CONTENTS

Page

Acknowledgement iii

Abstract in Thai iv

Abstract in English v

List of Tables ix

List of Figures x

List of Abbreviations xii

List of Symbols xiv

CHAPTER 1 Introduction 1

1.1 Research background 1

1.2 Objective 3

1.3 Study survey of this thesis 3

1.4 Scope of study 9

1.5 Expecting benefit 10

CHAPTER 2 Materials and methods 11

2.1 Soil analysis and soil samples 11

2.1.1 Physical properties 11

2.1.2 Chemical properties 12

2.2 Data pretreatments 14

2.2.1 Data screening or outlier detection 15

2.2.2 Data preprocessing 15

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

2.3 Chemometrics 16

2.3.1 Principal component analysis (PCA) 18

2.3.2 Self-organizing map (SOM) or Kohonen network 20

2.3.3 Multiple self-organizing maps (MSOMs) 25

2.3.4 Other classification models 28

2.4 Model optimization 29

2.4.1 Growing self-organizing map (GSOM) 30

2.5 Model validation 32

2.5.1 Bootstrap cross validation 32

2.5.1.1 Percentage of predictive ability (%PA) 33

2.5.1.2 Percentage of model stability (%MS) 34

2.5.1.3 Percentage of correctly classified (%CC) 34

CHAPTER 3 Results and discussion 36

3.1 Data pretreatment 36

3.2 Model optimization for classification models 36

3.2.1 Single and multiple self-organizing maps (SSOM and MSOMs) 37

3.2.2 Other linear pattern recognition methods 39

3.2.3 Other non-linear pattern recognition methods 39

3.3 Chemometrics 40

3.3.1 Exploratory data analysis (EDA) 40

3.3.1.1 Principal component analysis (PCA) 41

3.3.1.2 Single self-organizing map (SSOM) 46

3.3.1.3 Multiple self-organizing maps (MSOMs) 49

3.3.2 Classification models 53

3.3.2.1 Classification based on regions (NE, R1 and N, R2) 53

3.3.2.2 Classification based on provinces (P1-10) 55

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

CHAPTER 4 Conclusion and suggestion 58

4.1 Conclusion 58

4.2 Suggestion 58

References 60

Appendix 64

Curriculum Vitae 65

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

LIST OF TABLES

Page

Table 1.1 Some chemometric techniques used for soil analysis 7

Table 2.1 Averages and standard deviation of %Sand, %Silt and %Clay of the soil samples 12

Table 2.2 Average and standard deviation of soil pH, extractable element concentrations and %organic matter (%OM) of samples from 10 provinces 13

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 54

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 55

CHIANG MAI UNIVERSITY 1964

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

LIST OF FIGURES

Page

Figure 2.1 A chromatogram demonstrates retention times of compound No. 1-6 14

Figure 2.2 Standardization diagram 16

Figure 2.3 Visualization for class modelling method (a) and classification (b) 17

Figure 2.4 Example of principal component analysis (PCA) 18

Figure 2.5 PCA diagram 19

Figure 2.6 Initial SOM map with 4 parameters where *P* and *Q* = 4 (a) and training process (b) 21

Figure 2.7 Map visualization of SOM (a) unified distant matrix or u-matrix (b) supervised color sheading and (c) component plane 23

Figure 2.8 Overview of the single SOM for the soil classification 25

Figure 2.9 Supervised color shading of single SOM (a), component plane of parameter j^{th} based on single SOM (b) and multiple SOMs (c) 27

Figure 2.10 Overview of the multiple SOMs for the soil classification 28

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* 30

Figure 2.12 Diagram of bootstrap cross validation 33

Figure 3.1 The %CCs of the training sets of the MSOM models using different %MQEs as thresholds for the GSOM 37

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 38

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Figure 3.3 Score plot of the first three PCs where the samples were labelled according to the regions they were collected, blue represent northeast region and red represent north region. 41

Figure 3.4 %Eigen values of the 704 (a) and 330 (b) soil samples which are contained by each principal component (PC) 42

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) 43

Figure 3.6 Loading plot of PCA model based on 330 soil samples in PC1 44

Figure 3.7 Loading plot of PCA model based on 330 soil samples in PC2 45

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 Fig. 3.5 47

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. 48

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. 50

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 51

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
BDK	Bi-Directional Kohonen Network
BMU	Best Matching Unit
%CC	Percentage of Correctly Classified
CPN	Counter Propagation Network
EDA	Exploratory Data Analysis
EDXRF	Energy Dispersive X-Ray Fluorescence
GC-MS	Gas Chromatography Mass Spectroscopy
GSOM	Growing Self-Organizing Map
HSOM	Hierarchical Self-Organizing Map
KDML 105	*Oriza sativa* L. cv. Khao Dawk Mali 105
k-NN	*k*-Nearest Neighbors
LDA	Linear Discriminant Analysis
%MS	Percentage of Model Stability
MQE	Mean Quantization Error
%MQE	Percentage of Mean Quantization Error
MSOMs	Multiple Self-Organizing Maps
N	North Region of Thailand
NE	Northeast Region of Thailand
NIR	Near Infrared
OM	Organic Matter
%OM	Percentage Organic Matter
%PA	Percentage Predictive Ability
PCA	Principal Component Analysis

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

PCs	Principal Components
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis
PT 1	*Oriza sativa* L. cv. Pathumthani 1
QDA	Quadratic Discriminant Analysis
QE	Quantization Error
RBFN	Radial Basis Function Network
RMSECV	Root Mean Square Error of Cross Validation
SKN	Supervised Kohonen Network
SIMCA	Soft Independent Modelling of Class Analogy
SOM	Self-Organizing map
SQA	Soil Quality Assessment
SQIs	Soil Quality Indicators
SSOM	Single Self-Organizing map
U-matrix	Unified distance matrix
USDA	United States Department of Agriculture
WRB	World Reference Based for Soil Resources
XYF	*XY*-Fused Network

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

LIST OF SYMBOLS

A	Number of principal components
$\boldsymbol{c}_g$	Centroid of a class membership g
$\boldsymbol{E}$	Residual
$\mathrm{F}_{corrected}$	Number of time that a sample was correctly classified
F_{picked}	Number of time that the sample was picked as a test sample
I	Number of samples
J	Number of parameters
$\boldsymbol{P}$	Loadings
P	Number of rows of initial map
Q	Number of columns of initial map
T	Iterations
$\boldsymbol{T}$	Scores
$\boldsymbol{s}_{ij}$	Standardized of sample i^{th} of parameter j^{th}
$\boldsymbol{s}_{(\mathbf{w}_k,\mathbf{c}_g)}$	Dissimilarity between a centroid of class g and a map unit $\boldsymbol{w}_k$
$\boldsymbol{X}$	Data matrix
$\boldsymbol{x}_{ij}$	Sample i^{th} of parameter j^{th}
$\bar{x}_j$	Mean for variable j calculated over all I samples
$\boldsymbol{x}_r$	A random sample
$\boldsymbol{W}$	Weight matrix
$\boldsymbol{w}_k$	Weight vector of sample k

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved