

https://dspace.org/

Exploring issues of balanced versus imbalanced samples in mapping grass community in the telperion reserve using high resolution images and selected machine learning algorithms

Isoa, Mary Itohan

2018

https://hdl.handle.net/10539/25639

Isoa, Mary Itohan, (2018) Exploring issues of balanced versus imbalanced samples in mapping grass community in the telperion reserve using high resolution images and selected machine learning algorithm, University of the Witwatersrand, Johannesburg, https://hdl.handle.net/10539/25639.

Downloaded from a DSpace repository. https://dspace.org/



EXPLORING ISSUES OF BALANCED VERSUS IMBALANCED SAMPLES IN MAPPING GRASS COMMUNITY IN THE TELPERION RESERVE USING HIGH RESOLUTION IMAGES AND SELECTED MACHINE LEARNING ALGORITHMS

By

Isoa, Mary Itohan (1531061)

SUPERVISOR: Dr Elhadi Adam

A dissertation report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science (GIS and Remote Sensing).

MARCH 2018

Johannesburg, South Africa

ABSTRACT

Accurate vegetation mapping is essential for a number of reasons, one of which is for conservation purposes. The main objective of this research was to map different grass communities in the game reserve using RapidEye and Sentinel-2 MSI images and machine learning classifiers [support vector machine (SVM) and Random forest (RF)] to test the impacts of balanced and imbalance training data on the performance and the accuracy of Support Vector Machine and Random forest in mapping the grass communities and test the sensitivities of pixel resolution to balanced and imbalance training data in image classification. The imbalanced and balanced data sets were obtained through field data collection.

The results show RF and SVM are producing a high overall accuracy for Sentinel-2 imagery for both the balanced and imbalanced data set. The RF classifier has yielded an overall accuracy of 79.45% and kappa of 74.38% and an overall accuracy of 76.19% and kappa of 73.21% using imbalanced and balanced training data respectively. The SVM classifier yielded an overall accuracy of 82.54% and kappa of 80.36% and an overall accuracy of 82.21% and a kappa of 78.33% using imbalanced and balanced training data respectively.

For the RapidEye imagery, RF and SVM algorithm produced overall accuracy affected by a balanced data set leading to reduced accuracy. The RF algorithm had an overall accuracy that dropped by 6% (from 63.24% to 57.94%) while the SVM dropped by 7% (from 57.31% to 50.79%). The results thereby show that the imbalanced data set is a better option when looking at the image classification of vegetation species than the balanced data set.

The study recommends the implementation of ways of handling misclassification among the different grass species to improve classification for future research. Further research can be carried out on other types of high resolution multispectral imagery using different advanced algorithms on different training size samples.

DECLARATION

I, Mary Itohan Isoa (1531061), attest that this research is my unassisted work. It is being submitted for the degree of Master of Science at the University of Witwatersrand, Johannesburg. It has not been submitted at any other University for an examination or degree.

Signature: Date: 23rd day of March 2018

Mary Itohan Isoa

Dedication

I dedicate this research to my brother, Thomas, for his support all round in helping me to achieve this accomplishment. There cannot be enough thanks to you for your unfailing support and continuous encouragement throughout my years of study. This could not have been done without you.

Acknowledgments

All praise and honour go to God Almighty for His mercies and favours in my life, especially during this program.

My sincerest gratitude goes to my supervisor Dr. Elhadi Adam for his guidance and support not just throughout my research, but during the master's program. Your support gave me the opportunity to improve my knowledge on remote sensing.

TABLE OF CONTENTS

Acknowledgments……v
LIST OF FIGURES……viii
LIST OF TABLES……ix
LIST OF ABBREVIATIONS……ix

