purposes of the analyses such as data exploratory, design of experiments, calibration, classification and process monitoring. Each group of the methods is suitable for different data characteristic and research objectives [12]. Typically, classification techniques could be classified roughly into two groups including linear and non-linear classification model. The techniques with linear mathematical equations would be defied as linearity methods such as partial least square discriminant analysis (PLS-DA) when the methods with quadratic or non-linear function would be called non-linearity methods such self-organizing map which is one of the most famous artificial neuron network (ANN) methods. Moreover, ANN methods could be used for both linear and non-linear data correlation [13].

In this work, the method called multiple self-organizing maps (MSOMs) in supervised mode will be used to define group of the soil samples from twenty provinces in the north and northeast of Thailand. The classification mission was to produce provincial maps for the rice growing soils based on the soil physical and chemical parameters such as soil textures, pH, some nutrients and soil organic matter. The soil samples were collected from paddy fields specifically used for one of the most famous Thai jasmine rice variety Khao Dawk Mali 105 (*Oryza sativa* L. cv. KDML105) or KDML105. The predictive performance of MSOMs were compared to the conventional SSOM and also compared with some previously established Kohonen network methods such as counter propagation network (CPN) and supervised Kohonen network (SKN) as well as some classical linear classifiers such as linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA) and soft independent modelling of class analogy (SIMCA) and some basic non-linear classifiers such as *k*-nearest neighbors (*k*-NN) and quadratic discriminant analysis (QDA).

1.2 Objective

The aim of this work was to develop an algorithm called multiple self-organizing maps (MSOMs) for classifying soil samples collected from different areas in the north and northeast regions of Thailand. The predictive performance of developed MSOMs was compared with other classification methods such as linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), soft independent modelling of class analogy (SIMCA), k-nearest neighbors (k-NN), quadratic discriminant analysis (QDA), counter propagation network (CPN) and supervised Kohonen network (SKN).

1.3 Study survey of this thesis

Soil is a mixture between organic substances and inorganic materials or minerals which support existences of many organisms. Soil acts as an intermediate of many substance cycles such as recycling system for nutrients, organic matters and gases [14]. Moreover, soil is among the environmental factors that influence the success of cultivation, a human activity that produces food, clothes and medicine for supporting human daily life [3]. For this reason, it is possible to claim that the abundance of yield from the cultivation depended on soil properties playing an important role as environmental factors. Therefore, soil ecosystem supports plenty number of living things on the earth.

Soil classification is a methodology for identifying group of soil based on chemical and physical properties [4]. The criteria of soil classification depend on purpose of the usage of land. For example, physical properties such as soil texture, soil viscosity and consistency could be used for defining group of soil based on suitability for pottery, advisability for civil engineering or risking area for landslip [4]. Normally, soil from different areas could possess different physical and chemical characteristics. Natural factors such as their parent rock materials and variations in the climates could also contribute to the unique of soil property. Determining the right soils for the right plants could ensure the optimal yield and quality of the products, therefore, soil classification could be regarded as an important task before implantation [14]. Thus, soil classification, one part of soil quality assessment (SQA), relate to soil analysis approaches of soil science [3]. Although, SQA usually assigns the types of soil in relation to land used purposes like orchard, cropland or mining, the classification of soil would classify based on the physical and chemical properties of the soil [7].

Typically soil components could different characteristics if soil samples came from different geographical areas [1]. In 1900s, a simple way to classify types or characteristics of soil was to correct soil samples from field by burrowing or drilling and then analyzed soil morphology by comparing with the developed soil classification system such as United States department of agriculture (USDA) soil taxonomy [5] and world reference based for soil resources (WRB) [6]. The types of vegetable covers and other natural factors such as their parent rock materials and variations in the climates could be contributed the characteristic of soil properties. For example, soils in the northeastern region of Thailand are usually sandy with substantial salt deposits whereas darker clay soils are expected to be found in the northern region. Therefore, some study focus only on physical properties such as particle size, color and consistency, the strength, held together, or the resistance for deformation and rupture of soil samples [1, 3]. The development using triangular soil classification chart which based only on soil particle sizes was developed in 1990s [3]. The chart was popularly used for engineering but it was not suitable for making decision for cultivation. An alternative protocol to classify soil called gradient analysis was then developed [9-11].

Nevertheless, some activities like agricultural production should not be based on only physical property but also chemical properties such as pH, organic carbon content and element concentrations [1]. Therefore, some studies suggested that the soil chemical parameters should be used to indicate the chemical characteristic of the soil samples [14]. For example, pH, soil organic matter, iron oxides, amorphous or short-range-order aluminosilicates, carbonates and element concentrations were used to indicate the characteristic status of the soil [7]. For these reasons, several methods and criteria could be adopted to establish soil classification models such as pipet method which is used for particle size analysis, chemical digestion for elements extraction and spectroscopic instruments for concentration measurements [8]. Later, soil classification mission needed to deal with multivariate and complicate data, and multivariate statistical methods for data analysis such as cluster or factor analysis were introduced and applied in this area [9-11]. This approach could provide more accurate results because all the physical and chemical parameters could be incorporated into the modeling at the same time for the soil classification [10]. Groups of multivariate techniques, later called chemometrics, played an important role for soil classification.

Chemometrics is an application of mathematics methods to chemical data. There are many purposes for the use of chemometrics such as data exploratory, design of experiments and pattern recognition [12]. The main aim of exploratory data analysis (EDA) is to extract the chemical pattern existing in complicate data. Principal component analysis (PCA) is among the most common EDA methods [15]. PCA defines some new data dimension called principal components (PCs) which could be used to present the variation of the data while 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 will be projected nearby each other. Design of experiment (DOE) is a chemometric approach which could be used to systematically design a set of experiments [12]. Based on the statistical design, DOE targets to establish the smallest number of experiments where the maximum information about the chemical system could be achieved. In addition, the effects of studied parameters could be revealed using the coefficient values. Classification is among the most important techniques in chemometrics [16]. In this study, pattern recognition or classification approach would be focused.

For classification propose, it is possible to classified the multivariate classification methods based on three criterions. Firstly, classification due to linear and non-linear criterions. For example, linear pattern recognition methods are the techniques that generated each class boundary of general group of samples based on linear mathematical equation such as linear discriminant analysis (LDA) which the boundary between classes were calculated using Euclidean distant. On the other, the boundary of non-linear methods was created using quadratic or other non-linear mathematics equations such as quadratic discriminant analysis (QDA). Moreover, artificial neuron network (ANN) is also non-linear methods because mathematical functions for adaptive neuron network node usually are non-linear equations [12]. Secondly, pattern recognition methods sometime were divide as the class modeling and classification models, one of the most common criterions, sometime called one-class classification and multi-class classification models.

One-class classification model would generate each class model separately such as D-statistic and Q-statistic when multi-class classification techniques would establish every classes in the same space such as LDA and QDA [16]. Finally, classification techniques could be classified into unsupervised and supervised methods as well. The responds of the measured data were used for model establishing in supervised techniques such as partial least square discriminant analysis (PLS-DA) [12] when only the measurement data or the predictor would be used in unsupervised approaches.

According to different algorithm of pattern recognition models, the most appropriate method for each dataset should be considered carefully. Most classical classification methods usually get along well with not too complicated dataset or dataset with linear correlation such as LDA, QDA and PLS-DA. Nevertheless, these basic techniques could not deal with some complicated dataset [12]. For example, data with non-linear relationships between parameters. In this case, adaptive techniques could be applied. One of the most popular techniques is ANN or Kohonen network based methods. This approaches normally would be performed with huge and complicated dataset where the normal distribution of the data was not required [13]. Therefore, in this work, one of Kohonen network based methods called self-organizing maps (SOM) was applied. It is possible to defied SOM into smaller group based on different algorithms using two criterions. First, unsupervised and supervised SOM which are defined above. Second, single self-organizing map (SSOM) and multiple self-organizing maps (MSOMs) which are interested in this study. The algorithm of SSOM and MSOMs almost the same accept the number of SOM maps. SSOM, as always, there is only one map when MSOMs have the number of SOM maps equal to number of classes or groups of samples in the dataset [17]. In addition to real world datasets usually represent non-linearity pattern, therefore SOMs, one of the most famous artificial neural network, were applied in many studied which were reported by Olawoyin et al. [18], Bação et al. [19] and Lloyd et al. [17]. It is not easy to point out that which method is the best and could be used for every type of the data thus there are many approaches were applied as you can see some of examples in Table 1.1. The table shows many classification methods were applied with different kinds of multivariate soil data for soil classification such as Kaniu and Angeyo [9] investigated 115 EDXRF spectra of six different types of soil samples. The score plot of the PCA could distinguish soil samples into four different clusters.

Table 1.1 Some chemometric techniques used for soil analysis

References	**Methods**	**Data**	**Analysis purposes**
Levine et al., 1996	Artificial neural network (ANN)	Physical structures	Classification
Rammadan et al., 2005	Partial lest square (PLS), back-propagation neural networks (BP-ANN) and principal component analysis (PCA)	Microbial community DNA	Calibration
Viscarra et al., 2006	Partial lest square (PLS)	Vis-NIR-IR spectra	Calibration
Caglar and Arman, 2006	Artificial neural network (ANN)	Physical structures	Calibration
He and Song, 2006	Principal component analysis (PCA) and partial lest square (PLS)	NIR spectra	Data exploratory and calibration
Sielaff et al., 2007	Linear discriminant analysis (LDA) and principal component analysis (PCA)	Physical and chemical properties	Data exploratory and classification
Mele and Crowley, 2008	Artificial neural network (ANN) and self-organizing maps (SOMs)	Biological, physical and chemical properties	Data exploratory
Singh et al., 2011	Principal component analysis (PCA)	Elemental profiles	Data exploratory
Kaniu et al., 2012	Partial lest square (PLS) and artificial neural network (ANN)	EDXRF spectra	Classification
Olawoyin et al., 2013	Artificial neural network (ANN) and self-organizing maps (SOMs)	Physical and chemical properties	Classification
Vasques et al., 2014	Principal component analysis (PCA) and multiple linear regression (MLR)	Vis-NIR spectra	Data exploratory and classification
Ackerson et al., 2015	Partial lest square (PLS)	Vis-NIR spectra	Classification
Kaniu et al., 2015	Principal component analysis (PCA), partial lest square (PLS) and artificial neural network (ANN)	EDXRF spectra	Data exploratory and classification

