A Comparison of Classification Techniques for Monitoring and Mapping Land Cover and Land Use Changes in the Subtropical Region of Thai Nguyen, Vietnam

A thesis presented in partial fulfilment of the requirements for the degree of

Master of

Environmental Management

at Massey University, Palmerston North

New Zealand.

Tuyen Ha Van

2017

Abstract

Deriving land cover/land-use information from earth observation satellite data is one of the most common applications for environmental monitoring, evaluation and management. Many parametric and non-parametric classification algorithms have been developed and applied to such applications. This study looks at the classification accuracies of three algorithms for different spatial and spectral resolution data. The performance of Random Forest (RF) was compared to Maximum Likelihood (MLC) and Artificial Neural Network (ANN) algorithms for the separation of subtropical land cover/land-use categories using Sentinel-2 and Landsat 8 data. The overall, producers' and users' accuracies were derived from the confusion matrix, while local land use statistics were also collected to evaluate the accuracy of classified images. The accuracy assessment showed the RF algorithm regularly outperformed the MLC and ANN in both types of imagery data (>90%). This approach also exhibited potential in dealing with the challenge of separating similar man-made features such as urban/built-up and mining extraction classes. The ANN algorithm had the lowest accuracy among the three classification algorithms, while Landsat 8 imagery was most suitable for the classification of subtropical mixed and complex landscapes.

As the RF algorithm demonstrated a robustness and potential for mapping subtropical land cover/land-use, this study chose it to monitor and map temporal land cover/land-use changes in Thai Nguyen, Vietnam between 2000 and 2016. The results of this temporal monitoring revealed that there were substantial changes in land cover/land use over the course of 16 years. Agricultural and forest land decreased, while urban and mining extraction land expanded significantly, and water increased slightly. Changes in land cover/land-use are strongly associated with geographic locations. The conversion of agriculture and forest into urban/built-up and mining extraction land was detected largely in the Thai Nguyen central city and southern regions. In addition, further GIS analysis revealed that approximately 69.6% (100.2km^2) of new built-up areas had occurred within 2km of primary roads, and nearly 96% (137.6km^2) of new built-up expansion was detected within a 5-km buffer of the main roads. This study also demonstrates the potential of multi-temporal Landsat data and the combination of remote sensing, GIS and R programming to provide a timely, accurate and economical means to map and analyse temporal changes for long-term local land use development planning.

Keywords: Random forest; Land cover mapping; Remote Sensing; Vietnam

Acknowledgements

I would first like to express my gratitude and sincere thanks to my supervisors, **Mike Tuohy** and **Matthew Irwin** at Institute of Agriculture and Environment, Massey University for their guidance, valuable advice, encouragement, and endless patience through the entire process of this thesis research. They always allowed me to work on my own idea, but steered me on the right direction whenever I needed assistance.

My special thanks also go to all lecturers for their enthusiastic lectures and encouragement during my first year at Massey University, and to my classmates for cheering me up and giving constructive advice. Special thanks also devote to all Vietnamese friends and scholars at Massey University for their continuous encouragement and support.

I would also like to thank my friends back home in Vietnam who involved in fieldtrip survey for this thesis research, particularly Mr. **Tuan**, Mr. **Hung** and Mr. **Hoang**. This research would not be complete without their enthusiasm and assistance.

I would also love to take this opportunity to thank New Zealand government for sponsoring my home data collection, and for the scholarship to pursue the Master's degree in Environmental Management at Massey University. Special thanks also dedicate to **Jamie**, **Dave**, **Logan** and all staff at International Student Support Office, Massey University for making sure that all scholars like me could obtain the best academic achievements and enjoyable life experience while in New Zealand.

Finally, I must express my very profound gratitude to my family for providing me with continuous support and love throughout my years of study.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ROI: Region of Interest

DN: Digital Number

ETM+: Enhanced Thematic Mapper Plus

OLI/ TIR: Operational Land Imager/ Thermal Infrared

GIS: Geographic Information System

RS: Remote Sensing

MLC: Maximum Likelihood Classifier

ANN: Artificial Neural Network

RF: Random Forest

CART: Classification and Regression Tree

PCA: Principal Component Analysis

TOA: Top of Atmosphere

ENVI: Environment for Visualizing Images

NDVI: Normalized Differenced Vegetation Index

UTM: Universal Transverse Mercator

WGS84: World Geodetic System 84

NASA: National Aeronautics and Space Administration

USWGS: United States Geological Survey

VNIR: Visible/Near-infrared

SWIR: Shortwave Infrared

RGB: Red, Green and Blue

PCC: Post Classification Comparison

TNMT: Thai Nguyen Department of Natural Resource and Environment

1 Introduction

1.1 Background

Deriving land cover/land-use information from remotely sensed data has become a critical component for effective environmental monitoring, evaluation and management. Accurate and up-to-date land cover information is essential to understand and assess the consequences of environmental change. Remote sensing, with recent advances in the technology and an open-access data policy, together with increased temporal acquisition of data, can provide land cover/land-use information at a lower cost than traditional ground survey approaches (Szuster, Chen, & Borger, 2011). The analysis of these data can offer a better understanding of the subtropical landscape patterns and interactions between human activities and natural ecosystems (Rawat & Kumar, 2015). The expansion of urban and industrial areas on former cropland, grassland and forest could potentially cause significant consequences such as forest loss, biodiversity reduction and land degradation (Baker, Brazel, & Westerhoff, 2004; Zhao et al., 2006). Seto, Güneralp, and Hutyra (2012) estimated that three percent of the world biodiversity hotspots would be urbanized by 2030 due to rapid land cover/land-use change, and land cover/land-use changes are regarded as a main source of global warming emissions (Meyer & Turner, 1992). Given the scale and impact of land cover/land-use changes, the choice of classification algorithms is essential in accurately monitoring and assessing such dynamic changes for sustainable land use development planning.

Since the Landsat program was first launched in the 1970s, the derivation of land cover/land-use change information has been made possible by advances in computing technology and the development of software applications. Along with the rapid development of computer systems and machine learning algorithms, many parametric and non-parametric classification techniques (e.g., random forest, artificial neural network and maximum likelihood classifiers) have been developed and applied to extract land cover/land-use information and monitor temporal spatial changes. While parametric classification techniques are based on statistical assumption, non-parametric approaches are not dependent on such assumptions. The most widely used parametric techniques is the Maximum Likelihood Classification (MLC) algorithm, whereas newer alternative non-parametric classification algorithms such as Random Forest (RF) and Artificial Neural Network (ANN) are gaining popularity in the remote sensing community, particularly land cover/land-use studies (Huang, Davis, & Townshend, 2002). When producing land cover/land-use information from remotely sensed data , one of the challenging issues is the spectral mixture of different earth objects (Poursanidis, Chrysoulakis, & Mitraka, 2015).

Many studies have investigated the performance of different algorithms for land cover/land-use classification. Seto and Liu (2003), for instance, assessed the performance of the ANN algorithm with the MLC using Landsat TM imagery for urban change detection. Szuster et al. (2011) used a 15-m ASTER image to test the performance of SVM (support vector machine) against ANN and MLC algorithms in the coastal zones of Thailand, while Huang et al. (2002) used Landsat TM to compare the accuracy of SVM against the MLC and ANN algorithms. Pal and Mather (2003) tested the performance of decision tree methods against the MLC and ANN using Landsat ETM+ data, and this approach was also compared to discriminant analysis (DA) and support vector machine (SVM) using an airborne thematic mapper image (5-m). Most of these studies used Landsat to test the classification accuracies of algorithms. However, relatively little research has done to examine the performance of the RF algorithm against the MLC and ANN techniques for different spatial and spectral resolutions (e.g., Sentinel-2 and Landsat 8). Therefore, assessing the performance of the three classification algorithms is critically important in understanding the advantages and drawbacks for each technique in classifying multi-sensor data when deriving complex and mixed subtropical landscapes.

Currently, national and local land managers are frequently requested for reliable and up-to-date land cover/land-use information, relatively little research has conducted to monitor and map temporal land cover/land-use changes in subtropical environments (e.g., Vietnam). Although earth observation data for deriving land cover/land-use information has been demonstrated to be relatively efficient and accurate (Khiry & Csaplovics, 2007), the use of satellite data to monitor and map local land cover/land-use changes in subtropical regions was inadequate. As an example, Thai Nguyen, a northeast province of Vietnam, has been experiencing relatively substantial changes in land cover/land-use over the past two decades. The local government keeps practising traditional survey approach to map its land cover/land-use. Alternative monitoring and mapping these temporal changes using a reliable algorithm and remote sensing data are more efficient to provide timely and accurate information for sustainable local land use development planning.

The objectives of this work are; (1) to compare the performance of the RF algorithm against the MLC and ANN using Landsat 8 and Sentinel-2 data for subtropical land cover/land-use mapping, (2) to provide a recent perspective for land cover types and land cover changes that have taken place in the last 16 years in Thai Nguyen, Vietnam, (3) to integrate R programming and GIS with remote sensing data in land cover/land-use monitoring and mapping.

1.2 Land cover/ land-use in Thai Nguyen, Vietnam

After the adoption of a national economic reform (the Doi Moi) in 1986, Vietnam has substantially increased urban and industrial activities, and obtained impressive socio-economic achievements over the three decades. The annual average growth rate of the economy stabilized consistently at 7.5% between 1991 and 2005 (Vuong, 2014), while the country continued to expand its urban and industrial zones, and is considered one of the fastest urbanized countries in the southeast Asian region (T. McGee, 1995; T. G. McGee, 2008). The rapid growth of urbanization and industrialization has long been considered a sign of national and regional economic prosperity, but its tremendous changes in the economic system has brought a negative effect on the spatial structure and patterns of land cover/land use (Quang & Kammeier, 2002).

Vietnam has also witnessed a substantial change in land cover/land-use across the country over the past two decades, particularly in metropolitans and its suburbs. The transition of economic infrastructure from agriculture to industrial and urban services was constantly targeted by the Vietnamese government (Ministry of Planning and Investment, 2005). Massive rural landscapes are quickly converted and occupied by urban residential and industrial zones (Castrence, Nong, Tran, Young, & Fox, 2014). Ministry of Planning and Investment (2005) revealed that approximately 400 km^2 has been planned to allocate for the construction and development of industrial zones alone between 2005 and 2010, and this figure is expected to reach 800 km^2 by 2020. However, the actual amount of land mass was allocated to the construction and development of industrial zones alone by 2014 reached 810 km^2 (Phan Manh Cuong, 2015). Noticeably, these changes in land cover/land-use are unevenly distributed between geographic regions, and frequently occurred in neighbouring provinces of Hanoi capital city such as Thai Nguyen province.

Thai Nguyen province is located in the heart of southeast region of Vietnam, and it borders on Ha Noi city in the south and Bac Kan province in the north. Due to geographic terrain and climate factors, Thai Nguyen is largely covered by forest and agriculture in the north and south regions respectively. Developed areas are clustered in the Thai Nguyen central city and southern portions such as Pho Yen and Song Cong districts, while rural land characterizes the northern region and its surrounding areas. The local government identified six main land cover/land-use categories representing across the province, namely agriculture, non-agriculture, unused land, urban land, land for natural conservation and land for tourism development (Government, 2013). Among those land cover/land use types, agriculture and forest are the most prominent types of cover with an area of approximately 2830 km^2. Unused land accounted for a small area, and presents in mountainous areas due to timber extraction and gold mining.

In recent years, Thai Nguyen province has been experiencing a substantial urbanization and industrialization expansion. With the advantages of accessible transport and geography, Thai Nguyen has prioritized key economic sectors (e.g., industry and mining); therefore, its industrial and urban development improved significantly. However, the rapid development of industrial zones and urban infrastructure has resulted in the loss of a significant area of forest and cropland. As an example, the local government has allocated more than 6 km^2 of forest and agriculture for the construction and development of 20 industrial zones between 2002 and 2010 (Thai Nguyen People's Committee, 2010), and this figure increased approximately 41.2 km^2 in 2014 (Phan Manh Cuong, 2015).

The expansion of urban and industrial zones on rural landscapes is expected to continue in coming years. On 27th February 2015, the Vietnamese government issued the Decision No 260/QD-TTg in the approval of a comprehensive socio-economic development strategy for the Thai Nguyen province until 2020 with a vision for 2030. This comprehensive strategy was targeted to transform Thai Nguyen into a modernized, urbanized and industrialized province of the northeast region of Vietnam. With a primary focus on the industrial and urban development, the province has outlined and initialized the industrial zone development plan by 2020. Figure 1.1 shows the spatial distribution of industrial zones and major developments across the Thai Nguyen, and most of its industrial parks are located in the south and suburb areas of Thai Nguyen central city. This scheme is expected to allocate 638 km^2 (18% of the total area) for non-agricultural purposes by 2020, and would potentially increase the disappearance of forest and agriculture.

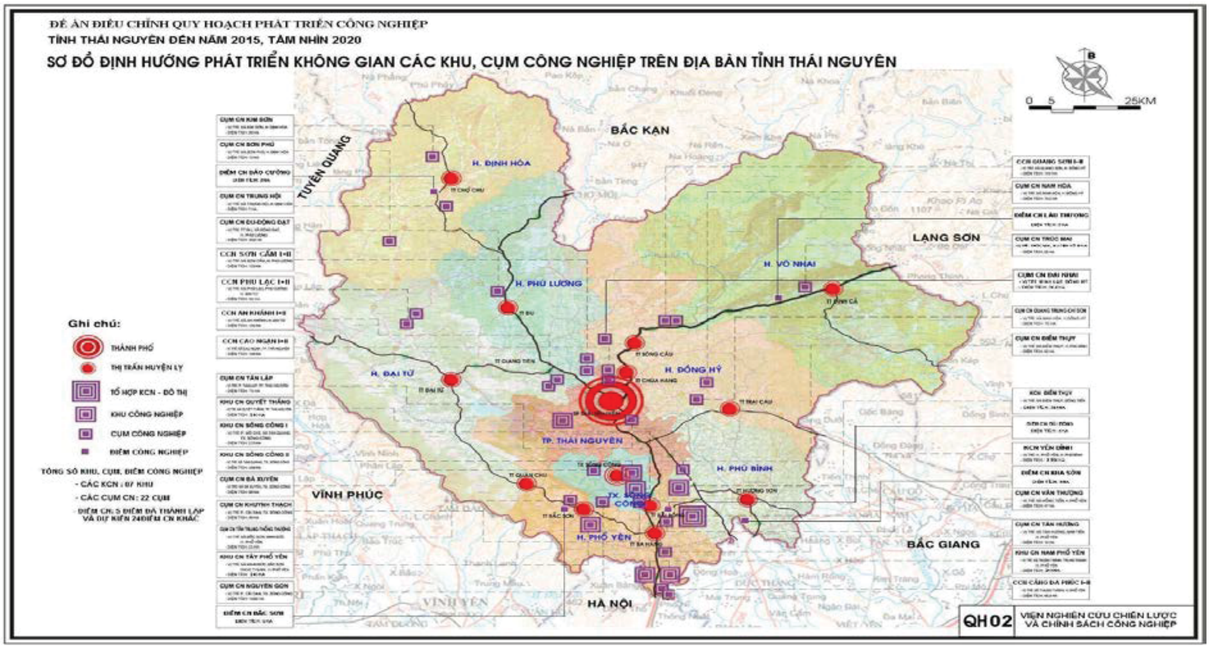


Figure 1.1 Spatial distribution of industrial zone development by 2020, Thai Nguyen

Source: Thai Nguyen's People Committee

A part from industrial developments, mining exploitation activities such as coal, iron and tungsten have been increasing in Thai Nguyen over the past decade. Most of these activities are taking place in forest or agriculture areas; and as a result, many fertile farming areas and natural forests were made way for mining extraction projects. Recently, the local government has granted 169 mining licenses for 79 organizations and individuals with a total area of around

55 km^2 (Vietnam Geology Society, 2014). The loss of farming and forest lands would potentially result in food security and environmental problems. Therefore, accurate mapping and monitoring of land cover/land-use information is a pivotal step in long-tern sustainable land use development.

While the local department of natural resource and environment is required by law to conduct regular ground surveys to monitor land cover/land-use change (Vietnam Government, 2013), little research has done to derive land cover/land-use information from earth observation satellite data over the entire province. It is obvious that ground surveys can provide accurate land cover/land-use information, but this approach is very costly, labour-intensive and time-consuming to produce geospatial data. Whereas there is an increasing availability of free-cost earth observation data and image processing software, this idea is clearly to produce massive updated and inexpensive land cover/land-use information over the same geographic region within a certain interval. Landsat and Sentinel-2 sensors, for example, continuously provide relatively high spatial and spectral images over the entire globe every 16 and 5 days respectively. This source of data will be valuable for local and national policy-makers and politicians in the consideration of its sustainable land use development planning.

1.3 Research problem

With the increasing availability of free and low-cost satellite data, the activities of the earth environment and its resources have become easier than ever before to monitoring and manage. Many environmental variables can be derived from satellite remote sensing data, of which land cover/land-use information is considered to be one of the most commonly derived variables. Over last few decades, there has been an increasing interest in machine learning algorithms for land remote sensing, along with the rapid development of computer systems. Many parametric and non-parametric classifiers have been developed and applied to land cover classification.

Although there is a need for a standard land cover classification technique, none of current classifiers has been internationally accepted (Anderson, 1976). This is partly because each algorithm has different capabilities to classify land cover surface, and each region usually represents various unique land cover complexes. Therefore, the comparison of classifiers plays an important role in the enhancement of the accuracies of land cover derived information, and minimizes time and costs. In this study, the focus is to explore the capability of the three land cover classification methods (maximum likelihood, artificial neural network and random forest) to classify subtropical land cover/land use categories using two different spatial and spectral Sentinel-2 and Landsat-8 data. The RF algorithm was then used to classify and monitor temporal land cover/land-use changes between 2000 and 2016 in Thai Nguyen using multi-temporal Landsat data due to its high accuracy and stability.

1.3.1 Research goals

This study aims to:

- Compare the performance of the RF against the MLC and ANN algorithms to derive land cover/land-use information using Sentinel-2 and Landsat 8 data;
- Provide recent perspective for subtropical land cover categories and its temporal changes in Thai Nguyen, Vietnam.

1.3.2 Research concerns

- ❖ Comparison of land cover classifiers

- What are advantages and disadvantages of the parametric and non-parametric classifiers for subtropical land cover detection?
- What are accuracies of each technique used? And which one yields the most accurate product?
- What are differences in overall accuracies between two different spatial and spectral satellite data?
- Which classifiers should local land managers choose in terms of overall accuracy and its accessibility?

❖ Land monitoring and mapping

- What were the trend and pattern of land cover change in Thai Nguyen over the last 16 years?
- How much land cover has been changed or converted? And what is the nature of its changes?
- What is the spatial distribution of its change in the study area?
- What are primary causes behind its changes?

1.4 The organization and structure of the research thesis

This research thesis is consisted of six chapters concerning various aspects of land cover/land-use, classification algorithms, accuracy assessment and temporal change detection. The first chapter introduces the research background, land cover/land-use status in Vietnam and Thai Nguyen in particular, and defines research goals and questions. Chapter 2 provides brief insights of image classification algorithms, land cover monitoring and mapping practices, role of remote sensing and GIS in land cover/land-use classification research, and change detection and accuracy assessment. Chapter 3 focuses on the description of study area and data collection, including satellite remote sensing and reference data. Chapter 4 deals with classification techniques used to classify the Sentinel-2 and Landsat data, accuracy assessment, and to monitor and map temporal land cover/land use changes. Chapter 5 discusses and presents the results of the research, and its findings. The final chapter summarises the main research findings, and recommendations for future studies.

2 LITERATURE REVIEW

2.1 Remote sensing and GIS

Remote sensing is the process of acquiring information about biophysical and biochemical properties of earth surface features without physical contact with it. The purpose of this acquisition process is to capture reflected or emitted energy from earth surface using a sensor, installed on a satellite platform or aircraft system.

Geographic Information System (GIS) is a computer-based tool for storing, managing, mapping and analysing various spatial data, particularly vector and raster data. Recently, the integration of remote sensing and GIS has proved useful in monitoring and managing the earth environment and natural resources as well as modelling of urban expansion.

2.1.1 Remote sensing

The term "remote sensing' originated in the 1960s by geographers at the US Office of Naval Research (Cracknell, 2007). Since then, there have been numerous definitions of remote sensing, but Campbell and Wynne (2011) stated that "Remote sensing is the practice of deriving information about the Earth's land and water surfaces using images acquired from an overhead perspective, using electronic magnetic radiation in one or more regions of the electronic-magnetic spectrum, reflected or emitted from the Earth's surface". Many earth observation programs have been developed and launched by various countries (e.g., USA, France, India and Japan) for monitoring and managing the earth environment. The Landsat program started with the launch of its first satellite in 1972, and has become the longest-running program for the acquisition of the earth observation satellite data. There has been a total of 8 satellites launched, of these 8 Landsat 6 failed to reach orbit in October 1993, while Landsat 7 ETM+ suffered from scan line corrector failure in 2003, and Landsat 8 is the current platform in orbit and expected to decommission in 2020 (NASA, 2017) the rest of the satellites reached the end of their operational life and were decommissioned.

With recent advancement in space technology and machine learning algorithms, some newer generations satellites with high spatial, spectral and temporal resolution have been developed and applied. Specifically, the Sentinel-2A satellite, a part of Copernicus program jointly developed by the European Space Agency and European Union, was first launched in June 2015. The Sentinel-2B was recently launched in March 2017, it provides 13 high spectral and spatial resolution bands (Figure 2.1), of which there are 10-m pixel size for bands 2, 3, 4 and 8, 20-m for bands 5, 6, 7, 8A, 11, and 12, 60-m for bands 1, 9, and 10, and with a 5-day revisit frequency respectively (Drusch et al., 2012; Spoto et al., 2012). The Sentinel-2 mission aims to provide systematic global acquisitions of high-spatial and spectral resolutions, and continuity for the current SPOT and Landsat satellite sensors with an expected design lifetime of 7 years. With a wealth of high-resolution images, the Sentinel-2 would extensively enhance our understanding of the earth surface, atmosphere and oceans, while it potentially provides critical information for local and national policy-makers and land managers to make wise decisions about their environment and resource management.

In recent years, commercial satellites have contributed greatly to the continuous provision of high-quality satellite data with the launch of IKONOS, QuickBird and OrbView-3 that have very high spatial resolutions ranging from 0.5-m to 5-m. These commercial sensors are very useful in urban planning and monitoring, emergency assistance, transportation, agriculture and forestry. However, it has associated limitations for land cover classification applications such as low coverage area and inaccessible affordability, and often records data in the visible and near-infrared electromagnetic region. GeoEye-1, for example, covers a narrow swath of 15.2 km, and is unavailable to the public at zero cost.

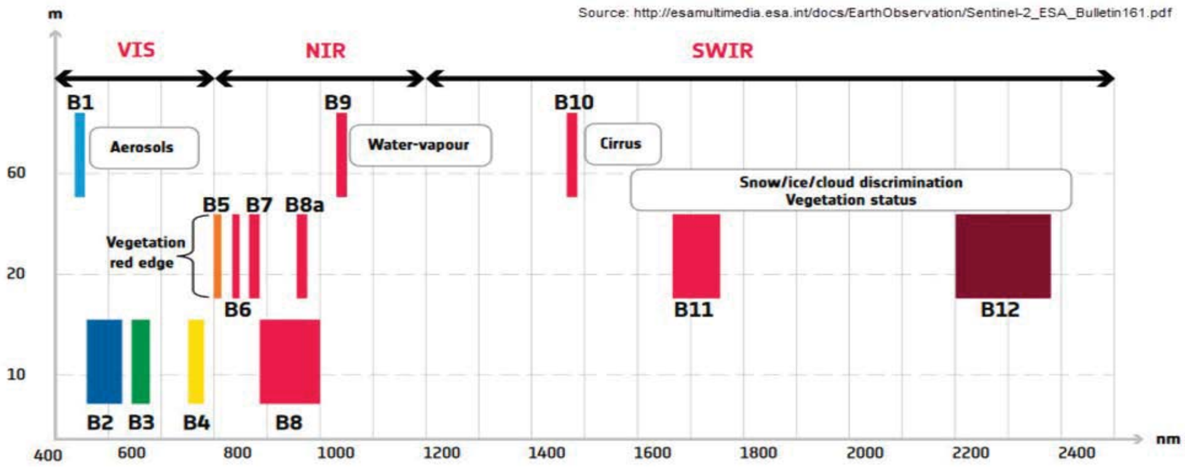


Figure 2.1 Sentinel-2 spectral and spatial resolution
Source: European Space Agency

Thanks to satellite technology and GIS development, many natural and human-made processes have been effectively monitored and detected. Dewan and Yamaguchi (2009a) used Landsat and socio-economic data to quantify land cover changes in Greater Dhaka, Bangladesh from 1975 to 2003. Their study revealed that the significant expansion of urban land was taking place in rural areas and water bodies, and largely correlated with elevation, population and economic growth. Similarly, Pham and Yamaguchi (2006) monitored and mapped temporal land cover/land-use changes in Ha Noi city, Vietnam between 1975 and 2003. The results of this study showed that the expansion of urban areas occurred largely along transport systems and transitioned to the southern and western regions. In addition, remote sensing is also used to monitor and provide rapid and accurate information about the nitrogen status in the agriculture practice to improve the growth of crops (Bausch & Duke, 1996). Many other applications of remote sensing in a wide range of disciplines have been documented by geographers, environmental managers, geologists, ecologists and cartographers (Govender, Chetty, & Bulcock, 2007; Green, Mumby, Edwards, & Clark, 1996; Sanyal & Lu, 2004).

2.1.2 Geographic Information System

Geographic Information System (GIS) is a system designed to capture, manipulate, analyse, manage and display spatial or geographic data. In a broader context, GIS is part of geographic information science and technology, and has recently become one of the fastest growing fields of study and application (David Dibiase, James L.Sloan, Ryan Baxter, Wesley Stroh, & King, 2017). Many government agencies, commercial organizations, private companies and universities have produced a wide range of spatial and remotely sensed data (e.g., elevation) for use in GIS.

GIS enables users to perform from basic tasks such as view, query and representation of spatial data to advanced statistical modelling such as modelling of natural and man-made processes (Unwin & Fisher, 2005). Many examples of GIS have been used in combination with remote sensing to assess land cover/land-use change as well as the analysis of urban growth. Weng (2002), for example, used GIS and temporal satellite data to monitor and map historical land cover/land-use dynamics in Zhujiang delta, China. Binh, Vromant, Hung, Hens, and Boon (2005) mapped the loss of forests and agriculture and expansion of shrimp farms in Ca Nuoc, Ca Mau province, Vietnam. GIS also used in many other disciplines such as ecology and agriculture. As an example, Joy and Death (2004) monitored and estimated the temporal changes of fish occurrence, and modelled the spatial habitat-suitability distribution of 14 freshwater fish species over the regional river network in New Zealand.

2.2 Image Processing

Digital image processing is referred to the use of computer algorithms to perform some operations on an image in order to either obtain an enhanced image or derive useful information from it. To effectively study the earth environment from remote sensing sensors, collected data should not be contained noise characteristics in radiometric and atmospheric conditions. However, the complexity of atmospheric conditions and technical limitations of sensors have frequently caused undesired noise in recorded images. Thus, pre-processing is an essential step to remove existing distortion inherent in an image, and to enhance the visualization and interpretation.

Natural resource and environment management decisions are based on accurate and informative inputs. The choice of classification algorithms plays an important role in deriving accurate land cover/land-use information to serve such purpose. Given the importance of classification choices, several comparative studies of algorithms usage in land cover classifications have been conducted (Hansen, Dubayah, & DeFries, 1996; Pal, 2005; V. F. Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012). However, little research has been done to compare the performance of the three classifiers, namely MLC, RF and ANN using Sentinel-2 and Landsat 8 data.

2.2.1 Pre-processing

Radiometric calibration and atmospheric correction are considered one of the most important components in digital image pre-processing. This process is to convert DNs to radiance at the sensor's aperture, and then continue to transform the radiance to Top of Atmosphere (TOA) reflectance. Surface reflectance at the ground is ultimately extracted by removing path radiance. Remote sensing data is processed to remove the effect of noise from atmosphere and other sources. The main aim of these operations is to enhance the visual interpretation, increase spectral separability of earth surface features and provide better inputs for further automated image processing algorithms (Maini & Aggarwal, 2010).

Visual enhancement is the improvement of brightness and stretching to assist the interpreters' visualization. Chuvieco (2016) defined criteria for visual interpretation and grouped them into a hierarchically according to their degree of complexity and spectral properties. Brightness and colour are the most prominent criteria for visual enhancement in terms of spectral resolution, while shape, size and texture can improve the spatial properties of the feature (Chuvieco, 2016). In multispectral sensors, colour composites play vital role in distinguishing objects by assigning each of the primary colours (RGB) to spectral bands. This is because each object has distinct spectral signature within the electromagnetic spectrum (EM). Vegetation, for example, is strongly reflected in the near-infrared region, while low reflectance can be seen in the visible spectrum (Figure 2.2).

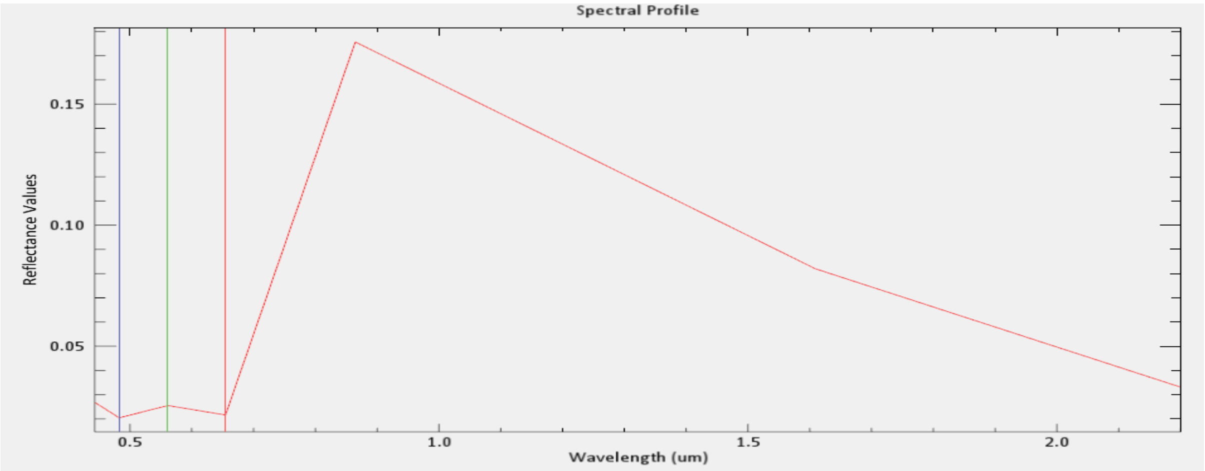


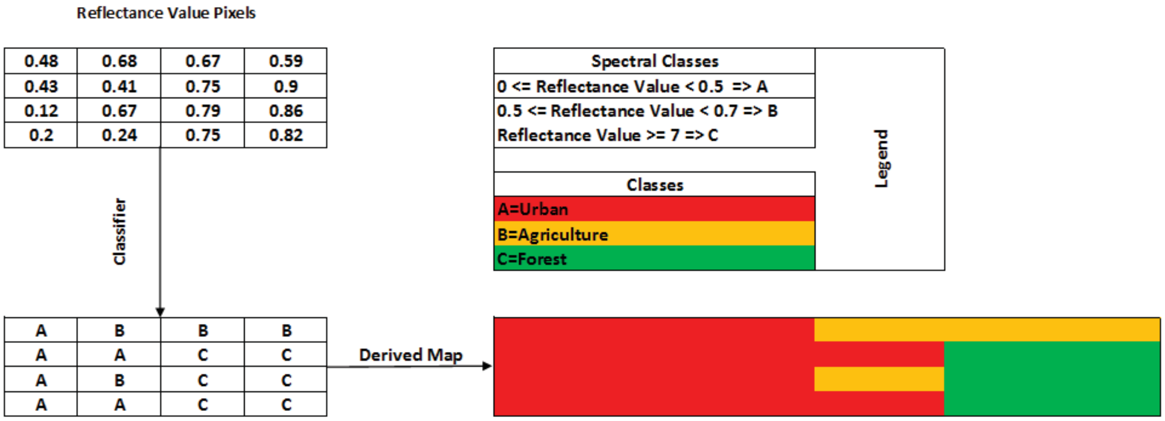
Figure 2.2 Spectral reflectance of green vegetation, Landsat 8 data

Image transformation reduces the redundancy of spectral information and existing noise in the images, while still maintaining the integrity of original data. Principal Component Analysis

(PCA) is widely used to transform original data to a new dataset with uncorrelated output bands. For automated classifiers, removing the redundancy of spectral information in each band and separating noise factors are important to create better classified output (Li & Yeh, 1998).

2.2.2 Classification algorithms

Deriving land cover/land information is one of the major remote sensing applications for environmental monitoring and management activities. Many automated digital image classification techniques have been developed and applied for organizing image datasets into classes based on their spectral properties using the similarity of spectral characteristics of each land surface (Figure 2.3). Supervised classification learning algorithms, for example, classify pixels into classes based on their spectral properties (reflectance values or DNs) with the selection of training data for each class manually defined by the interpreter (Campbell & Wynne, 2011; Chuvieco, 2016), and it is commonly used in the land remote sensing classification.



0.48	0.68	0.67	0.59
0.43	0.41	0.75	0.9
0.12	0.67	0.79	0.86
0.2	0.24	0.75	0.82

Spectral Classes
0 <= Reflectance Value < 0.5 => A
0.5 <= Reflectance Value < 0.7 => B
Reflectance Value >= 7 => C

Classes
🟥 A=Urban
🟨 B=Agriculture
🟩 C=Forest

A	B	B	B
A	A	C	C
A	B	C	C
A	A	C	C

Figure 2.3 An example of land cover classification using spectral class

Supervised classification algorithms are usually divided into two major approaches, namely parametric and non-parametric classifications. The traditional parametric methods (e.g., MLC and Minimum-Distance) are based on statistical assumption such as normal distribution of data. This assumption, unfortunately, is not always satisfied within the data. As an example, the distribution of reflectance values of training data in mining class for this study is skewed to right (Figure 2.4). Despite having the constraints of statistical assumption, the MLC considered among the most established algorithms for land cover change detection studies (Shalaby & Tateishi, 2007; Strahler, 1980).

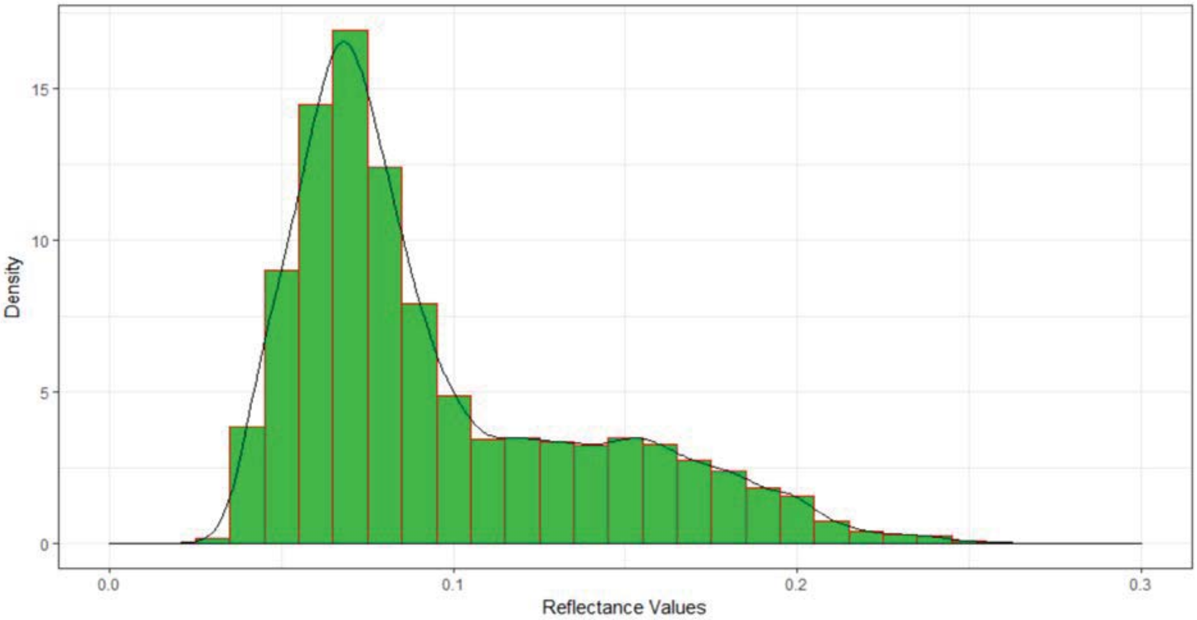


Figure 2.4 Right-skewed distribution of reflectance for mining class, Sentinel-2 band 3

To eliminate such limitations, non-parametric classification algorithms were developed and now are commonly used among the remote sensing community. Neural network and random forest are two examples of non-parametric classifications. The various developments of parametric and non-parametric classifiers resulted in a question of which classification technique should be chosen to provide the desired results. This concern has led to the comparative studies of various classifiers for land cover classification. Lu, Mausel, Batistella, and Moran (2004) compared four different classifiers, namely minimum distance (MD), maximum likelihood (MLC), extraction and classification of homogeneous objects (ECHO), and decision tree based on linear spectral mixture analyses (DTC-LSMA) to analyse multispectral data in Brazilian Amazon Basin. These were also investigated by Hansen et al. (1996), Szuster et al. (2011), Nangendo, Skidmore, and van Oosten (2007), and Pal (2005).

Although a substantial number of supervised classifiers were compared, little research has been done to investigate the accuracy of the three techniques: maximum likelihood, neural network and random forest using both Sentinel-2 and Landsat 8 data. This study attempts to examine the accuracy relationship between the classifiers, and to answer the question about the choice of classification algorithms, while the temporal land cover/land-use changes of Thai Nguyen province were mapped.

2.2.2.1 Maximum likelihood classifier (MLC)

Maximum likelihood (MLC) was originated from electrical engineering field of study (Nilson, 1925), whereas it has known in use for the applications of social sciences from the 1940s, and widely adopted in the field of pattern recognition in the following decades (Strahler, 1980). The MLC algorithm is based on statistical assumptions that the statistics for each training class in each band should be following Gaussian distribution or bell-shaped distribution. Mean and variance is calculated from each training class to form the probability distribution of each pixel in an image. An unknown pixel will be assigned to a specific class if it has the highest probability belonging to that class. A sufficient number of training data should be required for calculating mean and variance of each class (Richards & Richards, 1999). The technical procedure will be further discussed in Section 4.3.1.1.

Since the first initiation of Landsat satellite in 1970s, the MLC algorithm has become a popular approach for environment and earth scientists in monitoring and deriving physical earth information from remotely sensed data. Detecting land cover/land-use changes was probably the most widely used this classification technique. Dewan and Yamaguchi (2009b), for example, derived land cover/land use maps for Dhaka Metropolitan of Bangladesh between 1960 and 2005, and Muttitanon and Tripathi (2005) mapped temporal changes in land cover/land use from Landsat sensor TM in the coastal zone of Ban Don Bay, Thailand.

The MLC algorithm suffers from the fundamental drawbacks of all conventional classification procedures although this approach has several benefits, mainly connected with its theoretical simplicity and robustness (Davis et al., 1978). Most spectral information is lost in the process of transforming the remote sensing imagery to produce a classified image (Maselli, Conese, & Petkov, 1994), whereas it is frequently faced with the challenge of separating the mixed and spectral confused pixels. As a result, this traditional approach was usually considered less powerful than non-parametric classification algorithms (e.g., Random forest and artificial neural network). For instance, Erbek, Özkan, and Taberner (2004) evaluated the performance of the ANN and MLC algorithms using Landsat data and revealed that the ANN approach produced a higher overall classification accuracy than the MLC technique, and Kavzoglu and Mather (2003) also reached a similar conclusion. V. Rodriguez-Galiano, Chica-Olmo, Abarca-Hernandez, Atkinson, and Jeganathan (2012) conducted the comparison of classification performance between the RF and MLC algorithms, and found that the RF classifier consistently produced higher overall accuracy over the MLC.

2.2.2.2 Random Forest Classifier (RF)

Random forest is a powerful assemble learning algorithm, which has been increasingly used in the field of remote sensing. According to V. F. Rodriguez-Galiano et al. (2012), random forests are more robust because of its non-parametric nature and high accuracy. This approach has the capability to handle a large number of independent variables without variable removal and quantify the importance of variables in the classification. The detailed operation of RF will describe in Section 4.3.1.2.

A random forest is an ensemble of classification trees, where each tree is built on a subset of original data, and contributes with a single vote for the assignment of the most frequent class to the input data (Leo Breiman, 2001). To produce several trees, the RF adopted bagging or bootstrap approach to make the decision trees grow from different training subsamples. With the bootstrap method, each training subsample is created by randomly resampling from the original dataset with replacement. Each individual decision tree is grown on approximately two third of each selected subset, and the remaining is included as a part of model assessment called "out of bagging" (OBB). This bootstrapping process contributes an unbiased estimation of the model and reduce correlation between the individual trees. Also, the RF uses the Gini Index as a measure for selecting the best split at each node, which measures the impurity of a given element with respect to the rest of the classes (Leo Breiman, Friedman, Stone, & Olshen, 1984)

Land cover classification using a group of learning algorithms used to classify multi-spectral and hyperspectral satellite sensor imagery has received increasing interests. V. Rodriguez-Galiano et al. (2012) applied the RF method to map land cover/land-use using multi-seasonal imagery and texture in Spain. van Beijma, Comber, and Lamb (2014) investigated the use of the RF to map natural coastal salt marsh vegetation habitats in the Gower Peninsula, west of Swansea in South Wales. In addition, few authors have investigated the performance of the RF against other classification approaches such as MLC and SVM (support vector machine). (Gislason, Benediktsson, & Sveinsson, 2006) compared the performance of the RF against CART technique, and found that the RF outperformed the basic CART classifier by 4%.

Similarly, the performance of the RF was equally comparable to that of support vector machine (SVM) in terms of overall accuracy and training time, but required less number of user-defined parameters (Pal, 2005).

2.2.2.3 Artificial Neural Network Classifier (ANN)

The artificial neural network (ANN) is a machine learning algorithm, which was developed based on the inspiration of human brain networks. This approach was considered one of the latest added techniques in the collection of classifiers system, but it has been increasing adopted in the field of remote sensing studies. the ANN algorithm was widely recognized as an alternative for land cover/land-use classification because of its non-parametric nature and high overall accuracy (Kavzoglu & Mather, 2003; J. Paola & R. Schowengerdt, 1995). For example, there were many land cover/land-use studies published in major peer-reviewed international journals such as IEEE Transactions on Geoscience and remote sensing; International Journal of Remote Sensing; Photogrammetric Engineering and Remote Sensing and Remote Sensing of Environment (Figure 2.5).

Although the ANN algorithm is complexly described in its mathematical sense, it can be explained as a model of three layers: input layer, hidden layers and output layer (Kavzoglu & Mather, 2003). A set of random weights is assigned to the input layer, which will be passed to hidden layers with consideration of weightings correction. Number of hidden layers can be set by interpreter depending on land cover issues they are tackling. Output layer is the result of the network with activation function and will be repeated until an expected result is similar to the actual output.

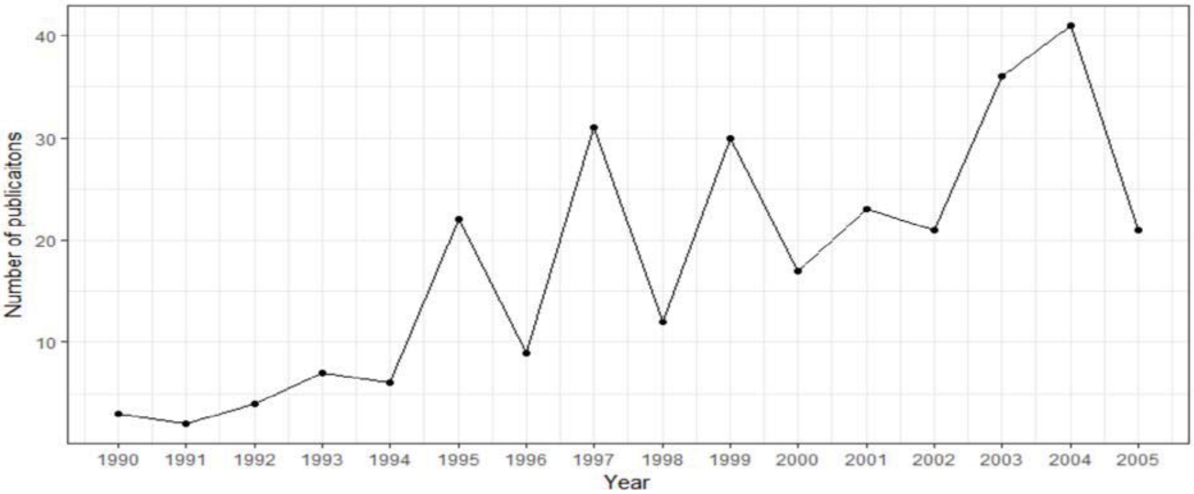


Figure 2.5 Land cover studies publications using the ANN approach between 1990 and 2005
Source: International Journal of Remote Sensing (J. F. Mas & Flores, 2008)

The ANN classifier requires no statistical assumptions of data normal distribution; it has been widely used in land cover classification studies over the past few decades. For instance, Civco (1993) used the artificial neural network learning algorithm as an alternative approach to derive land cover information from Landsat TM satellite. This study revealed that using the neural network technique to extract land cover maps from Landsat TM satellite data would provide more accurate and useful information for the integrated use with geographical information systems. G. F. Hepner (1990) investigated the use of the artificial network approach in processing satellite data with minimum number of training inputs. Although the ANN classifier in their study used a minimum set of training data, its classified output provided more accurate classification than conventional procedures.

The comparison of neural network with other classifiers in the remote sensing classification has been recently explored. J. D. Paola and R. A. Schowengerdt (1995), for example, described the comparison of the backpropagation neural network and maximum likelihood classifiers for detecting built-up land in various locations, namely Tucson, Arizona, Oakland, and California using Landsat TM imagery. The result of their study showed that the classified maps derived from neural network algorithm were more visually accurate in comparison with maximum likelihood. Similarly, Szuster et al. (2011) concluded that the ANN algorithm outperformed the MLC classifier with the respect of separating tropical coastal land covers. However, there are also several drawbacks associated with the ANN algorithm. It is frequently more computationally extensive and time-consuming than the MLC classifier (J. D. Paola & R. A. Schowengerdt, 1995). The ANN model also requires extensive amount of training input (Pal & Mather, 2003).

Although the performance of the ANN algorithm was extensively compared with the MLC classification approach, little research has been done to analyse the performance of land cover classification between the RF and ANN algorithms. Therefore, this study is in the hope of uncovering the capability of RF and ANN methods in classifying both Sentinel-2 and Landsat 8 data.

2.3 Temporal land cover and land-use monitoring and mapping

Before discussing land cover/land-use mapping in depth, the terms "land cover and land use' should be defined. In many instances, this term is used interchangeably, and so does this study. However, there are some distinctions between them. Land cover can be explained as the vegetation and man-made materials covering the earth surface such as forest, water and built-up infrastructure, while land use is commonly referred to specific purpose in which land serves such as rice fields, road networks and recreational parks. Meyer (1995) formally defined land cover as the physical condition of earth surface regarding to its natural and artificial feature categories, whereas describing land use as the interference of human activities on land surface.

Although there are differences in defining the land cover and land use, they were changed and influenced by human practices (Meyer & Turner, 1992). Many studies have been conducted to examine the relationship and interaction between land cover/land use and its dynamic environment. Meyer and Turner (1992) stressed that every parcel of land on the earth surface is unique in the cover it occurs and consistently linked to the developments of human activities. Anthropogenic activities have become recognized as the major force shaping the earth environment (Meyer & BL Turner, 1994) through the transformation and modification processes of land surface. Changes in land cover/land use may be caused by natural processes or human activities, but primarily by the conversion of natural land or forests to other land use purposes in which urbanization is the most obvious. Developing countries, for example, are undergoing the rapid growth of population and urbanization, which converts major agricultural and forestland to built-up uses (Meyer & Turner, 1992; Zhou et al., 2004), and even convert conservation zones or wild forests to urban development areas.

Changes in land surface may lead to serious consequences for both living animals and the environment. Many studies were carried out to estimate the impacts of land cover/land-use change on the environment and human health. (Foley et al., 2005), for example, revealed that changes in habitat environment would modify the transmission of infectious disease and lead to outbreaks. In Africa, Asia and Latin America where the increase of tropical deforestation was coincident with an upsurge of malaria is an example of this. Jha et al. (2005) investigated the impact of forest fragmentation on species diversity in India, and found a strong association between the loss of biodiversity and forest fragmented. Similarly, (Jetz, Wilcove, & Dobson, 2007) revealed that changes in land cover and global climate system could cause substantial species extinctions, while several endangered species are estimated to disappear over the next few decades. Dale (1997) also indicated that there was a strong association between climate change and land cover/land-use change, and its impact on ecosystem. Especially, land cover/land-use changes can alter atmospheric condition and cloud formations, which influence both local and global climate environment.

Therefore, the provision of up-to-date and accurate land cover/land use map is essential to understand and evaluate the environmental consequences of its changes. Land cover information provides critical inputs for local and national agencies to better monitor and manage their environment and resources, particularly regarding to the formulation of socio-economic development and planning policies (Anderson, 1976; Campbell & Wynne, 2011). With the increasing availability of remote sensing data, statistical software and GIS tools, monitoring land cover/land use changes will continue to provide essential information for government, non-government organizations and universities to make better decisions on their natural resource and environment management.

2.4 Change detection and Accuracy assessment

2.4.1 Change detection

Change detection is the process of identifying the differences of land surface using multi-temporal images acquired in the same extent of geographic area (Singh, 1989). Remote sensing data are repeatedly acquired over the same geographic coverage at regular time intervals. Therefore, the application of change detection techniques plays a vital role in monitoring and analysing the environmental, land cover, deforestation and disaster assessments.

Many change detection techniques have been developed and applied in the context of land cover classification. Among the different techniques, a post-classification comparison (PCC) was widely used to monitor and map land cover/land use change between two or more time intervals (Singh, 1989). After each remote sensing image is classified independently, two classified images are compared through the confusion matrix in order to create a change “from to” map. Many land cover monitoring studies have applied this technique and stressed its advantages. Shalaby and Tateishi (2007) emphasized that the comparison of two independently classified images with distinct dates using the post-classification detection is the most effective technique because it can minimize the problem of normalizing atmospheric and sensor variability. Another comparison study of change detection conducted by J.-F. Mas (1999) reported that post-classification not only outperforms over the other change detection techniques, but also provides the best information about the nature of its change. In addition, many other studies also reached the similar conclusions and prioritized post-classification comparison over other techniques (Liu & Zhou, 2004; Serra, Pons, & Sauri, 2003; Weismiller, Kristof, Scholz, Anuta, & Momin, 1977; Yuan, Sawaya, Loeffelholz, & Bauer, 2005).

2.4.2 Accuracy assessment

Accuracy assessment is considered the last step in the digital classification, which validates a thematic classified image and actual ground land cover type or reference data (Chuvieco, 2016). Validation is a critical stage to quantify how well our classified land cover maps compare to actual land cover on the ground. The validation is carried out using ground control data, and compare it to the classified resultants to produce a table, called confusion matrix or contingency table (Campbell & Wynne, 2011; Joseph, 2005; Richards & Richards, 1999). Based on the information from the confusion matrix, many measures can be derived to assess the fitness of a model to particular context, including an overall accuracy, producers’ and users’ accuracy for each class and Kappa statistic.

According to Congalton and Green (2008) and Anderson (1976), an overall accuracy is obtained by dividing the total sum of diagonal over the total number of samples, while Kappa statistic is calculated more sophisticated. Kappa coefficient takes chance agreement and covariance term into account in calculation process. The result of Kappa statistic ranges from 0 to 1, where values close to 0 indicate low agreement between two datasets, and values close to 1 indicate high agreement. These two coefficients indicate how well one classifier performs. Usually, the acceptable level of overall accuracy should be from or above 85% (Anderson, 1976; Foody, 2002; Thomlinson, Bolstad, & Cohen, 1999). Additionally, producers' and users' accuracies provide essential information to evaluate the performance of a classification algorithm with respect to the separation of certain land cover/land-use categories.

Although the confusion matrix is widely used in the field of remote sensing, it was still in debate and yet hardly adopted as an accuracy assessment standard (Foody, 2002). This is because Kappa statistic is not always appropriate due to chance agreement probability (Morris et al., 2008). Another aspect is that the target accuracies (overall accuracy>85%) commonly recommended by DeGloria et al. (2000); (Foody, 2002) are rarely achieved. Despite the drawbacks of the confusion matrix, it still plays an important role in validating the classified products derived from remotely sensed data (Foody, 2002).

3 STUDY AREA and MATERIALS

This chapter describes the study area and techniques that are applied in the collection, analysis and presentation of both satellite images and ground truth data.

3.1 Study Area

Thai Nguyen province is located in the northeast region of Vietnam and covers 9 local districts and parts of the Tam Dao national park with a total area of roughly 3534.45 km^2 (Figure 3.1). The study area is prominently characterized by subtropical climates (e.g., warm to hot in summers and cool to mild in winters) with strong inter-seasonal variability. Mean annual temperature and precipitation are comparably high, approximately 25ºC and 2250 mm respectively, while its elevation ranges from 4 meters in flat areas to 1591 meters in the Tam Dao mountain range (NASA & Ministry of Economy Trade and Industry of Japan, 2011). The climatic and topographic variabilities produced complex characteristics of subtropical landscapes such as semi-deciduous, scrubs and maintain ranges (Phuong, 2007). At present, forest and scrubs are predominant and cover most of the north and northwest, while agriculture and urban/built-up primarily characterize lower areas and drainage corridors in the south and central.

The historical vegetation of the study area was described as dense and diverse, with a prominence of the eucalyptus trees and shrubs (Hoang Ngoc Ha, 2008). Clement and Amezaga (2008) revealed that forest restoration programs implemented between 1985 and 2005 has significantly increased forest cover from 29.2% to 37.6% respectively. However, the recent expansion of industrial and urban zones and other infrastructure networks has increased pressure on local land-use. Conversion of agriculture and forest was largely occurred to serve industrial and urban development in the southern flat areas and suburb zones of districts and towns. As an example, Nam Thai urban zone was built in 2014 and is expected to accommodate from 25 to 30 thousand people

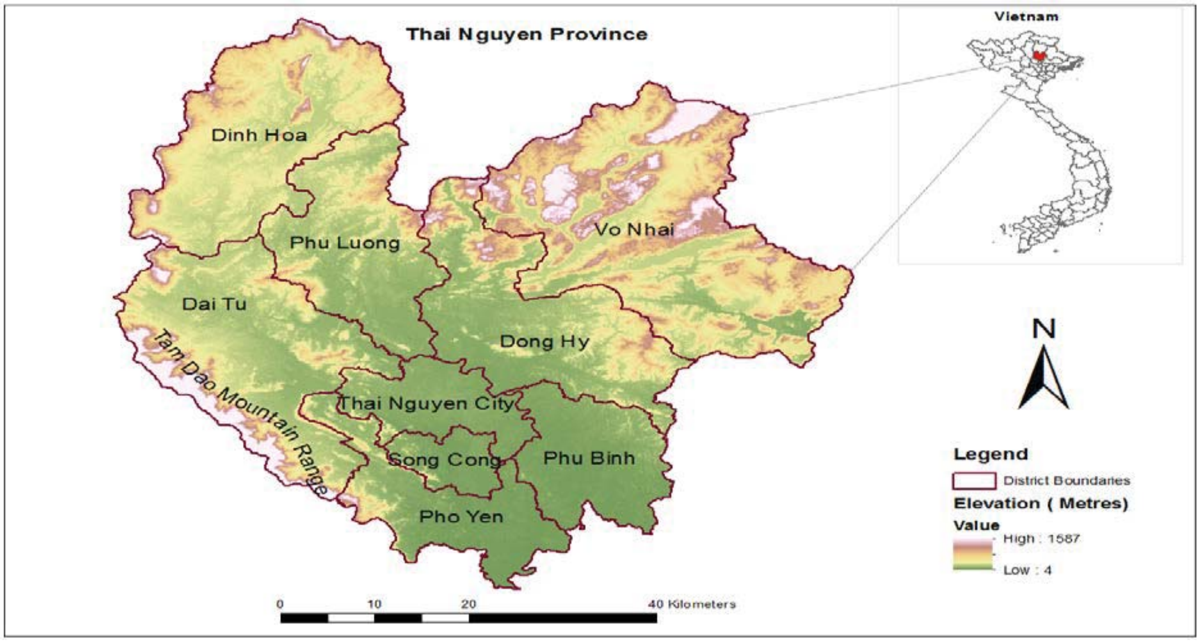


Figure 3.1 A map of the study area

The study area is home to nearly 1.2 million people in 2016, which represents roughly 1.3% of the Vietnamese population. The population of Thai Nguyen increased significantly over the past two decades (Hao Ho, 2015), and the percentage of rural population against urban population remained high. For example, the rural population in 2010 was approximately 74%, while urban residents only accounted for 26% (General Statistics Department of Vietnam, 2011). However, the movement of residential and commercial land use to rural areas at the periphery of urban areas has recently increased, and this trend is expected to continue occurring. The local government estimated that urban residents will account for 48% of the total population by 2025, whereas urban areas are expected to increase by 9% between 2015 and 2025 (Thai Nguyen People's Committee, 2009).

3.2 Biophysical and Scio-economic Characteristics

3.2.1 Topology and Hydrology

The topography of the Thai Nguyen province generally slopes from North to South with prominent high mountain landscapes and river networks (Figure 3.2). The structure of terrain is highly weathered with many caves, small valleys, rivers and lakes. In the southwest, Tam Dao mountain range spans along high mountains and the Van Lang plateau and Dai Tu paddy fields, while the north and northeast are primarily covered by lower mountains, and the south lies on relatively low land. Cau river, for example, flows through 7 provinces (e.g., Bac Kan, Cao Bang, Thai Nguyen and Bac Giang provinces) with 288 km in length (Nguyen, Everaert, Gabriels, Hoang, & Goethals, 2014). In addition, the development of aquaculture also increased in recent years. For example, the total area of aquaculture was 5.881ha, including small lakes, pounds and rivers. These days, the province is planned to develop integrated agriculture-aquaculture practices. In the words, traditional rice production is combined with aquatic development.

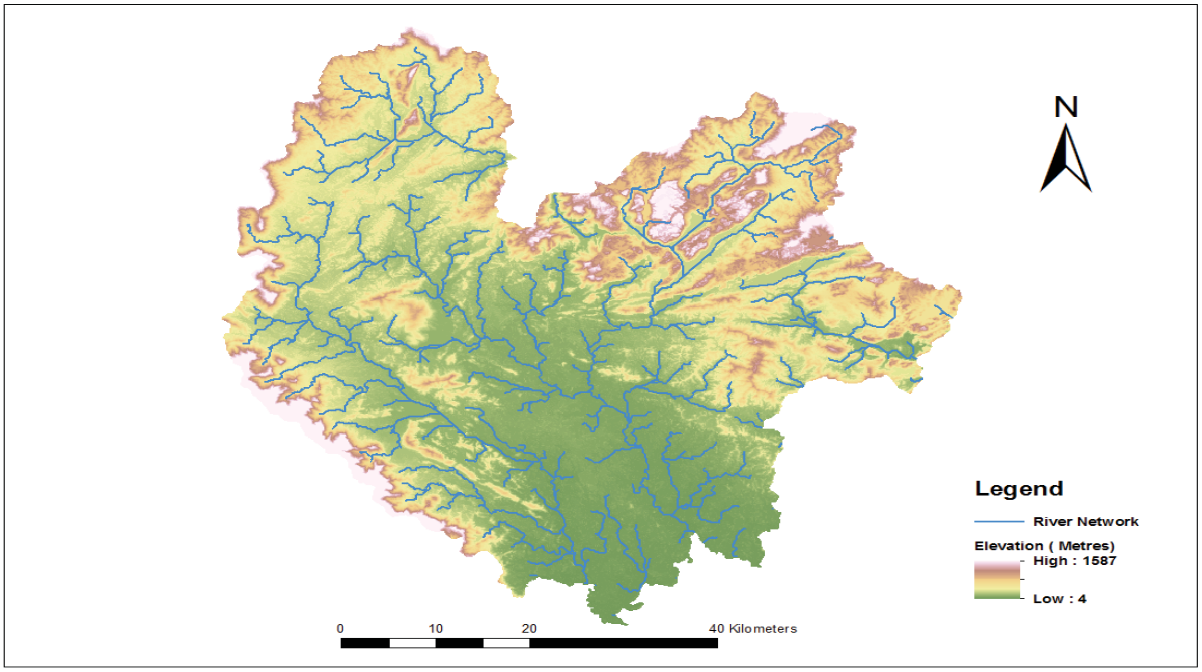


Figure 3.2: Elevation and water network for Thai Nguyen derived from DEM

3.2.2 Geology, Soil and Mineralogy

Most territory of Thai Nguyen was formed about 240 million years ago and ended 67 million years ago, but the biophysical process continued to undergo continent changes until around 50 million years. These natural change processes were resulted in shaping and dividing Thai Nguyen province into three main geological regions as today. Mountainous land is formed from the decaying of magma, rocks and sedimentary rocks with a major coverage, hilly land is made of condensed sand, clay and ancient alluvia while a small portion of plains is scattered along streams, rivers and lakes.

3.2.3 Flora and Fauna

The tropical and subtropical moist evergreen forests dominate the province and cover over Tam Dao range and into the northern landscape. The diversity of vegetative landscapes and elevation variations is home to many valuable medicinal plants and animal species. Huong, Anh, Yen, Thanh, and Thin (2012) conducted a study on the current status of medicinal plant species in Thai Nguyen province and found that there is a collection of 25 species of vascular plants (e.g., Anoectochilus calcareus, Stephania kwangsiensis and Tacca subflabellata) in the protected areas, of which 20 species are placed in endangered species list (Vietnam red book). In addition to plant species, wildlife animals are also found in some places such as Than Sa conservation area and Tam Dao mountain range. The final report of rapid assessment of mammals in the Tam Dao national park prepared by GTZ Office Hanoi (2005) revealed that 77 species are recorded in the park (e.g., Rodentia, Primates primates and Pingolin Pholidota), of which many species are considered to be threatened species.

3.2.4 Population and Economic Development

The population of Thai Nguyen has been increasing over the past decades and reached 1.156 million in 2016, an increase of 5.6% from 1997. The distribution and density of demography greatly varies between rural and urban areas. The major population lives in rural region with a low density, while urban residents are rapidly increasing from both seasonal and unseasonal flow of migrant farmers and workers.

Although urban population has been increasing over years, agriculture and livestock sectors are still traditional economic activities of the province. Recent economic development strategies have been transforming the traditional agricultural dominant land uses into a province with modern urbanized and industrialised zones. These development programs led to the emergence a series of industrial complex zones, from mineral mining and processing zones to high-tech corporations, investing in the province. Samsung launched an investment package of approximately 2 billion USD in 2012 to build a high-tech complex in the hope of producing approximately 100 million of products per year (Thai Nguyen Department of Commerce and Industry, 2012). Thai Nguyen local government approved 18 industrial zones with a total area of 620 ha in 2010.

3.3 Data Collection

A substantial number of satellite data and thematic maps (Table 3.1) were collected from NASA Landsat program, Copernicus program of European Space Agency, DIVA-GIS and Thai Nguyen Department of Land for this research project. Reference data were also captured between December, 2016 and January 2017 using a handheld GPS to enable the validation process of derived land cover/land-use maps, while Google Earth’s high spatial resolution imagery was also used to assist in the selection of training polygons.

Table 3.1 Data layers used in this study

Data	Data Category	Spatial Resolution (m)	Date of Acquisition	Data Source
Landsat ETM+	Imagery	30	4/11/2000	USGS[a]
Landsat OLI	Imagery	30	6/10/2016	USGS[a]
Sentinel-2	Imagery	10	6/10/2016	ESA[b]
Road	Vector			DIVA-GIS[c]
DEM	Elevation	30		USGS[a]
Local Land Statistics			12/2016	TNMT[d]

[a] U.S. Geological Survey

[b] European Space Agency, Copernicus Program

[c] DIVA-GIS

[d] Thai Nguyen Department of Natural Resources and Environment

3.3.1 Satellite Data and Thematic Maps

The Landsat Enhanced Thematic Mapper (ETM+) image was acquired on 4th November 2000 (mid-winter), while Landsat Operational Land Imager (OLI) and Sentinel-2 were acquired on 6th October 2016 (early winter). The acquired Landsat 7 and Sentinel-2 data had no clouds over the study area in both years 2000 (Landsat ETM+) and 2016 (Sentinel-2) respectively, but little cloud was found in Landsat 8 imagery over Tam Dao mountain range and other landscapes in the northeast (Figure 3.3). Fortunately, all this cloud was presented in the mountainous region so it did not affect the project outcomes.

The Landsat 7 carries the ETM+ sensor and has eight spectral bands, ranging from visible to mid-infrared and a portion of the electronic spectrum wavelengths, whereas the Landsat 8 carries OLI sensor to capture spectral signatures from the Earth's surface features within eleven spectral bands. These satellite systems were designed to collect data over the entire globe with a 185-km swath and 16-day revisit interval. Thai Nguyen province is entirely contained within one single Landsat ETM+ and Landsat OLI image path 127 and row 45 with Landsat Worldwide Reference System-2 (WRS-2).

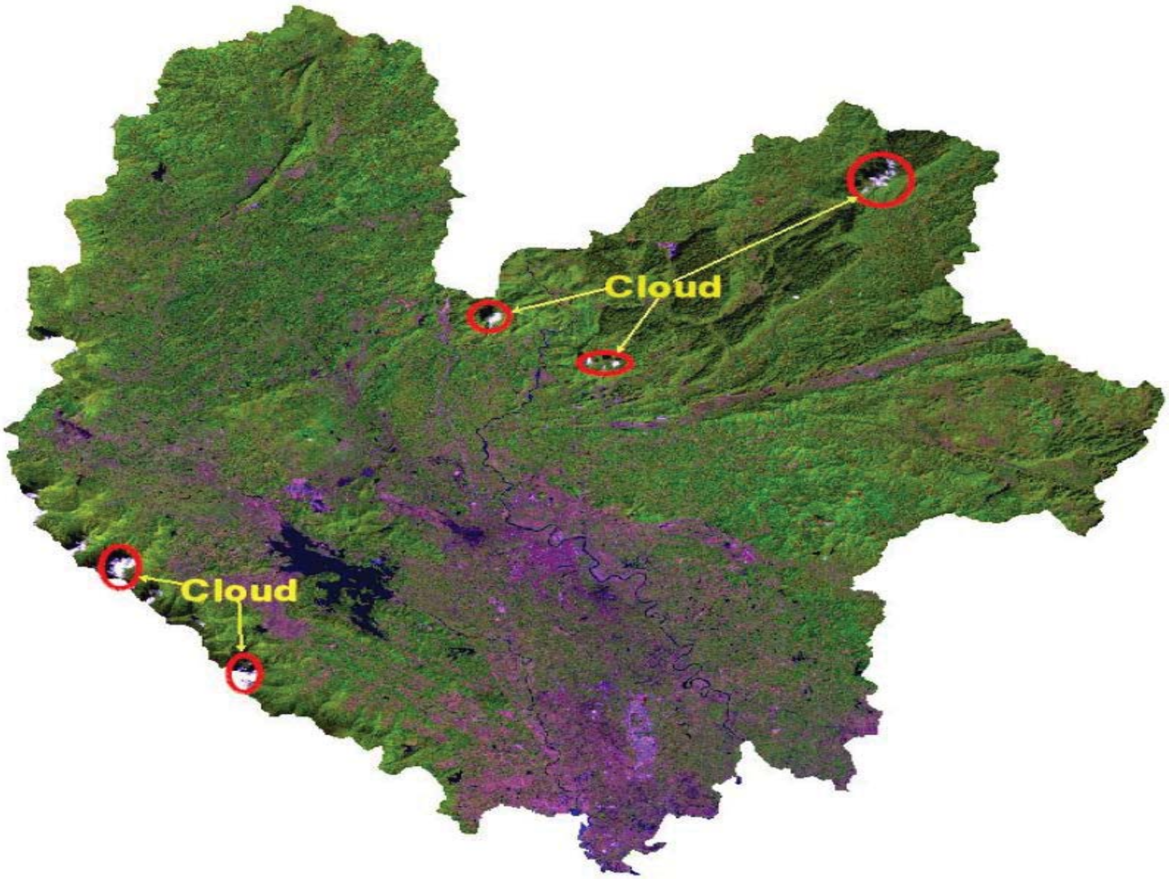


Figure 3.3 Clouds in the Landsat 8 image

Sentinel-2 mission, as a part of Copernicus program from European Space Agency, comprises twin polar-orbiting satellites (Sentinel-2A and Sentinel-2B) in the same orbit. The Sentinel-2A was launched in June 2015 consisting of a single multi-spectral instrument (MSI), while the Sentinel-2B (MSI) just recently launched in March, 2017. The Sentinel-2 provides relatively high spatial-resolution images with 13 spectral bands in visible/ near-infrared (VNIR) and shortwave infrared spectral range (SWIV), as discussed in Section 2.1.1. Compared to Landsat satellite system, Sentinel-2 was designed to collect high spatial images over the entire planet with a wider swath of 290 km and at a shorter revisit frequency of 5 days. The Sentinel missions are promising to support different application domains, from natural disaster management to humanitarian assistance and many others. There are four tiles of Sentinel-2 covering the entire Thai Nguyen province.

3.3.2 Training and Reference Data

The reference data were collected from 25th December, 2016 to 5th January, 2017 in accessible roads and areas using a handheld GPS receiver (Trimble Juno SB) from Institute of Environment and Agriculture, Massey University. To ensure the quality level of collected ground data, a random stratified sampling method was implemented and only comparably homogeneous areas were chosen with a visual estimation of at least four Landsat pixel cells (3600m^2) for any given ground point for assessing the accuracy of Landsat 8 and Sentinel-2 maps.

A total of 169 reference points was obtained over a large extent area, and it was used for accuracy assessment. According to Congalton (1991), a minimum number of around 75 to 100 ground points should be collected for each class. This study, however, was able to collect a total of 169 ground points due to the limitations of time and road accessibility. The distribution of ground control points for each land cover type ranged from 23 to 54 data points. For example, forest and water land classes had the lowest number of ground points, 23 and 24 points respectively while the largest number of ground points was in agriculture class with 54 data points. All the sample points were recorded in the world geographic coordinates of WGS-84.

Local land use statistics were collected from the Thai Nguyen Department of Natural Resource and Environment (TNMT) in 2016 and used for area match-up assessment of 2016 classified maps. Every year, the local government implements land statistics based on existing cadastral documents to monitor the changes among land cover/land-use types. The land statistics are undertaken in each administrative commune, and are aggregated to form a map of land use for both the district and the province. The land use statistics are essentially area assessments made by the administrators of each commune. Thai Nguyen province is required by law to implement a province-wide land survey program in every 5 years, and land statistics update every year. It is also important to note that planned areas for a specific purpose are included in land statistics document although these areas may not be occupied by built-up infrastructure on the ground.

Training data used in this study were selected manually using the ENVI 5.3 software. The pixels in each polygon were visually uniform and representative for that class. A total of 20383 and 183288 training data pixels was recorded for Landsat 8 and Sentinel-2 data respectively. The number of training pixels used for classifying Sentinel-2 image was much bigger than that of Landsat 8 because different pixel size between two images. The selected number of training data accounted for 0.29% of total pixels for each scene (Sentinel-2 and Landsat 8). While the ANN and MLC algorithms can accommodate a large number of training pixels to train its models in the ENVI software, this figure for the RF reduced to much smaller in the case of Sentine-2 data. This is because the RF algorithm in R programming cannot handle large datasets in memory. The details of distribution of training data points of Sentinel-2 and Landsat 8 data for each land cover/land-use are presented in Table 3.2.

Table 3.2 Training data statistics for Sentinel-2 and Landsat 8 data (pixels)

	Landsat 8			**Sentinel-2**		
	MLC	ANN	RF	MLC	ANN	RF
Agriculture	6204	6204	6204	55779	55779	4000
Forest	7183	7183	7183	64693	64693	4000
Mining	2518	2518	2518	22685	22685	4000
Urban	1750	1750	1750	15688	15688	4000
Water	2723	2723	2723	24443	24443	4000
Total	20383	20383	20383	183288	183288	20000

3.3.3 Software

Image pre-processing and processing was performed using SNAP (Zuhlke et al., 2015) software, ENVI (Guide, 2008) and R programing language (Bivand, Pebesma, Gomez-Rubio, & Pebesma, 2008), while Quantum GIS and ArcMap was used to produce the summary statistics and make map layouts of classified outputs (Logan, Hanson, & Seeger, 2014). Trimble Juno SB GPS was used to collect ground data for this study. All land cover/land-use classified maps used RF algorithm in this study were implemented in R programming language, and code can be found at github.com[1].

4 METHODS

4.1 Image Pre-processing

Pre-processing operations are very important before using satellite imagery for environmental applications. In most cases, some degree of pre-processing is needed to correct for any distortion or removing any cloud existing in the images (Chuvieco, 2016). The aim of these operations is to enhance visual interpretation, increase spectral separability of earth surface features and provide better inputs for further automated image processing algorithms (Maini & Aggarwal, 2010).

Many commonly used and established functions and techniques in the rectification of earth observation satellite data have been developed and tested by environmental scientists, geologists, cartographers, ecologists, biologists, oceanographers, foresters and computer engineers. Although the categorization of pre-processing operations is not clearly defined and sometimes interchangeable, Campbell and Wynne (2011) grouped such techniques into four main clusters: radiometric corrections, geometric corrections, enhancement and

[1] R code for RF algorithm at https://github.com/tuyenhavan/Sentinel-Data1

transformations. In this study, both raster and vector data are georeferenced to the WGS 1984 UTM 48 N coordinate system. Due to external and internal factors of noise caused by sensors and atmospheric conditions, the obtained remote sensing data may contain a certain amount of unwanted noise. To increase the quality of the images and visual interpretation, radiometric calibration and atmospheric corrections were applied to both Landsat and Sentinel-2 data. In addition, the process of combining multiple images from a single sensor to create a mosaic is also briefly described as it was applied to the Sentinel-2 tiles.

4.1.1 Radiometric Correction

Radiometric calibration is a common pre-processing step in remote sensing to compensate for radiometric errors from sensor defects, variations in scan angle and system noise; the aim is to produce true-spectral images at the sensor. The underlying theory of radiometric correction is to convert Digital Numbers (DNs) to Top of Atmosphere (TOA) radiance values using the bias and gain values specific to individual bands. The resulting radiance values are further transformed using irradiance values to TOA reflectance at the sensor. For Landsat and Sentinel-2 imagery, the process of converting DNs to surface reflectance can be done using the Semi-automatic classification plugin in QGIS (Congedo, 2013).

4.1.2 Enhancement and Transformations

Landsat data used in this study have moderate spatial resolution, (30-m), and since they have been acquired at two different times may contain internal and external factors of noise caused by sensor and atmospheric conditions. Therefore, enhancement of the satellite imagery is essential to increase visual discrimination between earth object features and remove noise in each image. The complementary capabilities of the human mind and computer tools are an excellent way to enable a visual interpretation of images from low to moderate spatial-resolution sensors (Shalaby & Tateishi, 2007).

There are a significant number of enhancement techniques in use to support the visualization of satellite images. Changing the band combination, from natural colour image to various false colour composites, is possibly the simplest technique to adjust images for human eye interpretation. As stated by Chuvieco (2016), the electromagnetic energy signals received by sensors from object features across different spectral regions vary with land cover categories and the biophysical and biochemical properties of surface features. Vegetation, for example, reflects strongly in the near-infrared region (NIR) while it is mostly absorbed in the visible region (VIS). A false colour composite (Figure 4.1) represents healthy vegetation as bright red in the Tam Dao mountain range and Vo Nhai district, while Thai Nguyen city and all the other urban areas are distinguished easily by the lighter tones in this image.

In multi-spectral remote sensing data, bands frequently contain redundant information either because some bands have similar spectral energy or because some features have similar radiances across spectral regions. Principal component analysis (PCA) has been developed to remove redundancies in the multi-band images without losing a substantial amount of original spectral information (Chuvieco, 2016). According to (Li & Yeh, 1998), PCA became one of the most widely used technique for producing uncorrelated output bands, separating noise factors and compressing remote sensing data. The process of principal component analysis is that the first principal component will store the highest variance, while the second principal component will describe the most of the remaining variance that is not explained by the first, and so forth (Taylor, 1977). This process may generate a large number of principal components, but the first three or four principal components will describe more than 95% of variance while the remaining individual raster bands can be dropped (Jensen, 1986).

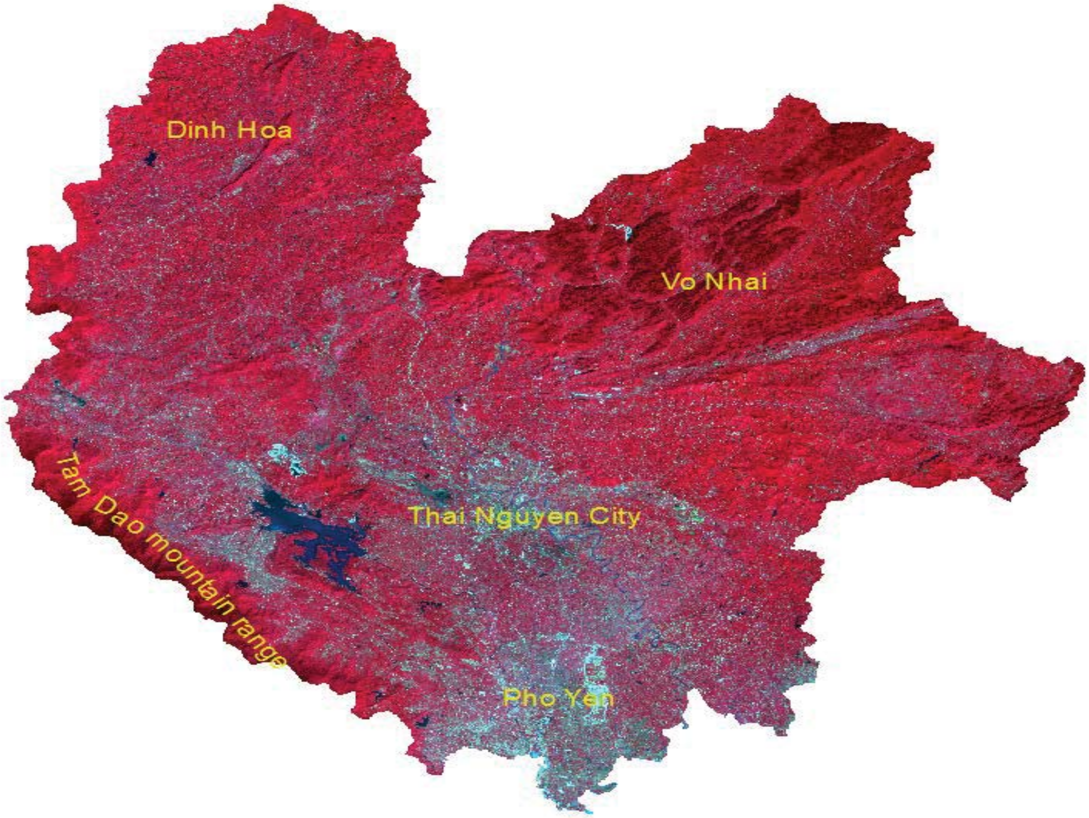


Figure 4.1 A false colour composite of the study area (Sentinel-2 bands 3, 4 and 8)

4.1.3 Mosaic from Multiple Images

Satellite images are often acquired in a designated geographic area, and must be mosaicked to cover larger areas. Images of a mosaic dataset may represent the issues of inconsistent colour tones and uneven brightness intensity (Figure 4.2 A). Therefore, colour balancing is essential to remove such problems and obtain a more consistent seamless image.

The mosaic operation is used to merge a collection of independent georeferenced raster datasets to a single seamless image. While Landsat data cover the entire extent of the study area, four Sentinel-2 tiles need to be collected to contain the entire Thai Nguyen province. In this study, Sentinel-2 data were pre-processed using SNAP and QGIS software to subset the tiles to cover only the extent of the province and contain the four 10m resolution bands before converting it from DIMAP to Geotiff format. The Geotiffs ware opened in ENVI software and a Seamless Mosaic operation applied with colour correction to create a single mosaic seamless image (Figure 4.2). Noticeably, after applying the seamless mosaic operation, the colour and contrast are more consistent between the four subsets although there is a small scatter of brightness stretching the south.

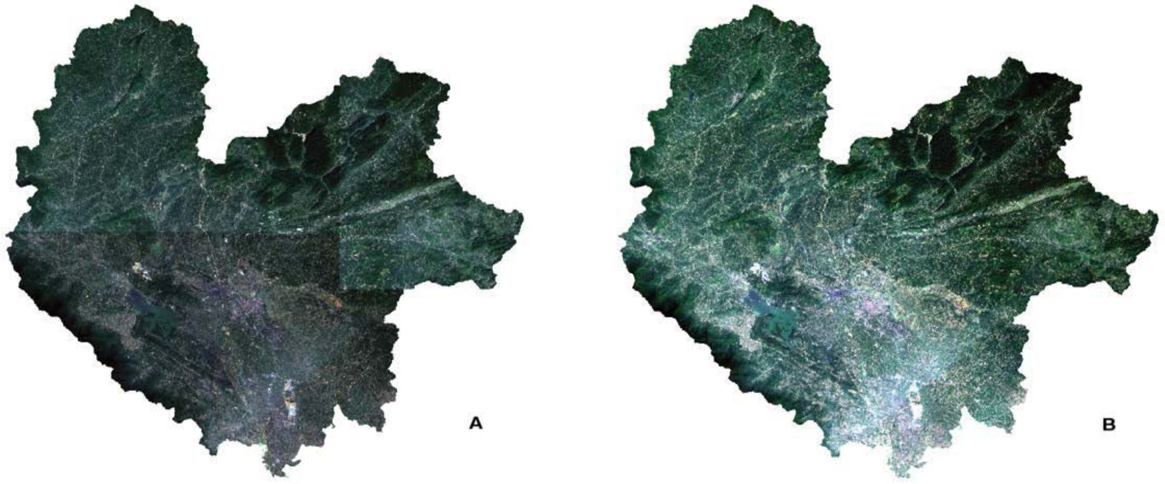


Figure 4.2 Natural colour Sentinel-2 imagery without colour balancing (A), and colour balancing (B)

4.2 Classification Scheme

The complexity of local landscape, topography and tropical climate results in a diverse pool of land cover. To form an initial classification scheme for the study area, the following process was followed. First, several band combinations for both Landsat and Sentinel-2 data were made to produce false colour composite images that could be interpreted visually. Secondly, the Normalized Difference Vegetation Index (NDVI) was calculated to simulate the density of vegetation surface (Figure 4.3). Finally, extensive field observation was carried out to verify and add or remove classes from the initial classification scheme.

Initially, seven land covers (built-up land, agricultural land, forest land, water bodies, bare land, mining extraction and tea plantation) were identified in the study area. However, during the field trip it was observed that the area of bare land and tea planation accounted for only a small portion. In addition, tea plantations were usually grown at the foot of mountains or under the canopy of other tall trees. The mixture of tea plantations, shrubs and forest resulted in spectral confusion, and in turn making it challenging to distinguish those similar features by the algorithms. Thus, the tea plantation was merged with forest whereas bare land was grouped into mining extraction.

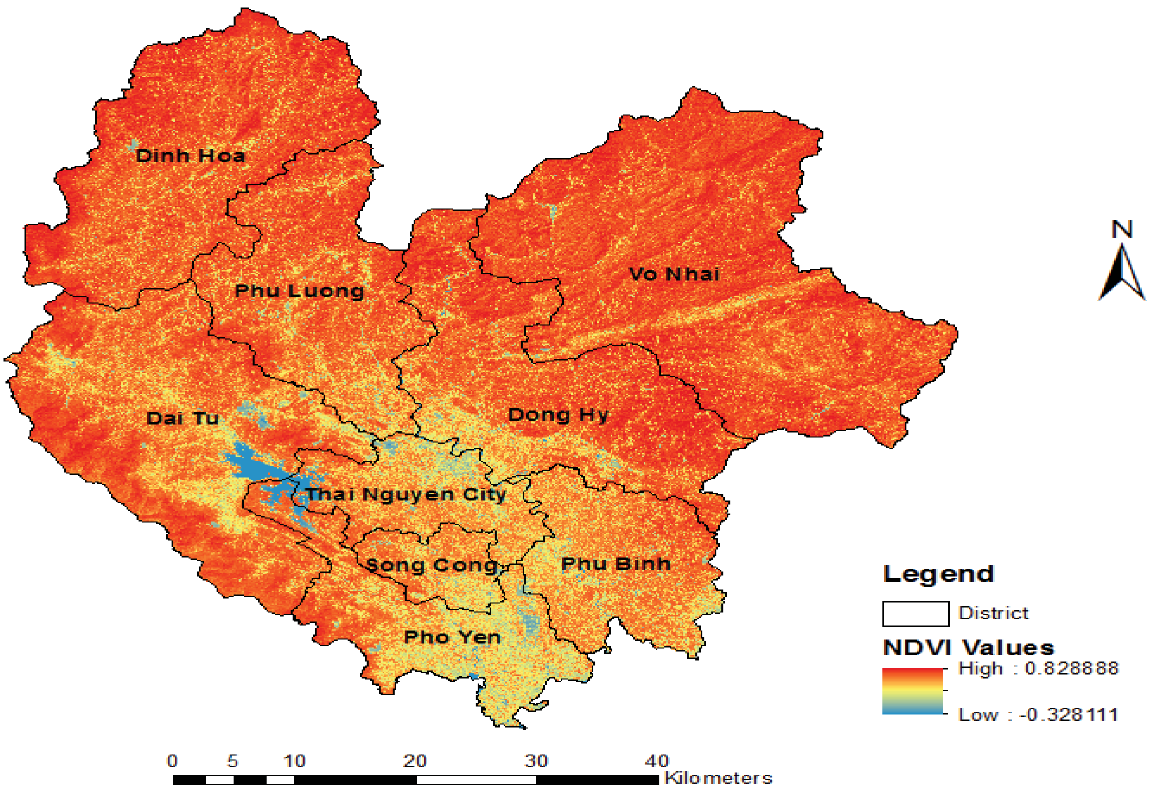


Figure 4.3 Normalized differenced vegetation index of the study area (Sentinel-2 bands 8 and 4)

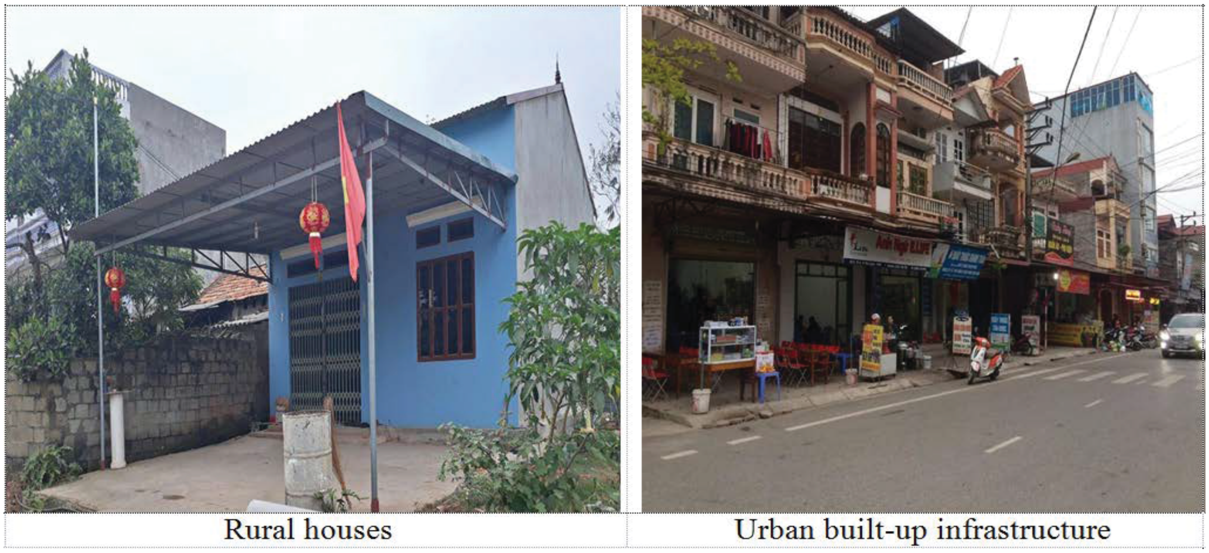
Up-to-date land use statistics such as the 2016 land use statistics and inventory from the local government, and land cover classification by Anderson (1976) were used as references for constructing the land cover scheme. The final classification settled on five land cover classes: urban/built-up land, agriculture, forest land, water bodies, and extraction. Although the Thai Nguyen 2016 land-use classification system did not have mining extraction, this study included it the classification scheme because of its large area. A detailed scheme of land cover classes for this study is presented in Table 4.1.

Table 4.1 Description of various land cover classes in Thai Nguyen, Vietnam

Classification Scheme	Description
Urban/Built-up Land	Rural houses and urban buildings Road network and utilities Industrial zones, Commercial Complexes and ongoing construction areas Mixed Urban or Built-up land Other impermeable surfaces
Agricultural Land	Crop and Pasture Orchards, Groves, Nurseries and horticultural Areas Other Agricultural Land
Forest Land	Deciduous Forest Land, Evergreen Forest Land Mixed Forest Land
Water	Streams, Rivers, Canals, Estuaries and Reservoirs Lakes and Ponds
Mining/Extraction	Iron mines and coal mines Bare Exposed Rocks, Transitional Areas Strip Mines, Quarries and Gravel Pits Mixed Barren Land

❖ Urban and built-up land

Built-up land is comprised of residential buildings, houses, industrial zones, urban and rural properties, road network and other infrastructures (Figure 4.4). Urban/built-up is mostly concentrated in Thai Nguyen city and central towns of each district with current urban and industrial development. Fewer urban areas are located in the north and the more other rural regions in the east and west. In recent years, built-up areas have merged in the southern parts, especially the Pho Yen and Song Cong districts. The expansion of urban and industrial services has removed green open spaces and other vegetation in recent years for constructing industrial and residential zones in the central districts and cities. Rural houses are often scattered along primary roads and sometimes covered by forest canopy, which make it challenging to separate those land cover classes.



Rural houses

Urban built-up infrastructure

Figure 4.4 Rural houses (left) and urban built-up infrastructure (right)

- Agricultural land

All land used for growing crops and feeding livestock is classified as agricultural land (Figure 4.5). During the field visit, it was observed that there were three main types of crops grown in the study area. Paddy rice is grown mainly in the southern region and other rural areas with access to water (e.g., lakes, irrigation network and wetland). Noticeably, a large rice growing area can also be seen in the west and far northwest due to the availability of water such as lakes. Corn, cassava and sweet potatoes are substantially found in rural areas and along major rivers, while short-rotation crops are also prominent in the south. Vegetables are mostly observed in suburban areas and along Cau river as these areas are planned to provide fresh vegetables for Thai Nguyen city.

Rice and other long-day crops are prominently grown across the province, but most rice fields have been harvested or replaced by corn or sweet potatoes and other short-rotation crops at the time of ground trothing. In many mountainous areas, local farmers and small business family often make use of open forest to grow cassava or combine cassava with commercial forests in order to increase the productivity. Similarly, fruit trees are also usually grown under the canopy of big trees, which resulted in similar cover between agriculture and forest (e.g., shrubs).

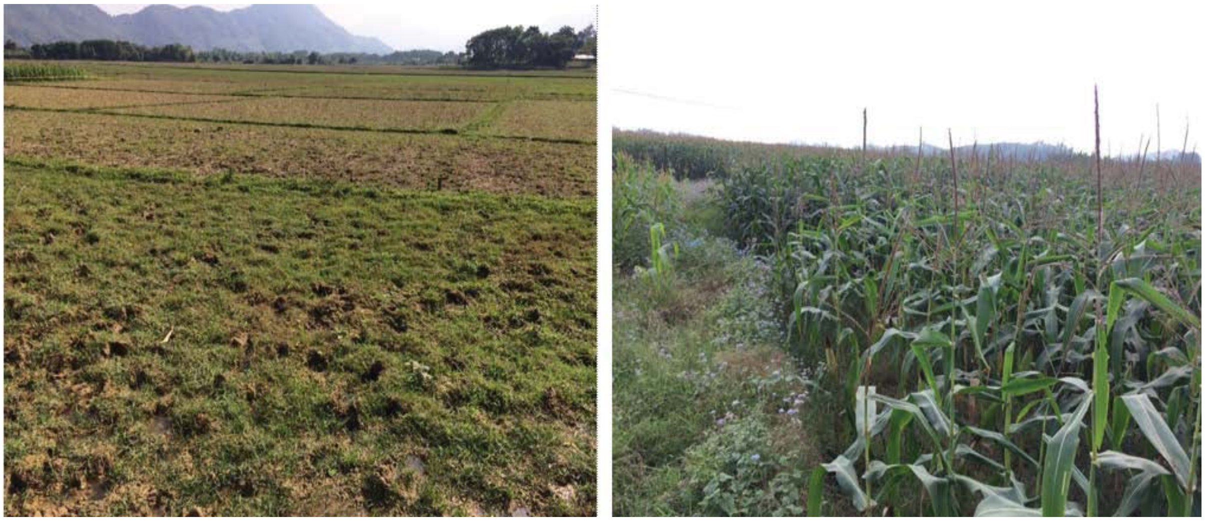


Figure 4.5 Examples of agriculture; harvested rice field (left) and corn fields (right)

❖ Forest land

Thai Nguyen is largely covered by forest, mostly in the north, east and the Tam Dao mountain range. There are still many big trees and dense forests in Tam Dao mountain range, while the other areas have been planted more commercial species (Figure 4.6).

Thai Nguyen has been known as a popular for tea plantation. These are located mostly in Dai Tu district and some less steep areas in the Tam Dao mountain range. Tea bushes (Figure 4.7) in Thai Nguyen are grown in medium size fields, although some small land-holders have cultivated teas intersecting among fruit trees or timber species.

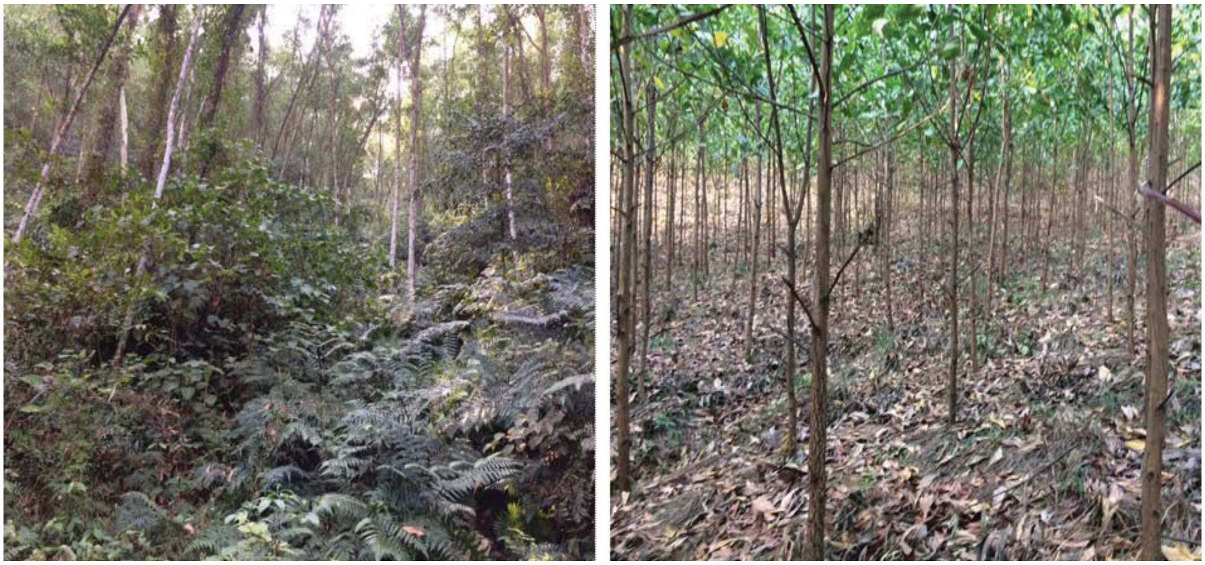


Figure 4.6 Mixed forests (left) and forest plantation (right)

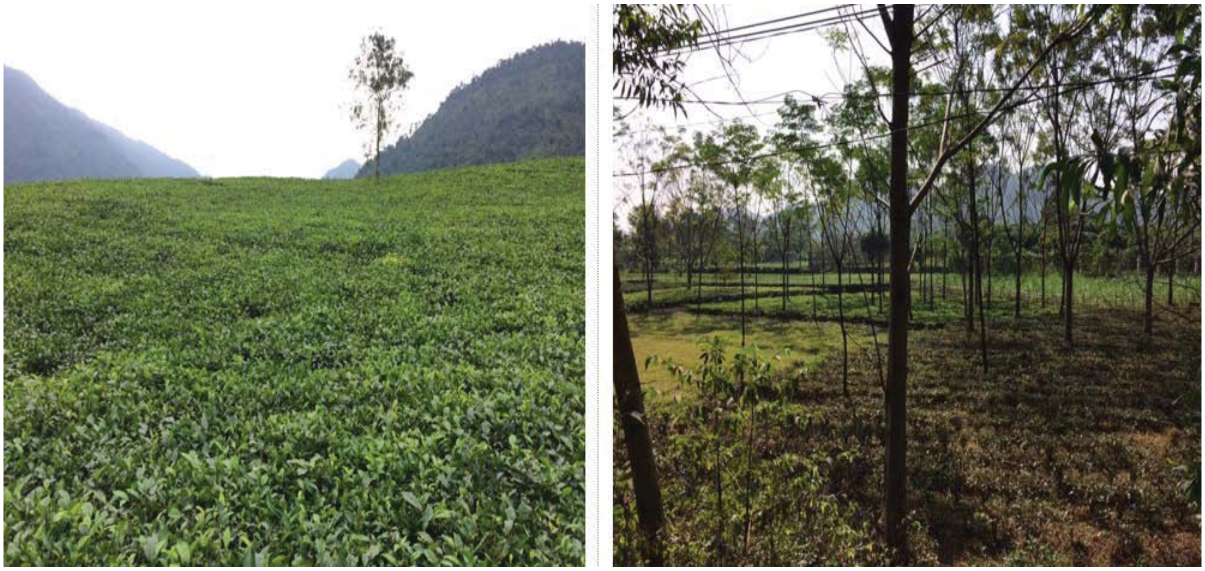


Figure 4.7 Tea plantations (left) and tea intercropping (right), Dai Tu district, Thai Nguyen

❖ Water

Water is identified as lakes, rivers, reservoirs, irrigation canals (Figure 4.8). There are many lakes and rivers in the province, but the significant water bodies are Coc lake and Cau river. Coc lake is the biggest lake in the province and covers a large area of land in the west, while Cau river flows across the province and provides much of the water for industrial and agricultural production.

Although Thai Nguyen province is substantially expanding its urban and industrial development, agricultural and animal farms are dominating in rural areas. Water, therefore, is essential for crop growing, especially rice. Thanks to the Coc lake and Cau river, rice and other crops are annually grown to provide the stable food for more than 1 million people of the province, of which 74% live in rural area.

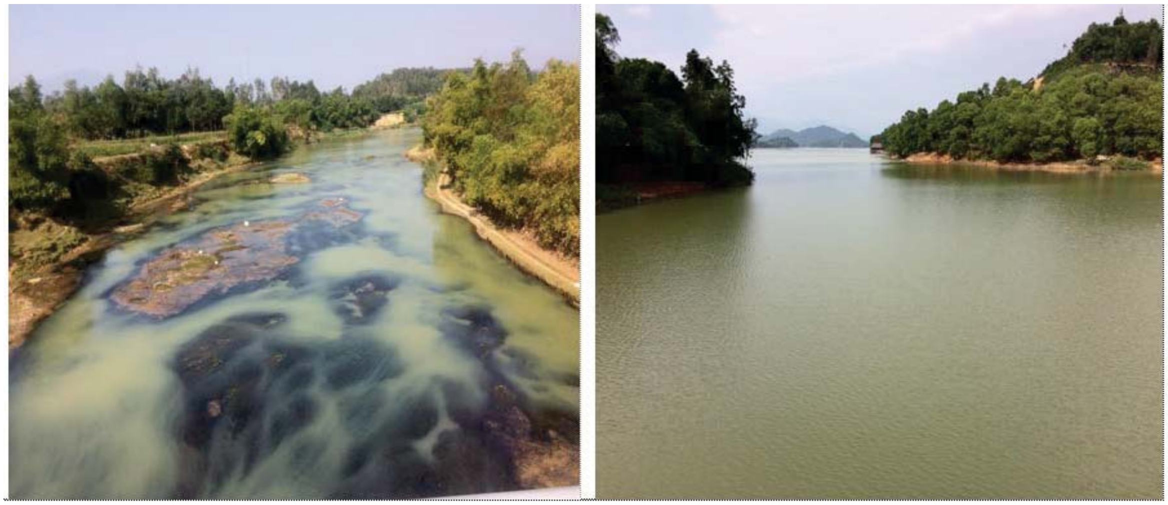


Figure 4.8 A part of Cong River (left) and Coc lake (right)

❖ Extraction

Extraction areas include all land that is used for mining activities (e.g., quarries, coal and iron) and other bare land. There are various active mines in Thai Nguyen with significant reserves. For example, coal mines (Figure 4.9) are located in Dai Tu and Phu Luong districts as well as outlying districts near Thai Nguyen city. Iron ore being mined in Trai Cau, and the Nui Phao mine is the second largest tungsten mine in the world.

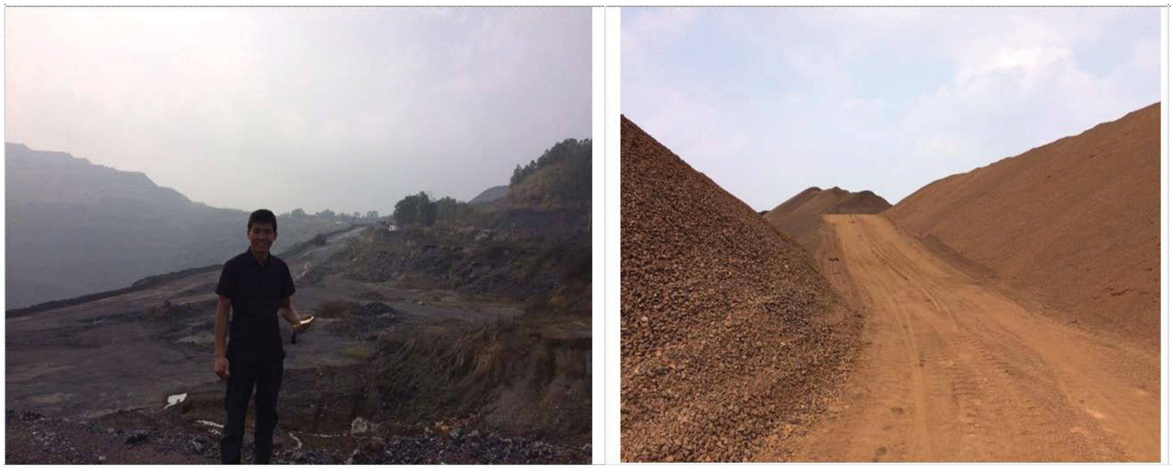


Figure 4.9 Coal mining activities in Dai Tu district (left) and iron exploitation (right) in Dong Hy district, Thai Nguyen province

4.3 Comparison of Land Cover Classifiers

4.3.1 Classification Algorithms

With the rapid development in machine learning algorithms, many supervised, unsupervised and object-based classifiers have been developed and used in the field of remote sensing, particularly for land cover classification. Many studies have been conducted to test the accuracy of parametric and non-parametric classification algorithms in the separation of land cover/land use classes using Landsat data (Dwivedi, Kandrika, & Ramana, 2004; Friedl & Brodley, 1997; Kavzoglu & Colkesen, 2009; Lu et al., 2004; V. F. Rodriguez-Galiano et al., 2012; Rogan, Franklin, & Roberts, 2002), and ASTER (Szuster et al., 2011), but not with both Landsat and Sentinel-2 satellite imagery. In addition, alternative classification algorithms are frequently reported to have a higher overall accuracy, but few studies were concerned with the association of accuracy among those non-parametric classification algorithms. Therefore, this study has compared the accuracy of the three different classification algorithms to classify Landsat 8 and Sentinel-2 data; MLC, ANN and RF classification algorithms.

4.3.1.1 Maximum Likelihood Classifier (MLC)

The maximum likelihood classifier is a statistic-based technique, which is one of the most widely used classifiers for land cover classification (Erbek et al., 2004; Otukei & Blaschke, 2010; Shalaby & Tateishi, 2007). According to Richards and Richards (1999), the algorithm is based on Bayes' theorem to calculate the likelihood of every pixel. The MLC assumes that each class in each band is normally distributed, and calculates the probability distribution if a given pixel belongs to a specific land cover class. A given user can define a threshold at which one pixel is assigned to unclassified if the probability of that pixel is below the threshold. Due to its simplicity and popularity, there are many open-source and commercial software packages supporting this classification algorithm such as ArcGIS and QGIS. I have used ENVI 5.3 software was used to produce classified land cover maps and associated accuracy the confusion matrices.

4.3.1.2 Random Forest (RF) Classifier

Random forests are a collection of decision trees aimed at improving the performance over a single decision tree. Figure 4.10 is an example of a decision tree model derived from Sentinel-2 data using the "rpart" library (Therneau, Atkinson, & Ripley, 2010) in R programming. A random forest is constructed by generating decision trees for subsamples of the original data. This process is known as a "random sampling with replacement or bagging" approach, which was developed by Leo Breiman (2001) and widely used in the statistical world today. With the bagging method, a single decision tree will behave differently from a random forest with a single tree. This means that the number of observations of each band in each class may not be always equal, and therefore some decision trees may perform better than others. However, on average they will produce a fairly unbiased and stable model, and this is also an advantage of the RF model.

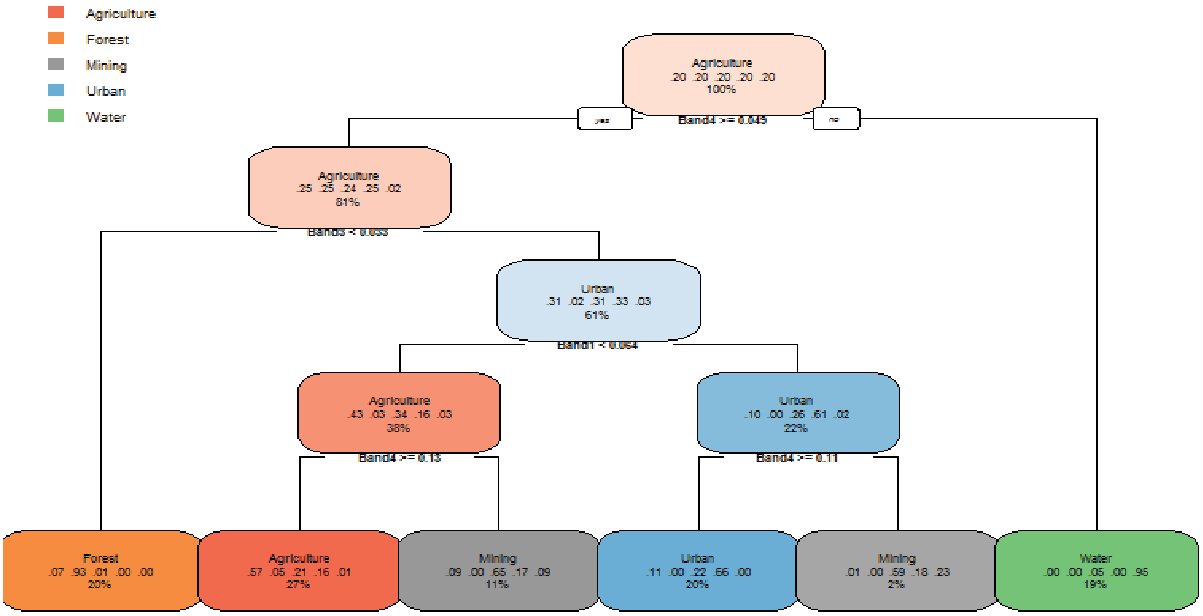


Figure 4.10 An example of decision tree

Random forest can contain an arbitrary number of *N* trees, where *N* is the number of trees to be grown. However, (Leo Breiman, Cutler, Liaw, & Wiener, 2011) advised that 500 trees is preferred to train the model. In this study, 100 trees were chosen based on number of trials and the graphical plot of effect of tree size to the model (in Section 5.1). Similarly, the number of variables is a user-defined parameter, and often set to the square root of number of inputs for selecting "best split" at each node (Leo Breiman, 2001). Interestingly, the algorithm is not sensitive to it, and the randomly selected number of variables substantially minimized the correlation between trees (Gislason et al., 2006) because each tree only uses a portion of the input variables in a random forest. This also reduced considerably the algorithm's computational intensity.

The analysis of the RF shows that its computational time is $cT\sqrt{M}\,\mathrm{N}\log(\mathrm{N})$ where c is a constant, T is the number of trees in the model, M is the number of variables and N is the number of observations in the dataset (L Breiman, 2003; Gislason et al., 2006). Noticeably, the

RFs are not computationally intensive, but they require a fair amount of memory as they store an N by T matrix in memory (Gislason et al., 2006). One advantage of the RF is that it can test itself after growing each tree. This means that approximately 2/3 data is used to train the model, while the remaining one is used to test the model, and its test result is known as "out of bag error". In addition, the RF also has capability to determine the importance of a variable *m*th. This can be estimated by randomly permuting all the values of the *m*th variable in the out of bag samples for each classifier. If an increased out of bag error is produced, that is an indicator of the importance of that variable (Gislason et al., 2006).

A "randomForest" package (Leo Breiman et al., 2011) in R programming language was used to train the model and derive variable importance indicators, effect of tree size and other information.

4.3.1.3 Artificial Neural Network (ANN)

ANNs are explained as a collection of nodes with lines (synapses) connecting to each other (Figure 4.11). The organization of ANNs is split in three main groups: one input layer, hidden layers of which the number may be small or large, and an output layer. To train the model, an initial set of randomly selected weights will be fed in the input and all of its weights will be calibrated by repeating two commonly used processes, forward and back propagation, to produce the output. In other words, the neural network repeats both forward propagation and back-propagation until the weights are accurately corrected to produce the output, which is comparable to the actual output.

According to Kavzoglu and Mather (2003), the success of the model is pretty much dependent on the choice of network parameters such as number of hidden layers, number of training iterations and minimum output activation threshold. In this study, all optimal parameters were selected to train the model based on a number of different trials. Especially, a total of 1000 iterations was established, while only one hidden layer was necessary. ANN algorithm was available in ENVI 5.3 software for this model to derive land cover classified maps and the associated accuracy confusion matrix

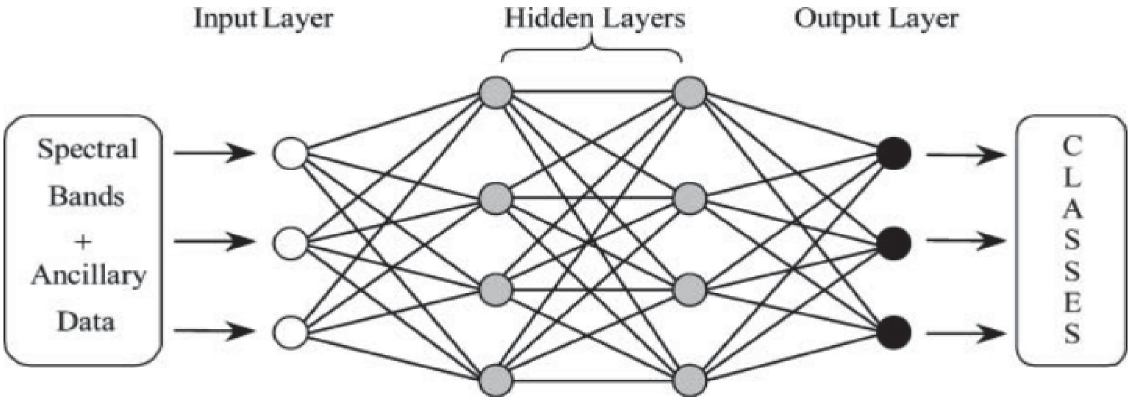


Figure 4.11: The operational procedure of feed-forward neural network

Source: International Journal of Remote Sensing (Kavzoglu & Mather, 2003)

4.3.2 Sentinel-2 and Landsat 8 Accuracy Assessment

Accuracy assessment was carried out using the confusion matrix method because of its simplicity and popularity, as discussed in Section 2.4.2. A total of 169 ground control points was collected from the fieldtrip, these were used in the accuracy assessment for Sentinel-2 and Landsat 8 classifications (Figure 4.12). Overall accuracy, users' and producers' accuracies and the Kappa statistic were then produced from the confusion matrix. Overall accuracy (OA) for a particular classified image was calculated by summing the number of correctly classified pixels and dividing the total number of pixels are located along the upper-left to lower-right diagonal of the confusion matrix (Story & Congalton, 1986).

$$OA = \frac{\sum_{i}^{p} N_i}{T}$$

Where p is the number of classes, N_i is the sum of correctly classified pixels, and T is the total number of reference pixels. The kappa coefficient (k) measures the agreement between classified map and reference values. A kappa value of +1 represents perfect agreement, while a value of 0 represents no agreement. The kappa coefficient is computed as follows:

$$k = 1 - \frac{1 - p_0}{1 - p_e}$$

Where p_0 represents the proportion of correct agreement in the test dataset and p_e is the proportion of agreement that is expected by chance. As discussed in Section 2.4.2, the confusion matrix also has some fundamental limitations. Therefore, the latest land use statistics and inventory (December 2016) from the Thai Nguyen department of natural resources and environment (TNMT) were used as reference statistics for further accuracy evaluation of the classified images (Thai Nguyen Department of Natural Resource and Environment, 2016a).

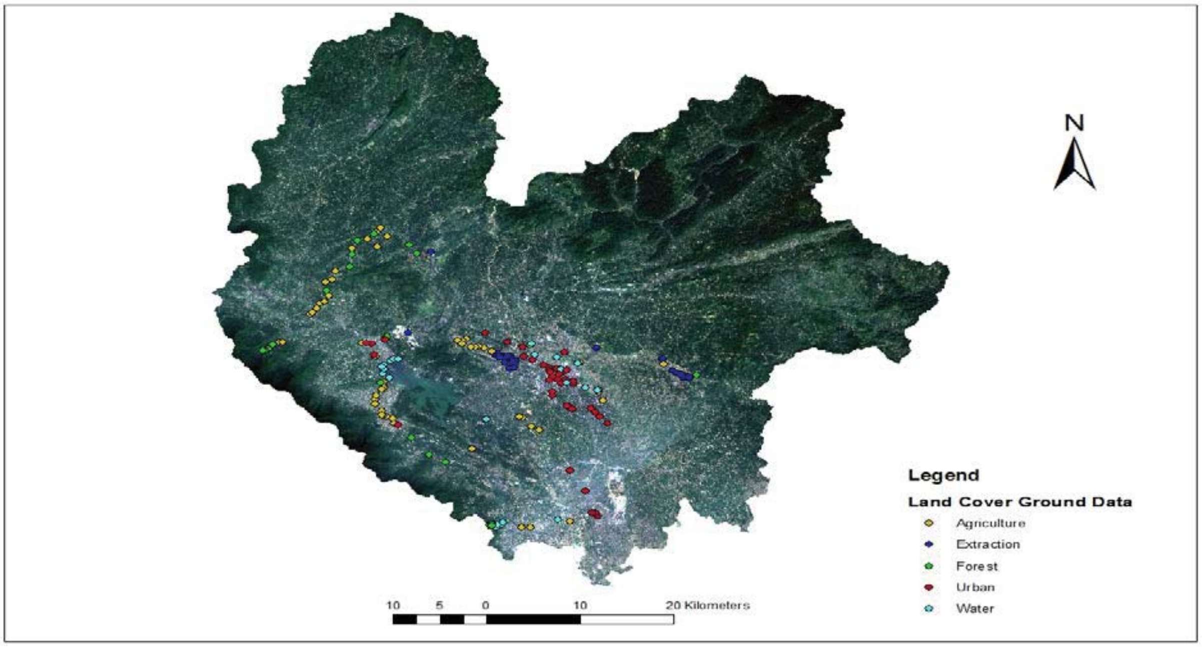


Figure 4.12 Reference data for assessing the Sentinel-2 and Landsat 8 classification

The TNMT data was constructed from traditional surveys and historical land use records. The local government has classified this data into many subclasses such as residential urban land

and residential rural land, while this study used only five land cover/land-use classes. Therefore, it was important to redefine and characterize the TNMT classification system to ensure that it was similar and compatible with this study's classification scheme. Based on characteristics as discussed in Section 4.2, five land cover/land use classes were built from the TNMT classification system (Appendix 1). As an example, residential, commercial and rural land uses were clustered into built-up/urban area, while annual crops, rice and short-rotation crops were grouped into the agriculture class. It was noted that the local land use classification system had no relevant mining class, and therefore the size of the mining area may not accurately represent the amount derived from this comparison. However, the TNMT statistic was generally considered unbiased, accurate and representative for land cover/land use types, and it was used with confidence, as reference statistics to compare with area statistics derived from Sentinel-2 and Landsat 8 data.

4.4 Land Cover Monitoring and Mapping for Thai Nguyen

Land cover/land-use monitoring and mapping is extremely important as it provides timely and accurate information for local planners and politicians to evaluate economic benefits and appreciate environmental aspects, particularly in the context of rapid growth of urbanization and industrialization in developing world. The RF classifier was used in this study to monitor and map temporal land cover/land-use changes between 2000 and 2016 due to its high overall accuracy, non-parametric nature and stability, which will be further discussed in Section 5.1. The use of the RF algorithm has demonstrated superior characteristics over conventional classification approaches, while it is also less intensive computationally and time-consuming than the ANN classifier.

4.4.1 Landsat 7 and 8 Classification Accuracy Assessment

The same reference data described in Section 4.3.2 was used for assessing the accuracy of land cover product derived from Landsat 8 data, while nearly 1000 reference data points were recorded based on Google Earth's high resolution imagery for assessing the accuracy of the Landsat 7 derived land cover map (Figure 4.13). While five land cover/land use classes were built from the local classification system for assessing three classification algorithms as discussed in Section 4.2, this section constructed only four land cover/land-use types (agriculture, urban/built-up, forest and water). The mining extraction class was merged with urban area to form a new class called urban/built-up class. This effort has provided a clear relationship between reference area statistics and Landsat-8 derived area statistics, which can be seen in Section 5.2.1.

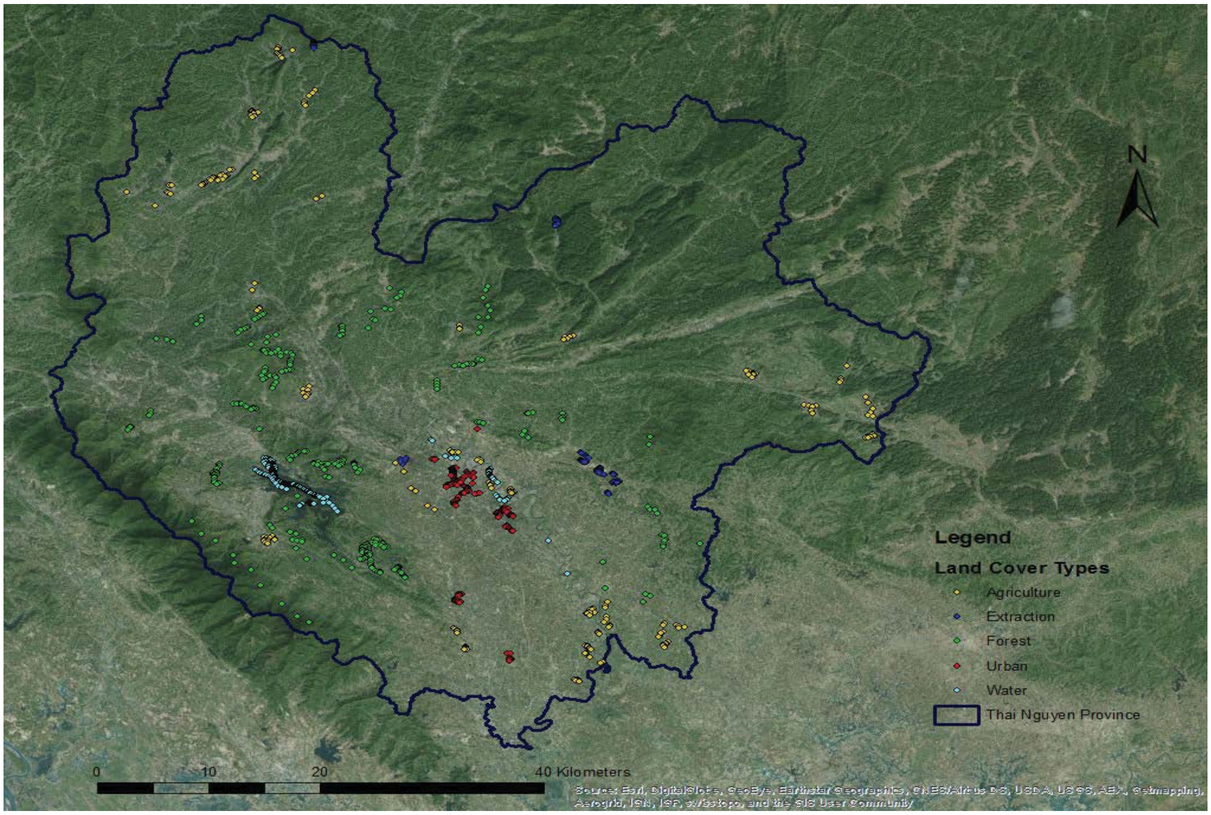

N
Legend
Land Cover Types
Agriculture
Extraction
Forest
Urban
Water
Thai Nguyen Province
0
10
20
40 Kilometers
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 4.13 Reference data for assess the accuracy of Landsat 7 (collected based on Google Earth and ESRI Imagery 2000)

The accuracy assessment assumed the sample data points selected are accurate and representative for the maps being evaluated. This study produced error matrices to show the contingency table with each pixel truly belongs (columns) on the map unit to which it is allocated by selected analysis (rows). Overall accuracy, users' and producer's accuracies, and Kappa statistic were then derived from the confusion matrix.

4.4.2 Change Detection

In this study, a post-classification change detection method was employed due to its advantages discussed in Section 2.4.1. Derived image pairs of two dates were compared using cross-tabulation in order to determine the nature of change and quantity of change between 2000 and 2016. In other words, this method provides a "from-to" change information, and identifies where such change has occurred, and how much has occurred (Stow, Tinney, & Estes, 1980). Post-classification comparison can minimize the problem of normalizing for atmospheric and sensor differences between two dates (Singh, 1989). ENVI 5.3 software was used to produce a change matrix and display spatial distribution of gain and loss of each land cover/land-use type for the same period.

5 RESULTS AND DISCUSSION

In this study, the Sentinel-2 and Landsat 8 data were classified into five land cover/land-use categories using the three different classification algorithms, namely MLC, ANN and RF classifiers. Each classification algorithm produced relatively different results both classification accuracy and area statistics. While the RF classification algorithm outperformed the other two classification algorithms (ANN and MLC) for both Landsat 8 and Sentinel-2 data, the MLC algorithm demonstrated a high overall accuracy for Landsat 8. The ANN algorithm, however, was not as accurate as the other two for both datasets.

The RF algorithm was used to classify Landsat 7 and Landsat 8 data to monitor and map land use/land cover changes in Thai Nguyen province, Vietnam between 2000 and 2016; it had high accuracy stability and was less time-consuming. The following sections demonstrate two of the above mentioned problems.

5.1 Comparison of Classifiers

5.1.1 Land Cover Maps Derived from Sentinel-2

Land cover classified maps for the 10-m spatial resolution Sentinel-2 data were produced using the MLC, ANN and RF classification algorithms (Figure 5.1). Overall and individual accuracy statistics for each classification technique were derived using a confusion matrix to analyse the performance of each classifier; these are summarized in Table 5.1. Area statistics for five different land cover/land-use categories derived from Sentinel-2 data using the three classification techniques were calculated and are presented in Table 5.2, while the TNMT statistics were summarised and clustered into five land cover characteristics, as discussed in Section 4.2, for further comparison.

In this comparison, overall accuracies for the MLC, ANN and RF classifiers were relatively low with 82.25% (kappa 0.77), 81.66% (kappa 0.76) and 89.94% (kappa 0.87) respectively. However, the RF algorithm showed the best performance over the two remaining classification techniques with the highest overall accuracy. Unexpectedly, while the ANN classification is usually reported to have a high overall accuracy and outperforms traditional parametric classification techniques (Civco, 1993; G. Hepner, Logan, Ritter, & Bryant, 1990; Szuster et al., 2011; Tayyebi, Pijanowski, Linderman, & Gratton, 2014), this approach turned out to be least accurate among the three classification algorithms in this comparison.

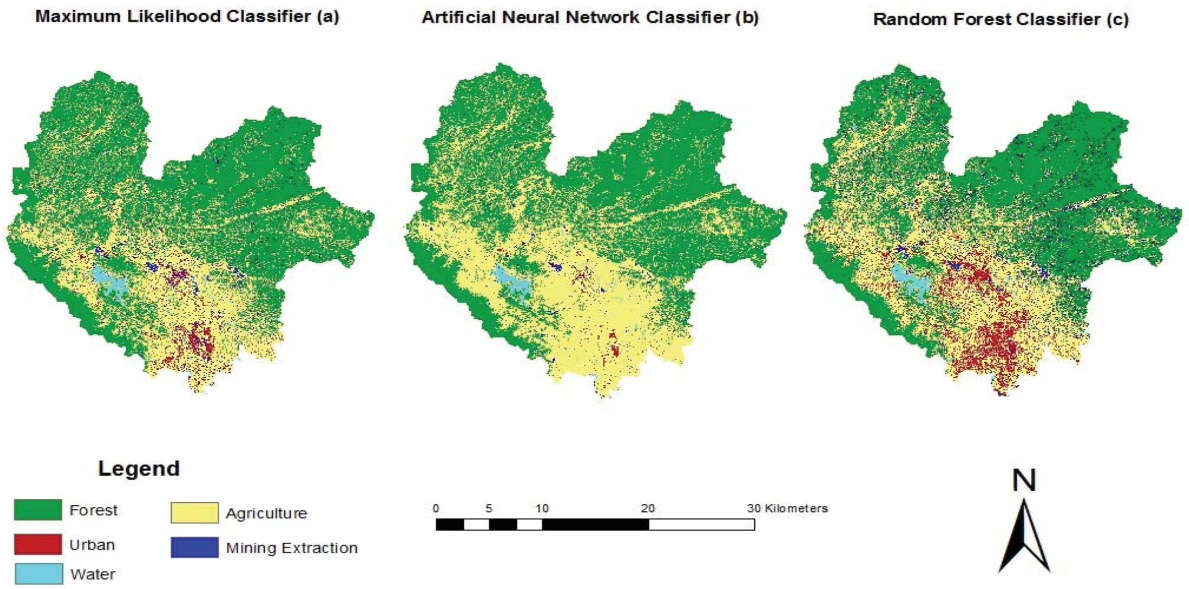

Maximum Likelihood Classifier (a)
Artificial Neural Network Classifier (b)
Random Forest Classifier (c)
Legend
Forest
Urban
Water
Agriculture
Mining Extraction
0
5
10
20
30 Kilometers
N

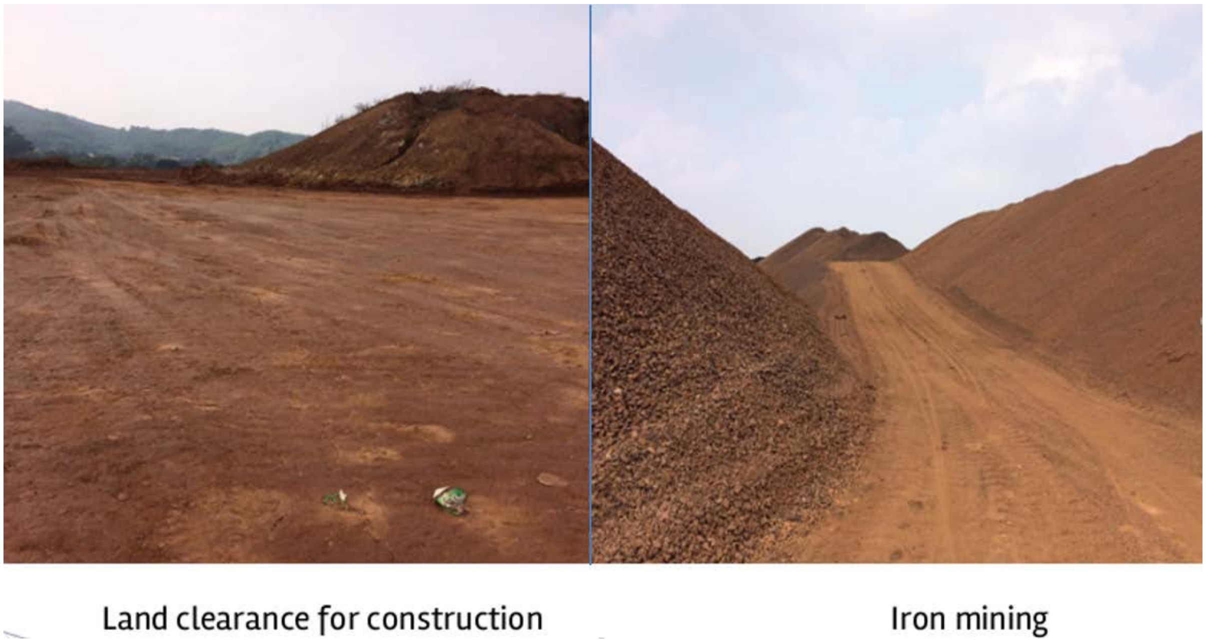
Figure 5.1 Land cover/land-use maps derived from Sentinel-2 data using MLC, ANN and RF algorithms

The low overall accuracies for all three classification algorithms using Sentinel-2 data are believed to be due to a heterogeneous landscape and the intersection of various land-use types in one field (Figure 4.7). In contrast these classification algorithms all produced high overall accuracy for Landsat 8 data, which will be seen in Section 5.1.2. While only 10-m resolution Sentinel-2 bands were used in this comparison, it was thought that there may some effects of spatial and spectral information on the overall accuracy of the Sentinel-2 derived maps. A study was carried out that included 8 Sentinel-2 bands (20-metre spatial resolution) in the hope of increasing overall accuracy, but it turned out to be not significantly different from the overall accuracy of classified images derived from 10-m Sentinel-2 bands (Appendix 2). Therefore, this study excluded this comparison, and focused on 10-m Sentinel-2 bands instead. After testing 20-m Sentinel-2 bands, it seemed that the finer spatial resolution of Sentinel-2 data together with heterogeneous landscapes were likely to affect the overall accuracy of the derived maps in this comparison.

Table 5.1 Summary of accuracy statistics for three classifiers using Sentinel-2 data (%)

Land covers	MLC		ANN		RF	
	Producer's	User's	Producer's	User's	Producer's	User's
Agriculture	94.44	87.93	96.30	69.33	90.74	84.48
Extraction	88.89	57.14	70.37	82.61	88.88	92.30
Forest	87.96	100.0	86.96	95.24	78.26	100.0
Urban	53.66	81.48	63.48	89.66	95.12	86.66
Water	91.67	100.0	87.50	100.0	91.66	100.0
Overall Accuracy	82.25		81.66		89.94	
Kappa Statistic	0.77		0.76		0.87	

Although overall accuracies of each classifier in this comparison were not very high, the individual class accuracies revealed interesting insights in the performance of each classification technique in the separation of certain individual land cover categories. The RF, for example, was separated the forest class poorly, but exhibited relatively accurate overall results. Agriculture, urban and water classes were well classified by the RF classifier with above 90% producer’s accuracy, while the separation of mining extraction was slightly less effective with 24 out of 27 pixels correctly classified (Appendex3). By contrast, the MLC and ANN classifiers produced similarly low overall accuracies, and classified less effectively in nearly every class, except for agriculture. While agriculture was classified accurately in both MLC and ANN classifiers at more than 94%, only 24 out of 41 pixels were classified correctly as urban and most misclassified pixels were in the forest and mining extraction classes for the ANN and MLC algorithms respectively (Appendix 3). This was expected and understandable in the context of rural land cover/land-use as forest and built-up areas are not clearly separated on the ground. Additionally, the removal of vegetation on the ground for constructing industrial, urban zones and mining activities make it challenging to distinguish between urban and mining classes due to its relatively similar surface (Figure 5.2) and confused spectral signatures on the image.



Land clearance for construction

Iron mining

Figure 5.2 The similarity in mining extraction and urban feature surface of the study area

The overall accuracy for land cover/land-use maps derived from remote sensing data is expected to be at least 85% (Anderson, 1976; Thomlinson et al., 1999) and preferably 90% (Lins & Kleckner, 1996). It is obvious that two out of the three classification algorithms used were unsatisfactory. However, it is also important to examine the area statistics of each classification technique derived in comparison to the locally measured land use statistics. General speaking, the statistics of areas of land cover maps derived from Sentinel-2 indicated that agriculture and forest are over the dominant covers in the study area, followed by urban and mining extraction. But each land cover classification algorithm produced slightly different areas for each individual class and total area for Thai Nguyen province (Table 5.2). For example, urban area derived from ANN and MLC was about 22.4 km^2 (0.6%) and 104.5 km^2 (3%) respectively, while that of RF was much larger with 268.02 km^2 (7.6%). The ANN classifier classified agriculture with an area of approximately 1530.1 km^2 (43.46%), while the other two algorithms found this class occupied an area of 1242 km^2 (MLC) and 1224 km^2 (RF). Clearly, the urban area derived from the RF classifier (Figure 5.1c) was visibly larger than that of the MLC and ANN algorithms (Figure 5.1a and 5.1b), while the area of agriculture in Figure 5.1b was largest in the ANN classification.

Table 5.2 Summary of area statistics for the three classifiers using Sentinel-2 data

Land Covers	MLC		ANN		RF	
	Area (km^2)	*Percent*	*Area (km^2)*	*Percent*	*Area(km^2)*	*Percent*
Agriculture	1242.4	35.29	1530.12	43.46	1223.59	34.75
Urban/Built-up	104.51	3.97	22.36	0.64	268.02	7.61
Mining	117.07	3.33	45.32	1.28	231.72	6.58
Forest	2001.86	56.86	1887.12	53.60	1730.63	49.15
Water	54.97	1.56	35.88	1.01	67.08	1.91
Total	3520.87		3520.80		3521.0	

Although each land cover classification algorithm classified Sentinel-2 data differently, there were some similarities and distinctions in area statistics when they were compared to TNMT statistics (Thai Nguyen Department of Natural Resource and Environment, 2016b). Figure 5.3 shows that the land use area statistics provided by the local government for urban and water classes were larger than that of land cover products derived from Sentinel-2 data, but areas of mining and agriculture were smaller than those derived from the three classification algorithms. The ANN classifier provided an area of mining extraction that is nearly the same as that reported by the local government, but there is a substantial difference in urban area. The RF classifier, on the other hand, overestimated the area of mining. It is also important to note that there is a relatively large difference between TNMT urban area and classification-based urban area. This can be due to the fact that the local government statistics included planned urban area, which may not happen on the ground. Although the RF classifier was not perfectly matched with all individual reference statistics, it was likely to be more representative of the TNMT statistics (Figure 5.3).

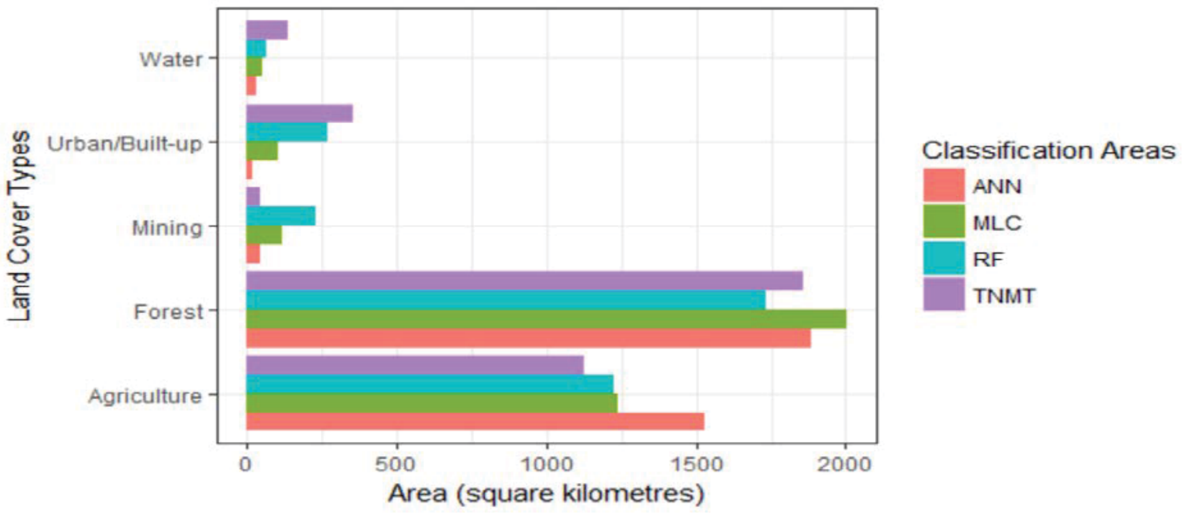


Figure 5.3 The comparison of area statistics reported between classification algorithms for Sentinel-2 data for the study area

The comparison of accuracy and area statistics indicates that the MLC and ANN algorithms have major advantages for the agricultural, mining and forest land classification, although there were substantial omission errors associated with these classes, except for forest. A significant difference between the MLC, ANN and RF algorithms is that RF had a higher overall accuracy and lower commission errors in nearly every individual class. While the MLC and ANN algorithms performed poorly on urban land classification, the RF classifier was much better for urban land classification.

The RF algorithm also demonstrated the capability of analysing the importance of variables and the effect of trees. In this comparison, the RF classifier identified that 100 trees would be adequate for the model (Figure 5.4), while Sentinel-2 band 8 (near infrared band) was identified as the variable that contributed most to the model. Overall, the comparison suggested that the RF algorithm should be favoured among the three classification algorithms for subtropical land cover classification using Sentinel-2 data.

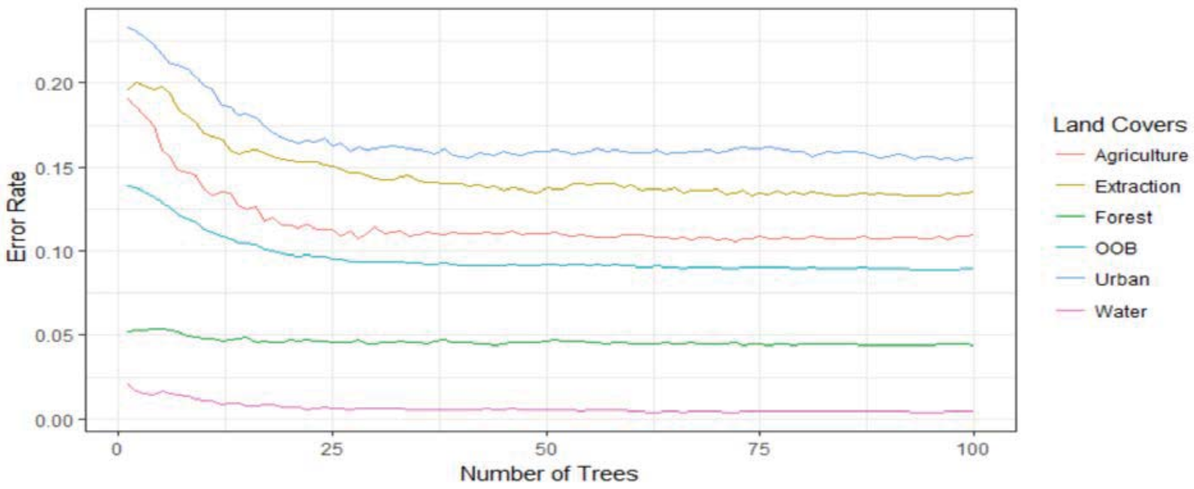


Figure 5.4 The effect of tree sizes on the RF model accuracy using Sentinel-2 data

5.1.2 Land Cover Maps Derived from Landsat 8

Three different land cover maps (Figure 5.5) were derived from the 30-m spatial resolution Landsat 8 image using the MLC, ANN and RF supervised classification algorithms. Accuracy assessment statistics for each classification technique were produced by adopting the confusion matrix approach to analyse the performance of each classifier. This was summarized in Table 5.3, while area statistics for each classification algorithm were summarised in Table 5.4. In this comparison, the overall accuracy for MLC, ANN and RF algorithms was 90.53% (kappa 0.88), 84.02% (kappa 0.79) and 94.10% (kappa 0.92) respectively. Although ANN had the lowest overall accuracy among three classifiers, all these overall accuracies were considered relatively high, particularly for MLC and RF classifiers with 90.53% and 94.10% respectively.

The difference in overall accuracy for the land cover maps derived from Landsat 8 is likely due to the moderate spatial resolution of the image and the underlying analysis for each algorithm. Among the three classification algorithms, the ANN produced the lowest overall accuracy with 84.02% and high error rates for mining extraction, forest and water. Commission and omission error rates of individual land cover/land-use classes for the ANN classifier were higher than that of the MLC and RF algorithms. As an example, commission error rate for agriculture by the MLC was 7.41%, while that of the ANN classifier was 22.86%. This high commission error could have contributed to the perfect producers' accuracy for urban and agriculture classes (Table 5.3). While the ANN has been found to have higher overall and individual accuracies than the MLC (J. D. Paola & R. A. Schowengerdt, 1995), this study revealed the opposite.

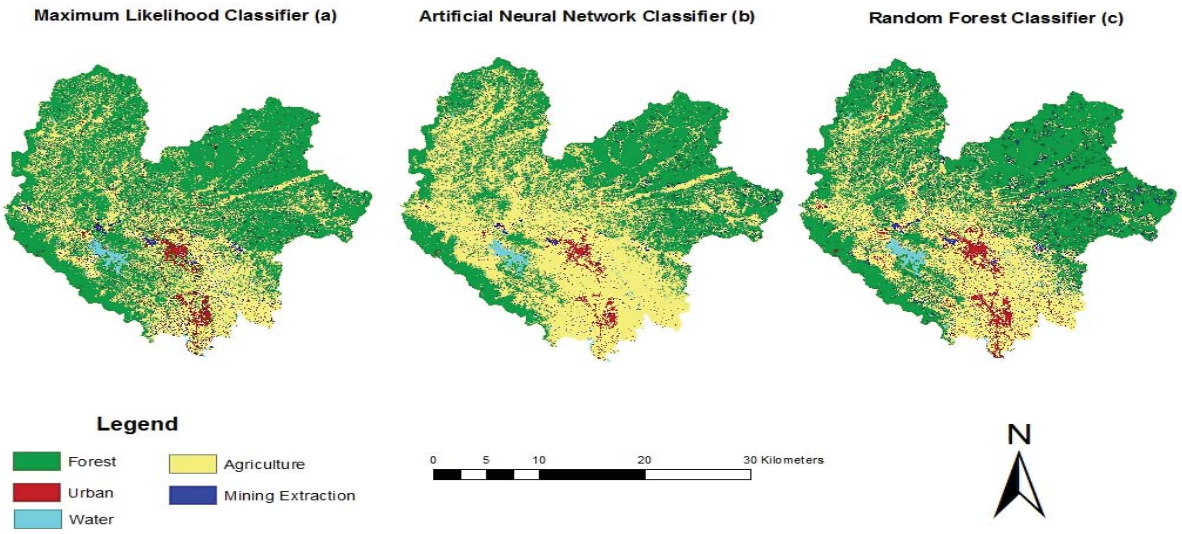

Maximum Likelihood Classifier (a)
Artificial Neural Network Classifier (b)
Random Forest Classifier (c)
Legend
🟩 Forest
🟥 Urban
🟦 Water
🟨 Agriculture
🟦 Mining Extraction
0
5
10
20
30 Kilometers
N

Figure 5.5 Land cover/land-use maps derived from Landsat 8 data using MLC, ANN and RF algorithms

The results of MLC showed that this conventional classification technique was relatively accurate in the separation of land covers using Landsat 8 data. While the MLC had a higher overall accuracy than the ANN, it was much lower than the RF classification. But the MLC was the best classifier for the separation of forest, and therefore would appear to be well suited for forest land classification. In addition, the MLC also separated agriculture, mining extraction and urban with minimum errors. Agriculture, for example, was correctly classified for 92.59% of the test pixels while urban class was correctly classified for 92.68% of test pixels. The water class had only 83.33% of pixels classified correctly and had the lowest individual accuracy when using the MLC algorithm.

Table 5.3 Summary of accuracy statistics for the three classifiers using Landsat 8 data (%)

Land covers	MLC		ANN		RF	
	Producer's	User's	Producer's	User's	Producer's	User's
Agriculture	92.59	92.59	100	77.14	98.15	92.98
Extraction	92.59	75.76	62.69	80.95	92.59	89.29
Forest	86.96	95.24	60.87	100	78.26	100.0
Urban	92.68	92.68	100.0	85.42	100.0	93.18
Water	83.33	100.0	66.67	100.0	91.66	100
Overall Accuracy	90.53		84.02		94.10	
Kappa Statistic	0.88		0.79		0.92	

The RF classifier produced the highest overall accuracy with an improvement of about 4% and 10% over the MLC and ANN algorithms respectively. The RF algorithm produced minimal commission and omission errors for most land cover classes, namely agriculture, urban, mining extraction and water. As an example, 22 out of 24 test pixels were classified correctly as water, while approximately 66.7% and 83.0% of test pixels were classified correctly as water with the ANN and MLC algorithms respectively. In addition, this approach separated the urban/built-up area from other land uses more effectively although there were substantial mixtures of spectral signatures between mining extraction class and urban features. The RF algorithm did not produce a high individual accuracy for the forest class like the ANN algorithm did. But the comparison of the local areas statistics and derived statistics for this land cover category revealed that there was a good match between local government-based statistics and classification-based area (Figure 5.6).

Figure 5.6 shows that the percentage of areas of each individual land cover category derived from Landsat 8 data using the RF algorithm was much closer to the locally acquired land use statistics (Appendix 1). For example, the percentage of forest land was nearly identical with statistics reported by the local government, while there was no significant difference in the percentage of agricultural area between actual statistics and the RF-based classified map.

Similarly, the MLC algorithm also produced similar statistics in relation to TNMT land use areas for each individual land cover type, except for forestland (Figure 5.6). By contrast, the ANN algorithm showed a substantial difference between the TNMT areas and classification-based areas. Notably, the TNMT percentage of agriculture was 35% of total area, while this approach classified nearly 50%. The ANN algorithm was likely to overestimate agriculture, while other land uses were usually underestimated (Figure 5.6), except for the mining extraction.

Table 5.4 Summary of area statistics for the MLC, ANN and RF algorithms using Landsat 8 data

Land Covers	MLC		ANN		RF	
	Area (km^2)	Percent	Area (km^2)	Percent	Area (km^2)	Percent
Agriculture	1182.56	35.59	1756.61	49.9	1240.92	35.24
Urban/Built-up	116.56	3.31	81.89	2.3	160.71	4.56
Mining	127.82	3.63	83.10	2.4	211.07	5.60
Forest	2055.27	58.37	1561.70	44.5	1848.67	52.50
Water	38.86	1.10	37.79	1.1	59.66	1.69
Total	3521.07		3521.09		3521.03	

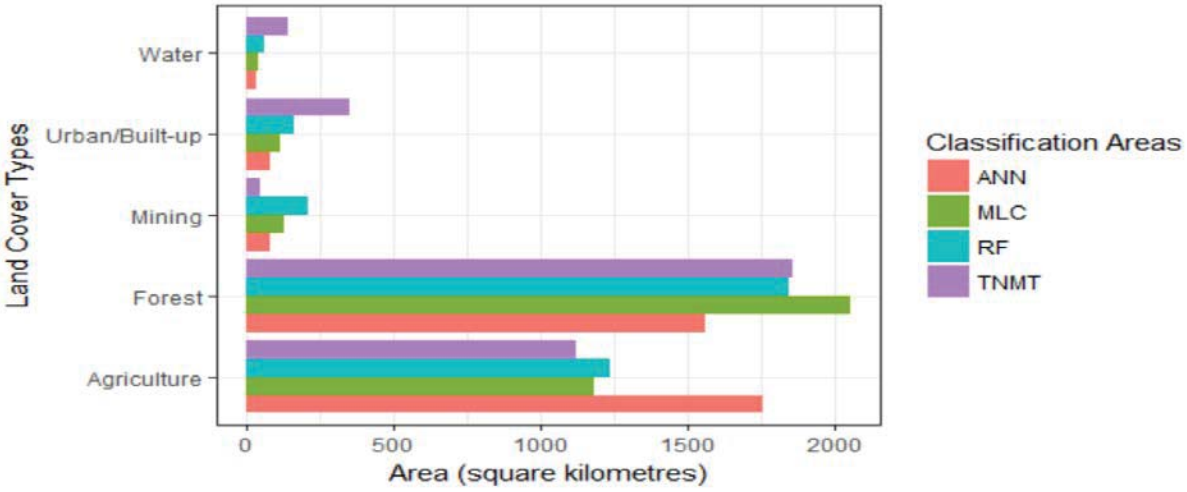


Figure 5.6 The comparison of area statistics between algorithm-based and TNTM statistics

With regard to the area statistics for each land cover/land-use type, Figure 5.7 showed some similarities and differences between the three classification algorithms. The area of agriculture classified from Landsat 8 data using the RF and MLC algorithms was relatively similar with 1200 km^2, while the ANN classifier estimated a much larger area of 1756 km^2. The RF algorithm produced a larger urban area of 160 km^2 (4.56%), whereas only 81.89 km^2 (2.33%) was classified as urban using the ANN algorithm. For Sentinel-2 data, similar comparisons revealed that the pattern of land cover/land-use classification using the MLC and RF algorithms was similar in individual area statistics to the Landsat 8 image, but there were substantial fluctuations in area for the ANN algorithm between Landsat 8 and Sentinel-2 data, especially for forest. It is, nevertheless, interesting to note that the surface area of water derived from both Sentinel-2 and Landsat 8 data using the MLC, ANN and RF algorithms was similar, ranging from 37 km^2 to 67 km^2.

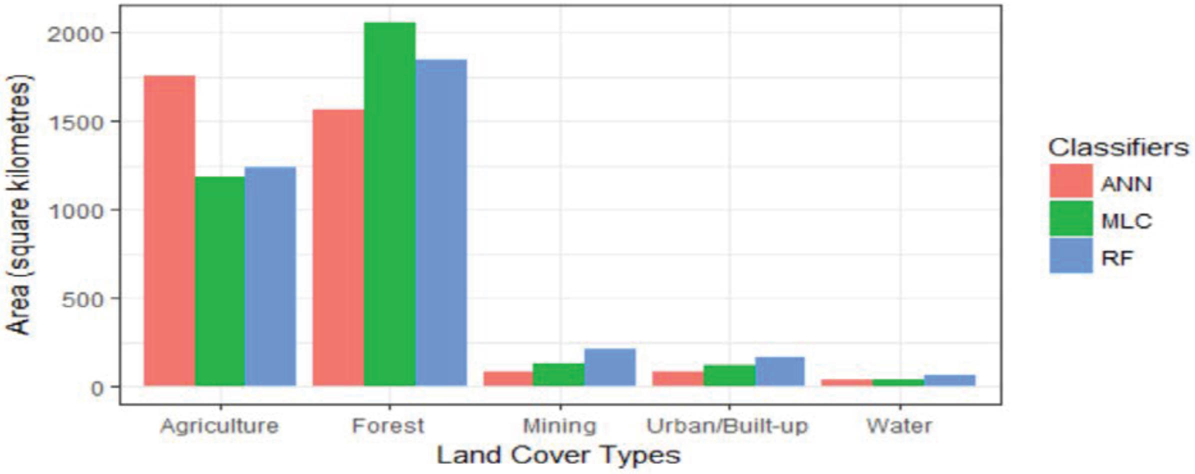


Figure 5.7 The comparison of area statistics reported between classification algorithms using Landsat 8

The comparison of accuracy and area statistics indicates that the RF and MLC algorithms have major advantages for classifying moderate spatial resolution Landsat data. For instance, the MLC algorithm outperformed the ANN classifier with respect to the separation of water, mining extraction and forest classes, while the RF algorithm well classified nearly every land cover category correctly, except for forest. As discussed in Section 5.1.1, the RF algorithm also has a major advantage for analysing the importance of variables and the effect of trees. In this comparison, the RF classifier identified that 100 trees would be adequate for the model (Figure 5.8), while Landsat-8 bands 1(Blue band) and 6 (Short-wave Infrared band) were determined as the variables that contributed the most to the model (Figure 5.9). The Landsat 8 blue band had the most significant importance because it enables the classification of those land cover types with a seasonal behaviour (V. Rodriguez-Galiano et al., 2012). It should be noted that the Landsat 8 used in this study was collected in early Winter when forests start to fall its leaves. This explains the importance of blue band in distinguishing soil, built-up and mining from vegetation.