Owing to PCA and PLS are methods based on linearity equation thus the performance for calibration and classification might less than non-linearity methods when dealt with non-linearity correlations of the data according to the study of Kaniu et al., Singh et al. and Olawoyin et al. [9, 15 and 18]. Therefore, the method named artificial neural networks

(ANN) gave higher predictive ability than PLS in some cases of soil quality indicators (SQIs), presented non-linearity behaviors. Owing to, soil data, used in this work, present non-linear correlations and very complicated thus SOM should suitable for classification of this dataset [20].

Self-organizing maps (SOM) is a neuron network technique that could be used for data exploratory, calibration and classification proposes [21]. SOM usually presents the structure of training samples into a two-dimensional map although three-dimensional map also could be used [22]. In addition to SOM algorithm, there are some parameters that should be optimized for the most presentable results of SOM including number of map units and number of iterations for training process [23]. In this study, the approach known growing self-organizing map (GSOM) was introduced. GSOM is a method for optimized the most suitable size of SOM map based on stopping criterion such as mean quantization error (MQE) or percentage of mean quantization error (%MQE) [24]. The mean quantization error (MQE) represent the different between dataset and SOM map. The higher MQE value, the more different between data and SOM map. Therefore, if the MQE is very high, it means the SOM map should have more number of map units for presenting the characteristic of the dataset. Then, GSOM algorithm would increase number of map units by insertion a new row or column of SOM map. Thus, the MQE value should be decrease continuously when SOM map was expanding until the MQE of the map lower than the value that was set as stopping criterion [25].

In this work, an extension of SOM, the method called multiple self-organizing maps (MSOMs) with developed algorithms for data exploratory and classification of soil dataset. MSOMs in unsupervised mode will be used to define group of the soil samples from twenty provinces in the north and northeast of Thailand. The classification mission was to produce provincial maps for the rice growing soils based on the soil physical and chemical parameters such as soil textures, pH, some nutrients and soil organic matter. The soil samples were collected from paddy fields specifically used for one of the most famous Thai jasmine rice variety Khao Dawk Mali 105 (*Oryza sativa* L. cv. KDML105) or KDML105.

Using these non-linear methods, we expected that the behaviors of soil samples in each group could be observed easily. Moreover, the classification performance of the developed algorithm for soil classification should be better than the previous approaches. Therefore, the predictive performance of MSOMs, applier in this study, were compared to the conventional SSOM and also compared with some previously established Kohonen network methods such as counter propagation network (CPN) and supervised Kohonen network (SKN) as well as some classical linear classifiers such as linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA) and soft independent modelling of class analogy (SIMCA) and some basic non-linear classifiers such as *k*-nearest neighbors (*k*-NN) and quadratic discriminant analysis (QDA) respectively. Therefore, in this study will focus on supervised classification method which is MSOMs and basic data exploratory method like PCA.

1.4 Scope of study

This work developed the chemometric algorithm called multiple self-organizing maps (MSOMs) for clustering of soil samples from 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. The cultivation areas were located within the upper north and northeast of Thailand. A total of 877 soil samples from 20 provinces were used in this study. Soil textures, organic matter (OM) level, pH, Extractable P, Ca, Mg, K, Na, Fe, Cu, Mn and Zn were analyzed by Ubon Ratchathani Rice Research Center, Ubon Ratchathani, Thailand. Thus, the performance of MSOMs using unsupervised training algorithm for soil classification were compared with other classification methods such traditional SOM algorithm in both unsupervised and supervised mode, called SSOM and SKN respectively, linear methods such as LDA, PLS-DA, SIMCA, and non-liner methods such as *k*-NN, QDA and CPN.

1.5 Expecting benefit

The developed method could be used to identify the difference among the soil samples. It is possible to build a provincial map for the soil samples in different parts of Thailand. This information will be applied to investigate the parameters influencing the aromatic quality of Thai Jasmine rice and examine their behaviors.

In addition to the soil samples, it is possible to apply the developed methods to other research topics or the research objects. For example, classification of rice samples or other agriculture products such as oil palm and orange.