After analysing the accuracy and area statistics for both Sentinel-2 and Landsat 8 data, this study found that the RF algorithm should be favoured among the three classification algorithms for subtropical land cover classification regardless of the type of imagery.

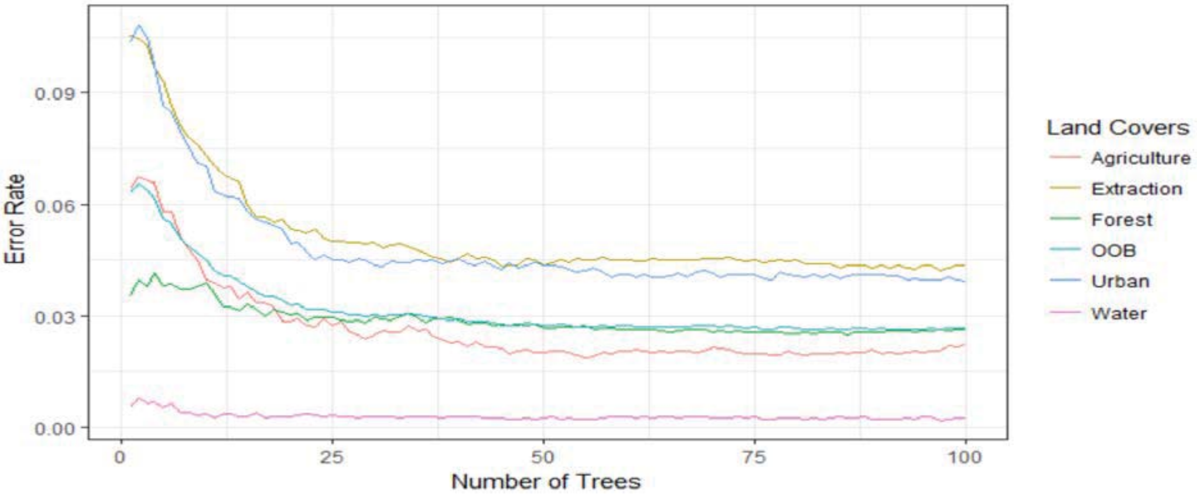


Figure 5.8 The effect of tree sizes on the model accuracy using Landsat 8

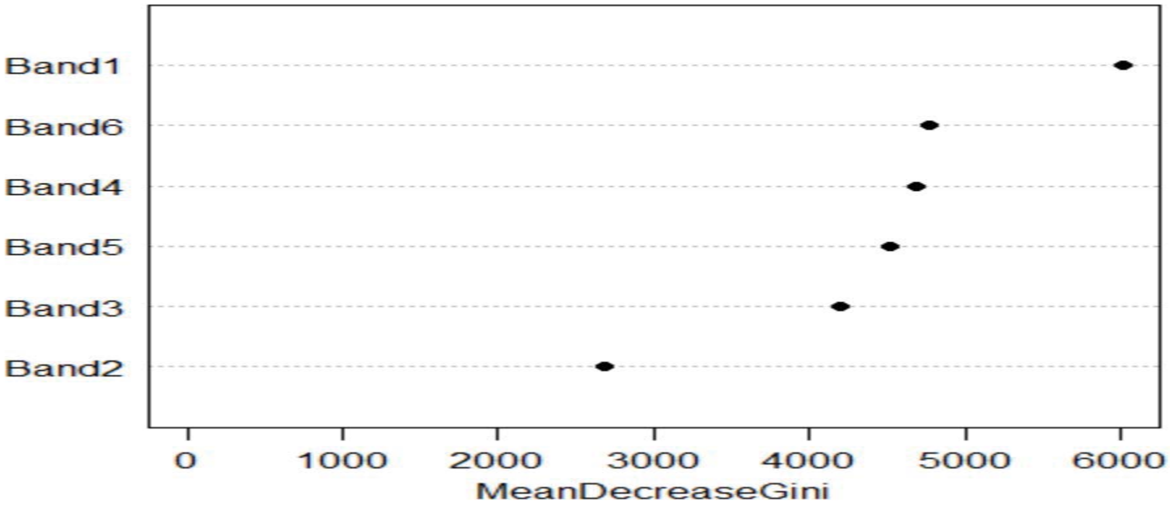


Figure 5.9 The variable importance identified by the RF model for Landsat 8 bands

5.2 Temporal Land Cover Change Monitoring

5.2.1 Land cover/land use change

Land cover maps were produced for two years (Figure 5.10), 2000 and 2016 respectively, and the statistics of individual class areas and changes between years were derived and summarised in Table 5.5. Overall, there was a substantial increase in the urban and mining areas from 2000 to 2016, while the forest land decreased significantly over the same period and less agricultural land was also mapped.

Over the past 16 years, the study area witnessed a relatively substantial change in land cover/land-use, and these changes occurred largely in Thai Nguyen central city and southern regions (Figure 5.10). This can be seen through variability in the area statistics for each land cover between 2000 and 2016 (Figure 5.11). For example, the built-up area experienced an unprecedented expansion, from 30.6 km^2 (0.9%) to 160.7 km^2 (4.6%) between 2000 and 2016 respectively, which represents an increase of 130.1 km^2 (424.5%) in the urban/built-up class. Similarly, the area in mining increased by 172.1 km^2 (440.9%), while there was a small increase in the area of water by 8.1 km^2 (15.6%) over the same period.

Table 5.5 Summary of classification statistics for 2000 and 2016

Land Cover Class	2000		2016		Change (2000-2016)	
	Area (km^2)	%	Area (km^2)	%	Area (km^2)	%
Agriculture	1267.9	36.0	1240.9	35.3	-27	-2.1
Built-up	30.6	0.9	160.7	4.9	130.1	424.5
Extraction	39.1	1.1	211.0	5.6	171.9	440.9
Forest	2131.5	60.5	1848.7	52.1	-282.8	-13.2
Water	51.6	1.5	59.7	1.8	8.1	15.6
Total	3520.7		3521.0			

By contrast, the forest and agriculture all decreased. The loss of forest and agriculture land between two periods is around 282.8 km^2 and 27 km^2 respectively (Table 5.5). Unexpectedly, water area increased over the period. This could be due to variations in precipitation, water levels of lakes and possible classification errors. However, this finding reflected the current demands for aquaculture growth and an increased area of fish farms in recent years (Department of Agricultural and Rural Development, 2015), and therefore classification errors and water variations are less likely to account for this increase.

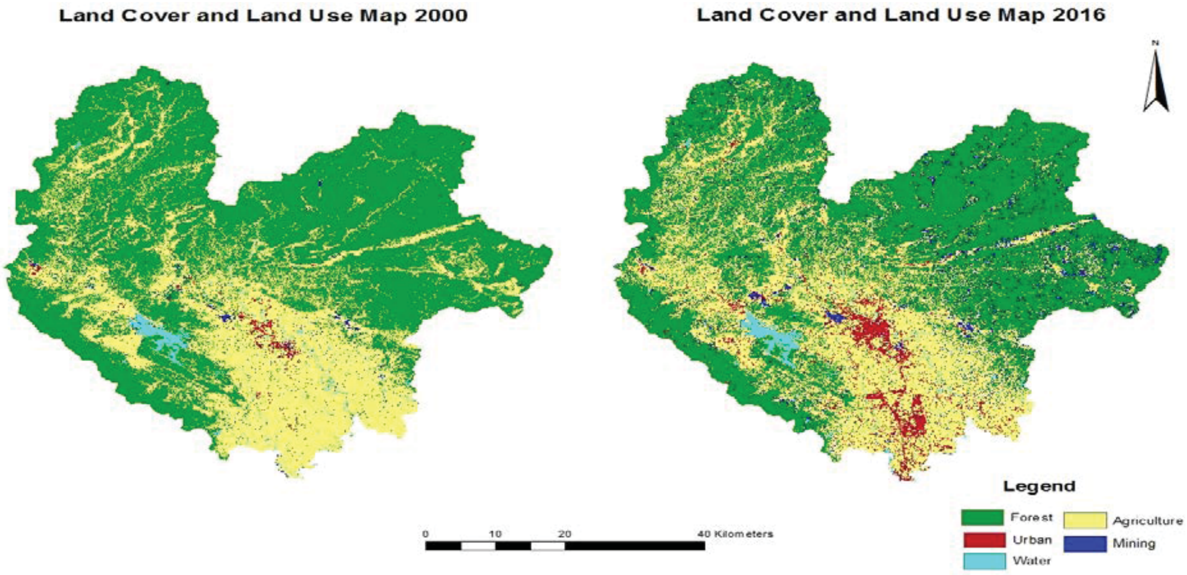

Land Cover and Land Use Map 2000
Land Cover and Land Use Map 2016
N
0
10
20
40 Kilometers
Legend
Forest
Agriculture
Urban
Mining
Water

Figure 5.10 Land cover classification products derived from Landsat data for 2000 and 2016

While the accuracy assessment of the land cover map derived from Landsat 7 data was solely based its confusion matrix assessed from a comparison with a visual interpretation of high resolution imagery on Google Earth, the most up-to-date land cover/land-use statistics from local government were collected to complement the validation process of the accuracy of the land cover map 2016 derived from the classification of 2016 Landsat 8 data. It was observed that the area statistics were similar in all classes (Figure 5.12), except for the water class. As discussed in Section 5.1.2, land cover map products derived from Landsat data agreed well with the statistics of land covers reported by the local government. These similarities may not be coincident as it is in relation to high accuracy of 2016 Landsat derived land cover map, which will be analysed in detail in Section 5.2.2.

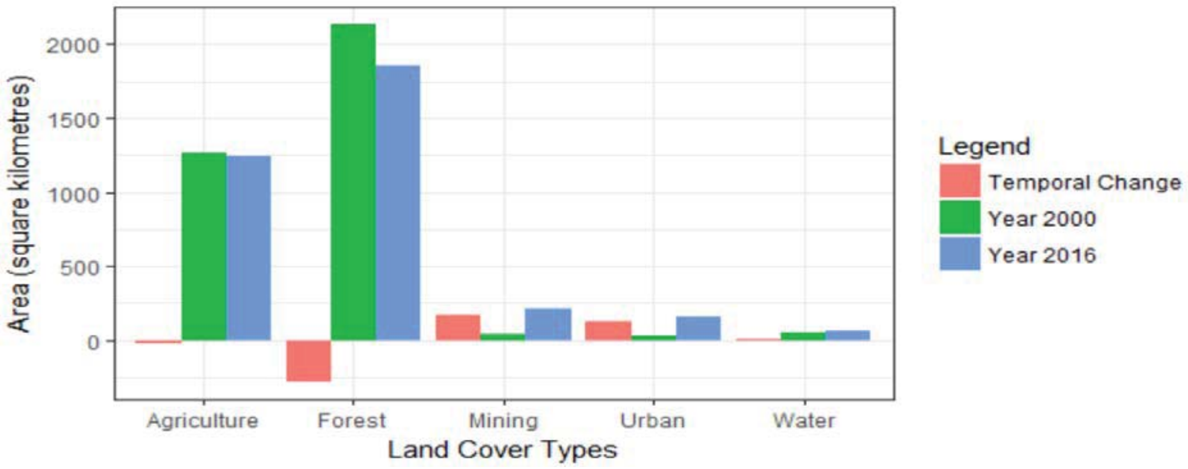


Figure 5.11 Area statistics by land cover for 2000 and 2016 and the areas of change

5.2.2 Classification accuracy assessment

Overall accuracy, Kappa statistic, producers' accuracy and users' accuracies were derived from the confusion matrices to assess the accuracy of classified maps from Landsat data and are summarised in Table 5.6. According to (Gislason et al., 2006), random forest classifier was commonly reported as an alternative for land cover classification because of its high accuracy derived products and variable importance. In this study, the overall accuracies for 2000 and 2016 were high, at 96.83% and 94.10% and Kappa coefficients were also good at 0.95 and 0.92

respectively. Users' accuracies of individual classes for 2000 were all high, with only the mining extraction lower at 88.10%, while producers' accuracies are also excellent for all classes with a range from 93.7% to 100%. Similarly, producers' and users' accuracies of individual classes in 2016 were constantly high, except for producer's accuracy of the forest class with 78.26%.

The overall accuracy for land cover/land-use maps, as mentioned earlier in Section 5.1.1, is often regarded as acceptable above 85% (Anderson, 1976; Thomlinson et al., 1999), while Lins and Kleckner (1996) set a higher standard requiring 90% accuracy. Compared to those studies, the overall accuracies for both 2000 and 2016 derived images were even better. This showed that land cover information derived from remote sensing data using the RF algorithm has performed better than that produced by traditional parametric classifiers. This study also had a higher overall accuracy than that produced by Shalaby and Tateishi (2007) using a maximum likelihood classifier on Landsat data. Furthermore, the results of this study in terms of overall accuracy were compatible with previous land cover classification research using Landsat data (Erbek et al., 2004); Kavzoglu and Mather (2003); (J. D. Paola & R. A. Schowengerdt, 1995).

Table 5.6 Summary of Landsat ETM+ and Landsat OLI classification accuracies (%)

Land cover class	**Landsat data**			
	2000		2016	
	Producer's	User's	Producer's	User's
Agriculture	97.70	93.81	98.15	92.98
Forest	95.99	98.50	78.26	100.0
Mining extraction	100	88.10	92.59	89.29
Built-up Area	93.68	99.44	100	93.18
Water	99.38	100.0	91.67	100.0
Overall accuracy	96.83		94.10	
Kappa statistic	0.95		0.92	

Although overall accuracies were high for both Landsat 7 and Landsat 8 derived products, some misclassification errors occurred, notably for forest land in 2016 at 78.26%. This was expected as forest and farming areas are sometimes confused on the ground. For example, rural farmers often grow cassava crops in forest land or practise slash and burn agriculture hilly areas. But the RF algorithm was very effective in the separation of urban/built-up and mining areas although there was substantial spectral confusion between these land surface features.

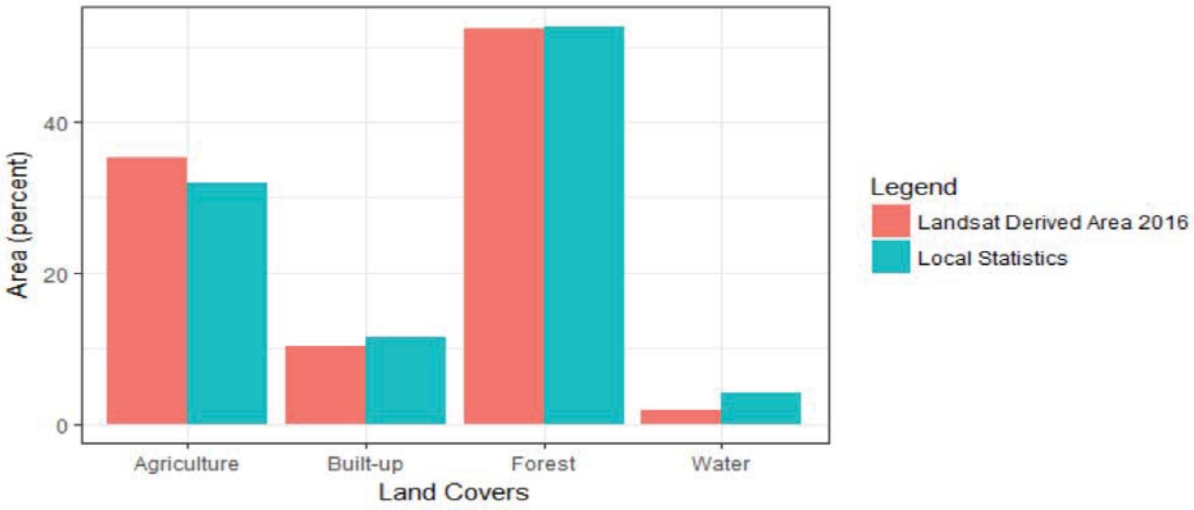


Figure 5.12 Variations in the percentage of each land cover between local statistics and the RF classification derived from the Landsat 8

5.2.3 Land cover change patterns

The advantages of products derived from satellite remote sensing include the calculation of the statistics and the capability to display the distribution of temporal changes. In this study, a matrix of land cover/land-use changes (Table 5.7) and a map displaying the spatial distribution of these changes (Figure 5.13) were created. Figure 5.13 shows that built-up/urban uses increased markedly between 2000 and 2016 in Thai Nguyen central city and southern areas (e.g., Pho Yen and Song Cong districts), while the forest in the north and east regions declined. Forest, agriculture, mining extraction and urban are the four main land cover categories; they represent 98% of total area.

The rapid growth of urban and mining land uses come from declines in both forestry and agriculture. For example, 139.7km^2 of urban/built-up land was gained from agriculture and forest land, while mining land encroached 67.8 km^2 of agriculture and 131.9 km^2 of forest during 16 years (Table 5.7). In addition, GIS analysis revealed a strong relationship between newly-developed area expansion and proximity to highways. Approximately 69.6% (100.2 km^2) of newly-built-up areas in this land cover classification occurred within 2km from main roads (Figure 5.13), while nearly 96% (137.6 km^2) of urban expansion was within 5 km from primary roads. Also, 16.5% (23.6 km^2) of new built-up area was detected within a 5 km radius the centre of Thai Nguyen city.

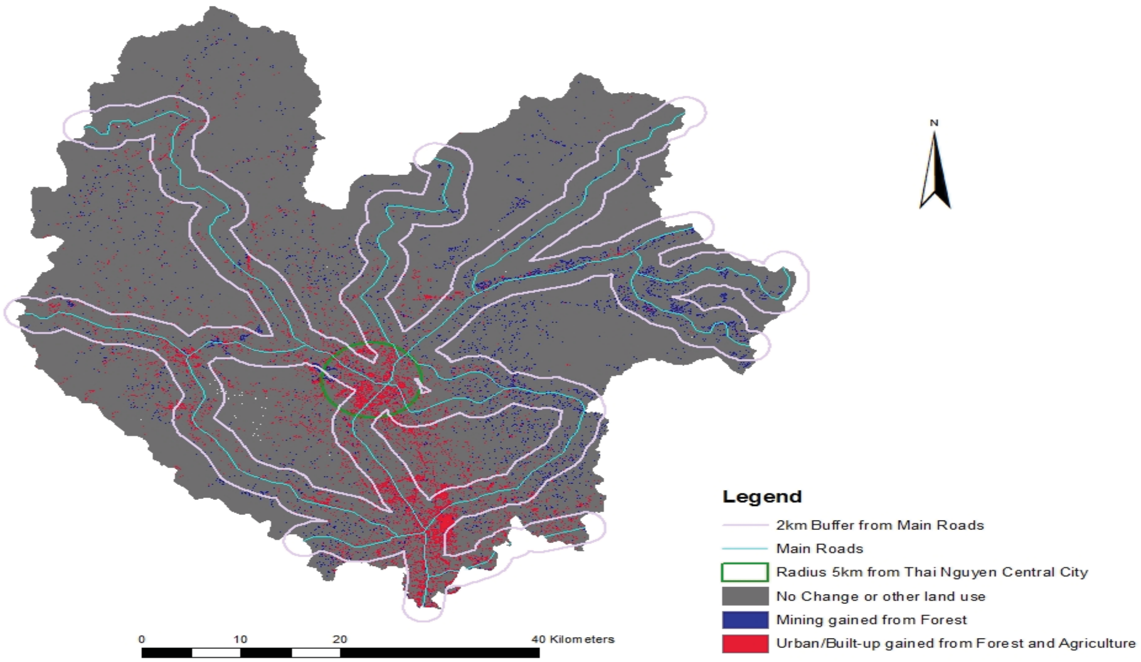

N
Legend
2km Buffer from Main Roads
Main Roads
Radius 5km from Thai Nguyen Central City
No Change or other land use
Mining gained from Forest
Urban/Built-up gained from Forest and Agriculture
0
10
20
40 Kilometers

Figure 5.13 Major changes in two land cover/ land-use categories using Landsat data between 2000 and 2016

Almost all changes in land cover/land-use in Thai Nguyen have taken place in the central city (middle) and the southern regions (Song Cong and Pho Yen districts). Interestingly, growth was largely concentrated in a strip from the southern perimeter following the Hanoi-Thai Nguyen national highway QL3. This highway is a major connector between Hanoi capital, Thai Nguyen and other northeast provinces of Vietnam. This pattern clearly reflected recent developments of the province, which focused on urbanization and industrialization. Between 2010 and 2016, industrial zones and infrastructure construction have been significant in Thai Nguyen city, Pho Yen and Song Cong districts. Samsung Electronics Vietnam Thai Nguyen Company Limited, for example, has occupied about 150 ha (primarily converted from agriculture and forest), while more than 50 other industrial zones have also occupied what was rural land[2]. Also, the rapid growth of urban population and university infrastructure in Thai Nguyen city has resulted in the expansion of the road network and residential areas. For example, there are 7 universities and 25 colleges in the province, which is considered the third largest educational provider in Vietnam.

Table 5.7 Matrix of land cover/land-use and changes (area in km^2) between 2000 and 2016

	2016					
2000	Agriculture	Forest	Mining Land	Urban	Water	Class Total
Agriculture	940.4	117.6	67.8	131.8	10.2	1267.9
Forest	263.5	1721.2	131.9	8.9	4.9	2131.5
Mining Land	19.8	6.7	6.9	2.7	2.9	39.0
Urban	10.2	0.5	2.3	16.7	0.9	30.6
Water	6.7	1.9	2.0	0.7	40.3	51.6
Class Total	1240.6	1847.9	210.9	160.8	59.2	3520.6

Long-term development policies of the province (Vietnam Government, 2007) were targeted to transform Thai Nguyen into a modernized and industrialized province by 2020. The industry, construction and service activities were targeted to account for about 87% of province's GDP, while agriculture, fisheries and forestry were predicted to occupy 13% of the land by 2020.

[2] Industrial development zones at http://enternews.vn/quy-hoach-phat-trien-kinh-te-xa-hoi-thai-nguyen-vung-nen-tang-chac-tuong-lai.html

This policy will continue to facilitate the conversion of agriculture and forest land into built-up area (e.g., industrial zones and urbanized infrastructure) if land cover/land-use information is not adequate or taken into account. This explains the importance of integrating GIS and remote sensing in monitoring and mapping land cover/land-use changes to provide timely and accurate information for sustainable land use development.

6 CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

6.1.1 Comparison of land cover classification techniques

Overall, the results of this study indicated that the RF algorithm performed best in the separation of subtropical land cover/land-use information on both Landsat 8 and Sentinel-2 data in comparison to the ANN and MLC algorithms. The RF technique also produced stable overall and individual accuracies for most classes. The ANN and MLC algorithms were less accurate in classifying Sentinel-2 data, but they did provide a higher overall accuracy for Landsat 8 classification, especially the MLC.

Non-parametric classification techniques provided higher accuracies over traditional classification approaches in this study, supporting the results of (Erbek et al., 2004; Pal & Mather, 2003; J. D. Paola & R. A. Schowengerdt, 1995). Understanding of the advantages of each technique is an important aspect of land cover classification as more advanced and accurate classifier are often required, and therefore could potentially improve the quality of classified maps (Szuster et al., 2011). As urbanization and industrialization become more prominent, and are changing local landscapes, a good classification approach may be to adopt such as the RF technique.

The MLC and ANN seemed to be less suitable for classification of Sentinel-2 data (10-m resolution) at regional scale due to the problems of smaller, mixed pixels and heterogeneous landscapes. But the MLC is proved to be suited for classifying Landsat data. In this study, the RF algorithm is recommended as the most suitable for land cover classification in subtropical regions.

6.1.2 Temporal land cover monitoring and mapping

The derived land cover products from Landsat data indicated that Landsat data can be used successfully to map and monitor land cover/land-use changes with a high accuracy. Overall, a major change in land cover/land-use has taken place in Thai Nguyen province, particularly near Thai Nguyen central city and in southern regions over last 16 years. Agriculture has been converted into built-up land, mining extraction has expanded into forest land while water had little change. The main causes of land cover changes are due to recent development resulting in the expansion of mining activities, industrial and residential zones in formerly agricultural and forest lands.

The combination of satellite remote sensing, R programming and GIS demonstrated the potential of rapid data acquisition over large areas and the informative display of spatial changes to provide timely and accurate land cover information for efficient land management and policy decisions. The RF algorithm has advantages in producing more accurate and stable overall and individual accuracies, and therefore offers the opportunity for better resource management and sustainable land development.

6.2 Recommendations

6.2.1 Comparison of land cover classification techniques

After analysing the output of different classifiers, this study showed that the MLC could be used for extracting land cover/land use information when using Landsat data as it produced relatively high overall accuracy. Also, the MLC algorithm is simple, fast and available in QGIS which is free to use. The ANN algorithm is not recommended for classifying Sentinel-2 data as it resulted in lower overall accuracies in this study. In addition, the ANN algorithm is time-consuming, as well as computationally and mathematically intensive. However, this classifier could be used to classify Landsat 8 data as it provided a relatively good overall accuracy, and is available for use in ENVI 5.3.

The RF algorithm demonstrated its superior performance in classifying both Landsat 8 and Sentinel-2 data with high and stable accuracies. This approach should be preferred for Sentinel-2 and Landsat 8 land cover classification. However, it is not available in ENVI 5.3 (at the time of writing this thesis), so users may not have access to this classifier. Another way to apply the RF classifier for any satellite data is to use R or Python to carry out land cover classification, but some skill in programming is needed.

6.2.2 Land cover monitoring and mapping

With the increasing availability of free satellite imagery and open software like R programming, QGIS and Python, making use of these technologies can offer benefits for land cover classification with high accuracy. Local land managers and decision-makers should adopt these classification approaches as they are both cost-effective and provide good accuracy.

References

Anderson, J. R. (1976). *A land use and land cover classification system for use with remote sensor data* (Vol. 964): US Government Printing Office.

Baker, L. A., Brazel, A., & Westerhoff, P. (2004). Environmental consequences of rapid urbanization in warm, arid lands: case study of Phoenix, Arizona (USA). *WIT Transactions on Ecology and the Environment, 72*.

Bausch, W., & Duke, H. (1996). Remote sensing of plant nitrogen status in corn. *Transactions of the ASAE-American Society of Agricultural Engineers, 39*(5), 1869-1878.

Binh, T. N., Vromant, N., Hung, N. T., Hens, L., & Boon, E. (2005). Land cover changes between 1968 and 2003 in Cai Nuoc, Ca Mau peninsula, Vietnam. *Environment, Development and Sustainability, 7*(4), 519-536.

Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., & Pebesma, E. J. (2008). *Applied spatial data analysis with R* (Vol. 747248717): Springer.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Breiman, L. (2003). RFtools–two-eyed algorithms, Invited talk at SIAM International Conference on Data Mining.

Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2011). Package'randomForest'. *software available at URL: http://stat-www. berkeley. edu/users/breiman/RandomForests*.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*: CRC press.

Campbell, J. B., & Wynne, R. H. (2011). *Introduction to remote sensing*: Guilford Press.

Castrence, M., Nong, D. H., Tran, C. C., Young, L., & Fox, J. (2014). Mapping urban transitions using multi-temporal Landsat and DMSP-OLS night-time lights imagery of the Red River Delta in Vietnam. *Land, 3*(1), 148-166.

Chuvieco, E. (2016). *Fundamentals of Satellite Remote Sensing: An Environmental Approach*: CRC press.

Civco, D. L. (1993). Artificial neural networks for land-cover classification and mapping. *International journal of geographical information science, 7*(2), 173-186.

Clement, F., & Amezaga, J. M. (2008). Linking reforestation policies with land use change in northern Vietnam: Why local factors matter. *Geoforum, 39*(1), 265-277.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of Environment, 37*(1), 35-46.

Congalton, R. G., & Green, K. (2008). *Assessing the accuracy of remotely sensed data: principles and practices*: CRC press.

Congedo, L. (2013). Semi-automatic classification plugin for QGIS. *Sapienza University, Rome*.

Cracknell, A. P. (2007). *Introduction to remote sensing*: CRC press.

Dale, V. H. (1997). The relationship between land-use change and climate change. *Ecological applications, 7*(3), 753-769.

David Dibiase, James L.Sloan, Ryan Baxter, Wesley Stroh, & King, B. F. (2017). The Nature of Geographic Information. Retrieved from https://www.e-education.psu.edu/natureofgeoinfo/node/1672

Davis, S. M., Landgrebe, D. A., Phillips, T. L., Swain, P. H., Hoffer, R. M., Lindenlaub, J. C., & Silva, L. F. (1978). Remote sensing: the quantitative approach. *New York, McGraw-Hill International Book Co., 1978. 405 p.*

DeGloria, S., Laba, M., Gregory, S., Braden, J., Ogurcak, D., Hill, E., . . . Beecher, J. (2000). *Conventional and fuzzy accuracy assessment of land cover maps at regional scale.* Paper presented at the Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.

Department of Agricultural and Rural Development. (2015). Aquaculture Development Retrieved from http://sonnvptnt.thainguyen.gov.vn/-/san-luong-nuoi-trong-thuy-san-at-7-778-tan

Dewan, A. M., & Yamaguchi, Y. (2009a). Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization. *Applied Geography, 29*(3), 390-401.

Dewan, A. M., & Yamaguchi, Y. (2009b). Using remote sensing and GIS to detect and monitor land use and land cover change in Dhaka Metropolitan of Bangladesh during 1960–2005. *Environmental Monitoring and Assessment, 150*(1), 237-249.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., . . . Martimort, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment, 120*, 25-36.

Dwivedi, R., Kandrika, S., & Ramana, K. (2004). Comparison of classifiers of remote-sensing data for land-use/land-cover mapping. *CURRENT SCIENCE-BANGALORE-, 86*(2), 328-334.

Erbek, F. S., Özkan, C., & Taberner, M. (2004). Comparison of maximum likelihood classification method with supervised artificial neural network algorithms for land use activities. *International Journal of Remote Sensing, 25*(9), 1733-1748.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., . . . Gibbs, H. K. (2005). Global consequences of land use. *science, 309*(5734), 570-574.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote sensing of Environment, 80*(1), 185-201.

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote sensing of Environment, 61*(3), 399-409.

General Statistics Department of Vietnam. (2011). *Population and Employment* Vietnam: Statistical Publishing House Retrieved from http://www.gso.gov.vn/default.aspx?tabid=512&ItemID=11973

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover classification. *Pattern Recognition Letters, 27*(4), 294-300.

Govender, M., Chetty, K., & Bulcock, H. (2007). A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa, 33*(2), 145-151.

Government, V. (2013). Short and Long-Term Land Use Planning by 2020 for Thai Nguyen.

Green, E., Mumby, P., Edwards, A., & Clark, C. (1996). A review of remote sensing for the assessment and management of tropical coastal resources. *Coastal management, 24*(1), 1-40.

GTZ Office Hanoi. (2005). *Rapid Assessment Of Mammals In The Tam Dao National Park*. Retrieved from Vietnam:

Guide, E. U. s. (2008). ENVI on-line software user's manual. *ITT Visual Information Solutions*.

Hansen, M., Dubayah, R., & DeFries, R. (1996). Classification trees: an alternative to traditional land cover classifiers. *International Journal of Remote Sensing, 17*(5), 1075-1081.

Hao Ho. (2015). Local Geography of Thai Nguyen. Retrieved from https://sites.google.com/site/dialitinhthainguyen/home/dan-so

Hepner, G., Logan, T., Ritter, N., & Bryant, N. (1990). Artificial neural network classification using a minimal training set- Comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing, 56*(4), 469-473.

Hepner, G. F. (1990). Artificial neural network classification using a minimal training set. Comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing, 56*(4), 469-473.

Hoang Ngoc Ha. (2008). Status and Solutions for forest development in Hoa Binh Commune, Dong Hy District, Thai Nguyen.

Huang, C., Davis, L., & Townshend, J. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing, 23*(4), 725-749.

Huong, T. T. L., Anh, T. N. T., Yen, T. N. N., Thanh, T. N., & Thin, N. N. (2012). The Status of Precious Medical Plant Species in Thai Nguyen, Vietnam. *Journal of Natural Science and Technology, Vietnam National University, 28*, 173-194.

Jensen, J. R. (1986). *Introductory digital image processing: a remote sensing perspective*. Retrieved from

Jetz, W., Wilcove, D. S., & Dobson, A. P. (2007). Projected impacts of climate and land-use change on the global diversity of birds. *PLoS Biol, 5*(6), e157.

Jha, C., Goparaju, L., Tripathi, A., Gharai, B., Raghubanshi, A., & Singh, J. (2005). Forest fragmentation and its impact on species diversity: an analysis using remote sensing and GIS. *Biodiversity and Conservation, 14*(7), 1681-1698.

Joseph, G. (2005). *Fundamentals of remote sensing*: Universities Press.

Joy, M. K., & Death, R. G. (2004). Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. *Freshwater Biology, 49*(8), 1036-1052.

Kavzoglu, T., & Colkesen, I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation, 11*(5), 352-359.

Kavzoglu, T., & Mather, P. M. (2003). The use of backpropagating artificial neural networks in land cover classification. *International Journal of Remote Sensing, 24*(23), 4907-4938.

Khiry, M. A., & Csaplovics, E. (2007). *Appropriate methods for monitoring and mapping land cover changes in semi-arid areas in North Kordofan (Sudan) by using satellite imagery and spectral mixture analysis.* Paper presented at the Analysis of Multi-temporal Remote Sensing Images, 2007. MultiTemp 2007. International Workshop on the.

Li, X., & Yeh, A. (1998). Principal component analysis of stacked multi-temporal images for the monitoring of rapid urban expansion in the Pearl River Delta. *International Journal of Remote Sensing, 19*(8), 1501-1518.

Lins, K., & Kleckner, R. (1996). Land cover mapping: An overview and history of the concepts. *Gap analysis: A landscape approach to biodiversity planning*, 57-65.

Liu, H., & Zhou, Q. (2004). Accuracy analysis of remote sensing change detection by rule-based rationality evaluation with post-classification comparison. *International Journal of Remote Sensing, 25*(5), 1037-1050.

Logan, A. A., Hanson, B. A., & Seeger, C. J. (2014). Introduction to QGIS.

Lu, D., Mausel, P., Batistella, M., & Moran, E. (2004). Comparison of land-cover classification methods in the Brazilian Amazon Basin. *Photogrammetric Engineering & Remote Sensing, 70*(6), 723-731.

Maini, R., & Aggarwal, H. (2010). A comprehensive review of image enhancement techniques. *arXiv preprint arXiv:1003.4053*.

Mas, J.-F. (1999). Monitoring land-cover changes: a comparison of change detection techniques. *International Journal of Remote Sensing, 20*(1), 139-152.

Mas, J. F., & Flores, J. J. (2008). The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing, 29*(3), 617-663.

Maselli, F., Conese, C., & Petkov, L. (1994). Use of probability entropy for the estimation and graphical representation of the accuracy of maximum likelihood classifications. *ISPRS Journal of Photogrammetry and Remote Sensing, 49*(2), 13-20.

McGee, T. (1995). The urban future of Vietnam. *Third World Planning Review, 17*(3), 253.

McGee, T. G. (2008). The urban future of Vietnam reconsidered.

Meyer, W. B. (1995). Past and Present Land Use and Land Cover in the U. S. A. *Consequences: The nature and implications of environmental change, 1*(1).

Meyer, W. B., & BL Turner, I. (1994). *Changes in land use and land cover: a global perspective* (Vol. 4): Cambridge University Press.

Meyer, W. B., & Turner, B. L. (1992). Human population growth and global land-use/cover change. *Annual review of ecology and systematics, 23*(1), 39-61.

Ministry of Planning and Investment. (2005). Industrial Zone Planning and Development between 2006 and 2020 Retrieved from http://www.chinhphu.vn/portal/page/portal/chinhphu/noidungcackhucongnghiepkhuchexuat?categoryId=879&articleId=10001189. http://www.chinhphu.vn/portal/page/portal/chinhphu/noidungcackhucongnghiepkhuchexuat?categoryId=879&articleId=10001189

Morris, R., MacNeela, P., Scott, A., Treacy, P., Hyde, A., O'Brien, J., . . . Drennan, J. (2008). Ambiguities and conflicting results: The limitations of the kappa statistic in establishing the interrater reliability of the Irish nursing minimum data set for mental health: A discussion paper. *International journal of nursing studies, 45*(4), 645-647.

Muttitanon, W., & Tripathi, N. (2005). Land use/land cover changes in the coastal zone of Ban Don Bay, Thailand using Landsat 5 TM data. *International Journal of Remote Sensing, 26*(11), 2311-2323.

Nangendo, G., Skidmore, A. K., & van Oosten, H. (2007). Mapping East African tropical forests and woodlands—a comparison of classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing, 61*(6), 393-404.

NASA. (2017). Landsat Missions: Imaging the Earth Since 1972. Retrieved from https://landsat.usgs.gov/landsat-missions-timeline

NASA, & Ministry of Economy Trade and Industry of Japan. (2011). Global Elevation Map. Retrieved from https://asterweb.jpl.nasa.gov/gdem.asp

Nguyen, H. H., Everaert, G., Gabriels, W., Hoang, T. H., & Goethals, P. L. (2014). A multimetric macroinvertebrate index for assessing the water quality of the Cau river basin in Vietnam. *Limnologica-Ecology and Management of Inland Waters, 45*, 16-23.

Nilson, N. J. (1925). *Learning Machines: Foundations of trainable pattern-classifying systems*: McGraw-Hill.

Otukei, J. R., & Blaschke, T. (2010). Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *International Journal of Applied Earth Observation and Geoinformation, 12*, S27-S31.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217-222.

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote sensing of Environment, 86*(4), 554-565.

Paola, J., & Schowengerdt, R. (1995). A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery. *International Journal of Remote Sensing, 16*(16), 3033-3058.

Paola, J. D., & Schowengerdt, R. A. (1995). A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification. *IEEE Transactions on Geoscience and remote sensing, 33*(4), 981-996.

Pham, M. H., & Yamaguchi, Y. (2006). Monitoring land cover change of the Hanoi city center under impacts of urbanization by using remote sensing.

Phan Manh Cuong. (2015). Industrial Zone Sustainable Development in Thai Nguyen Province. doi:62340101

Phuong, V. T. (2007). Forest environment of Vietnam: features of forest vegetation and soils. *Forest Environments in the Mekong River Basin*, 189-200.

Poursanidis, D., Chrysoulakis, N., & Mitraka, Z. (2015). Landsat 8 vs. Landsat 5: A comparison based on urban and peri-urban land cover mapping. *International Journal of Applied Earth Observation and Geoinformation, 35*, 259-269.

Quang, N., & Kammeier, H. D. (2002). Changes in the political economy of Vietnam and their impacts on the built environment of Hanoi. *Cities, 19*(6), 373-388.

Rawat, J., & Kumar, M. (2015). Monitoring land use/cover change using remote sensing and GIS techniques: A case study of Hawalbagh block, district Almora, Uttarakhand, India. *The Egyptian Journal of Remote Sensing and Space Science, 18*(1), 77-84.

Richards, J. A., & Richards, J. (1999). *Remote sensing digital image analysis* (Vol. 3): Springer.

Rodriguez-Galiano, V., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P. M., & Jeganathan, C. (2012). Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote sensing of Environment, 121*, 93-107.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing, 67*, 93-104.

Rogan, J., Franklin, J., & Roberts, D. A. (2002). A comparison of methods for monitoring multitemporal vegetation change using Thematic Mapper imagery. *Remote sensing of Environment, 80*(1), 143-156.

Sanyal, J., & Lu, X. (2004). Application of remote sensing in flood management with special reference to monsoon Asia: a review. *Natural Hazards, 33*(2), 283-301.

Serra, P., Pons, X., & Sauri, D. (2003). Post-classification change detection with data from different sensors: some accuracy considerations. *International Journal of Remote Sensing, 24*(16), 3311-3340.

Seto, K. C., Güneralp, B., & Hutyra, L. R. (2012). Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences, 109*(40), 16083-16088.

Seto, K. C., & Liu, W. (2003). Comparing ARTMAP neural network with the maximum-likelihood classifier for detecting urban change. *Photogrammetric Engineering & Remote Sensing, 69*(9), 981-990.

Shalaby, A., & Tateishi, R. (2007). Remote sensing and GIS for mapping and monitoring land cover and land-use changes in the Northwestern coastal zone of Egypt. *Applied Geography, 27*(1), 28-41.

Singh, A. (1989). Review article digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing, 10*(6), 989-1003.

Spoto, F., Sy, O., Laberinti, P., Martimort, P., Fernandez, V., Colin, O., . . . Meygret, A. (2012). *Overview of Sentinel-2*. Paper presented at the Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International.

Story, M., & Congalton, R. G. (1986). Accuracy assessment: a user's perspective. *Photogrammetric Engineering and Remote Sensing, 52*(3), 397-399.

Stow, D., Tinney, L., & Estes, J. (1980). Deriving land use/land cover change statistics from Landsat-A study of prime agricultural land.

Strahler, A. H. (1980). The use of prior probabilities in maximum likelihood classification of remotely sensed data. *Remote sensing of Environment, 10*(2), 135-163.

Szuster, B. W., Chen, Q., & Borger, M. (2011). A comparison of classification techniques to support land cover and land use analysis in tropical coastal zones. *Applied Geography, 31*(2), 525-532.

Taylor, P. J. (1977). *Quantitative methods in geography: an introduction to spatial analysis*: Houghton Mifflin.

Tayyebi, A., Pijanowski, B. C., Linderman, M., & Gratton, C. (2014). Comparing three global parametric and local non-parametric models to simulate land use change in diverse areas of the world. *Environmental Modelling & Software, 59*, 202-221.

Thai Nguyen Department of Commerce and Industry. (2012). Samsung Invested in Thai Nguyen. Retrieved from http://congthuongthainguyen.gov.vn/

Thai Nguyen Department of Natural Resource and Environment. (2016a). Land Use Statistics.

Thai Nguyen Department of Natural Resource and Environment. (2016b). *Thai Nguyen Land Use Statistics*. Department of Natural Resource and Environment. Thai Nguyen

Thai Nguyen People's Committee. (2009). *Decision on the approval of Urban System Development Planning*. Thai Nguyen, Vietnam: Thai Nguyen Portal.

Thai Nguyen People's Committee. (2010). *Industrial Zone Development Strategy between 2011 and 2015*. Thai Nguyen: Thai Nguyen News.

Therneau, T. M., Atkinson, B., & Ripley, M. B. (2010). The rpart package.

Thomlinson, J. R., Bolstad, P. V., & Cohen, W. B. (1999). Coordinating methodologies for scaling landcover classifications from site-specific to global: Steps toward validating global map products. *Remote sensing of Environment, 70*(1), 16-28.

Unwin, D. J., & Fisher, P. (2005). Research Agenda. *Re-presenting GIS*, 277.

van Beijma, S., Comber, A., & Lamb, A. (2014). Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data. *Remote sensing of Environment, 149*, 118-129.

Vietnam Geology Society. (2014). Thai Nguyen Mining Monitoring Management. Retrieved from http://dgmv.gov.vn

Vietnam Government. (2007). *Approval of Socio-economic Development Strategy for Thai Nguyen province till 2020*. Vietnam Retrieved from http://www.chinhphu.vn/portal/page/portal/chinhphu/hethongvanban?class_id=1&mode=detail&document_id=23936.

Vietnam Government. (2013). *Land Law*. Vietnam.

Vuong, Q. H. (2014). Vietnam's Political Economy in Transition (1986-2016).

Weismiller, R., Kristof, S., Scholz, D., Anuta, P., & Momin, S. (1977). Change detection in coastal zone environments. *Photogrammetric Engineering and Remote Sensing, 43*(12).

Weng, Q. (2002). Land use change analysis in the Zhujiang Delta of China using satellite remote sensing, GIS and stochastic modelling. *Journal of Environmental Management, 64*(3), 273-284.

Yuan, F., Sawaya, K. E., Loeffelholz, B. C., & Bauer, M. E. (2005). Land cover classification and change analysis of the Twin Cities (Minnesota) Metropolitan Area by multitemporal Landsat remote sensing. *Remote sensing of Environment, 98*(2), 317-328.

Zhao, S., Da, L., Tang, Z., Fang, H., Song, K., & Fang, J. (2006). Ecological consequences of rapid urban expansion: Shanghai, China. *Frontiers in Ecology and the Environment, 4*(7), 341-346.

Zhou, L., Dickinson, R. E., Tian, Y., Fang, J., Li, Q., Kaufmann, R. K., . . . Myneni, R. B. (2004). Evidence for a significant urbanization effect on climate in China. *Proceedings of the National Academy of Sciences of the United States of America, 101*(26), 9540-9544.

Zuhlke, M., Fomferra, N., Brockmann, C., Peters, M., Veci, L., Malik, J., & Regner, P. (2015). *SNAP (Sentinel Application Platform) and the ESA Sentinel 3 Toolbox*. Paper presented at the Sentinel-3 for Science Workshop.

Appendices

Appendix 1: TNMT land use statistics

- TNMT land use statistics

TNMT Land Use Type	Total Area (ha)	Redefined Land Use Classes	Total Area (ha)
Total Area (ha)	352666		352666
Agriculture	**112673**	Agriculture	112673
Annual Agriculture	51064		
Rice	45067		
Other annual agriculture	16322		
Other farming land	219		
Forestry	**185922**	Forest	185922
Production forest	109717		
Defensive forest	36846		
Special forest	39359		
Land for Aquaculture	**13954**	Water	13954
Rivers, streams, canals	5651		
Special water surface	3662		
Aquaculture farms	4641		
Residential land	**12135**	Urban/built-up	35337
Rural residential land	9907		
Urban residential land	2228		
Special-use Land	**23202**		
Land for construction of offices	143		
Land for national defence	3473		
Land for national security	479		
Land for construction of state facilities	1234		
Land for non-agricultural business	5093		
land for public use	11808		
Land for religious organizations	79		
Land for other religious facilities	68		
Land for cemetery and funeral services	813		
Other non-agricultural land	12		
Unused land	**4780**	Mining	4780
Flat unused land	1084		
Mountainous unused land	1534		
Bare land	2162		

- Five land cover/land-use extracted from the TNMT area statistics

Land Use Type	Total Area (km^2)	Percent (%)
Agriculture	1126.7	31.9
Forest	1859.2	52.7
Built-up/Urban	353.4	10.0
Mining Extraction	47.8	1.4
Water	139.6	4.0
Total Area (ha)	3526.7	100.0

Appendix 2: Accuracy statistics for 20-m Sentinel-2 data using the three classification algorithms

- Accuracy statistics for 20-m Sentinel-2 data using MLC

Overall accuracy: 82.8%

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	51	2	2	2	0	57	89.5
Forest	1	21	0	0	0	22	95.5
Mining	1	0	23	16	1	41	56.1
Urban	1	0	1	23	1	26	88.5
Water	0	0	1	0	22	23	95.7
Total	54	23	27	41	24	169	
Producers' Accuracy	94.4	91.3	85.2	56.1	91.7		

- Accuracy statistics for 20-m Sentinel-2 data using ANN

Overall accuracy: 86.3%

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	48	1	0	2	0	51	94.1
Forest	2	22	0	0	0	24	91.7
Mining	0	0	20	0	7	27	74.1
Urban	4	0	7	39	0	50	78.0
Water	0	0	0	0	17	17	100
Total	54	23	27	41	24	169	
Producers' Accuracy	88.9	95.7	74.1	95.1	70.8		

- Accuracy statistics for 20-m Sentinel-2 data using RF

Overall accuracy: 89.9%

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	52	4	0	0	0	56	92.9
Forest	0	19	0	0	0	19	100
Mining	0	0	23	0	4	27	85.2
Urban	2	0	4	38	0	44	86.4
Water	0	0	0	0	20	20	100
Total	54	23	27	38	24	166	
Producers' Accuracy	96.3	82.6	85.2	100	83.3		

Appendix 3: Accuracy statistics for 10-m Sentinel-2 data using the three classification algorithms

- Accuracy statistics for 10-m Sentinel-2 data using MLC

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	51	3	1	2	1	58	87.9
Forest	0	20	0	0	0	20	100
Mining	0	0	24	17	1	42	57.1
Urban	3	0	2	22	0	27	81.5
Water	0	0	0	0	22	22	100
Total	54	23	27	41	24	169	
Producers' Accuracy	94.4	86.9	88.9	53.7	91.7		

- Accuracy statistics for 10-m Sentinel-2 data using ANN

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	52	3	6	12	2	75	69.3
Forest	1	20	0	0	0	21	95.2
Mining	0	0	19	3	1	23	82.6
Urban	1	0	2	26	0	29	89.7
Water	0	0	0	0	21	21	100
Total	54	23	27	41	24	169	
Producers' Accuracy	96.3	86.9	70.4	63.4	87.5		

- Accuracy statistics for 10-m Sentinel-2 data using RF

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	49	5	2	1	1	58	84.5
Forest	0	18	0	0	0	18	100
Mining	0	0	24	1	1	26	92.3
Urban	5	0	1	39	0	45	86.7
Water	0	0	0	0	22	22	100
Total	54	23	27	41	24	169	
Producers' Accuracy	90.7	78.3	88.9	95.1	91.7		

Appendix 4: Accuracy statistics for Landsat 8 data using the three classification algorithms

- Accuracy statistics for Landsat 8 data using MLC

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	50	3	1	0	0	54	92.6
Forest	1	20	0	0	0	21	95.2
Mining	2	0	25	3	3	33	75.8
Urban	1	0	1	38	1	41	92.7
Water	0	0	0	0	20	20	100
Total	54	23	27	41	24	169	
Producers' Accuracy	92.6	87	92.6	92.7	83.3		

- Accuracy statistics for Landsat 8 data using ANN

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	54	9	3	0	4	70	77.1
Forest	0	14	0	0	0	14	100
Mining	0	0	17	0	4	21	80.9
Urban	0	0	7	41	0	48	85.4
Water	0	0	0	0	16	16	100
Total	54	23	27	41	24	169	
Producers' Accuracy	100	60.8	62.7	100	66.7		

- Accuracy statistics for Landsat 8 data using RF

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	50	3	1	0	0	54	92.6
Forest	1	20	0	0	0	21	95.2
Mining	2	0	25	3	3	33	75.8
Urban	1	0	1	38	1	41	92.7
Water	0	0	0	0	20	20	100
Total	54	23	27	41	24	169	
Producers' Accuracy	92.6	86.9	92.6	92.7	83.3		

Appendix 5: Accuracy statistics for Landsat 7 data using the RF classification algorithms

- Accuracy statistics for Landsat 7 data using RF

	Agriculture	Forest	Mining	Urban	Water	Total	Users' Accuracy
Agriculture	212	10	0	3	1	226	93.8
Forest	4	263	0	0	0	267	98.5
Mining	0	1	74	9	0	84	88.1
Urban	1	0	0	178	0	179	99.4
Water	0	0	0	0	160	160	100
Total	217	274	74	190	161	916	
Producers' Accuracy	97.7	95.9	100	93.7	99.2		