1.1. General Introduction 2
1.2. Problem Statement 5
1.3. Aims and objectives 6
2.1. Mapping grass communities using remote sensing 8
2.1.1. Importance and principle of mapping grass communities 9
2.2. Mapping grass communities using multispectral remote sensing 9
2.3. Mapping grass community using hyperspectral remote sensing 10
2.4. Mapping grass communities using new advanced multispectral data 11
2.5. Ground and training sampling for mapping grass communities 11
2.6. Machine learning classifiers for mapping grass communities 12
2.7. Conclusion 14
3.1. Study area 16
3.2. Remote sensing data acquisition and pre-processing 18
3.2.1. Sentinel-2 Multispectral Instrument (MSI) image acquisition 18
3.2.2. RapidEye image acquisition 19
3.3. Remote sensing data pre-processing 20
3.4. Field data collection 20
3.5. Image classification 22
3.5.1. Support Vector Machines 23
3.5.2. Random forest classifiers 28
3.6. Accuracy assessment 30
4.1. Optimization of RF parameters 32
4.1.1. Sentinel-2 MSI imagery 32
4.1.2. RapidEye imagery 33

4.2. Parameter tuning of SVM 35
4.2.1. Sentinel-2 MSI 35
4.2.2. RapidEye imagery 35
4.3. RF and SVM performance in mapping grass community 35
4.3.1. Sentinel-2 MSI imagery (imbalanced training data) 35
4.3.2. Sentinel-2 MSI imagery (balanced training data) 36
4.3.3. RapidEye imagery (imbalanced training data) 37
4.3.4. RapidEye imagery (balanced training data) 38
4.4. RapidEye and Sentinel-2 bands significance 39
4.4.1. Sentinel-2 MSI imagery (balanced training data) 39
4.4.2. Sentinel-2 MSI imagery (imbalanced training data) 40
4.4.3. RapidEye Imagery (balanced training data) 42
4.4.4. RapidEye imagery (imbalanced training data) 43
4.5. Accuracy assessment 45
4.5.1. Sentinel-2 MSI (balanced dataset) 45
4.5.2. Sentinel-2 MSI (imbalanced dataset) 45
4.5.3 RapidEye imagery (balanced dataset) 50
4.5.4. RapidEye imagery (imbalanced dataset) 50
5.1. Discussion 58
5.2. Conclusion 59
References 61

LIST OF FIGURES

Figure 1: The location of Telperion Nature Reserve……...….17

Figure 2: Support vector machine linear classifier……24

Figure 3: Support vector machine non-linear classifier…….…25

Figure 4: The Main idea of support vector machine……..26

Figure 5: Workflow and main idea of Random forest……...27

Figure 6: RF parameter optimization for imbalanced and balanced data of Sentinel-2……...……..….33

Figure 7: RF parameter optimization for imbalanced and balanced data of RapidEye……34

Figure 8: Vegetation mapping classification using RF and SVM classification algorithm for Sentinel-2 (imbalanced)…….36

Figure 9: Vegetation mapping classification using RF and SVM classification algorithm for Sentinel-2 (balanced)…….37

Figure 10: Vegetation mapping classification based using RF and SVM classification algorithm for RapidEye (imbalanced)…….……...38

Figure 11: Vegetation mapping classification using RF and SVM classification algorithm for RapidEye (imbalanced)……...……...39

Figure 12: Sentinel-2 band significance in vegetation classification for all vegetation species and for each vegetation species (balanced)……...……..…….40

Figure 13: Sentinel-2 band significance in vegetation classification for all vegetation species and for each vegetation species (imbalanced)…….....……....41

Figure 14: RapidEye band significance in vegetation classification for all vegetation species and for each vegetation species (balanced)…...……...43

Figure 15: RapidEye band significance in vegetation classification for all vegetation species and for each vegetation species (imbalanced)……...……...…44

LIST OF TABLES

Table 1: Spectral bands of Sentinel 2 A imagery……..……18

Table 2: Spectral bands of RapidEye imagery……...19

Table 3: Training and test data for grass species (Imbalanced)…….21

Table 4: Training and test data for grass species (Balanced)…….22

Table 5: Confusion matrix for Sentinel-2 using Random Forest (Imbalanced)…….46

Table 6: Confusion matrix for Sentinel-2 using Support vector machines (Imbalanced)…….47

Table 7: Confusion matrix for Sentinel-2 using Random Forest (Balanced)……48

Table 8: Confusion matrix for Sentinel-2 using Support vector machines (Balanced)…….49

Table 9: Confusion matrix for RapidEye using Random Forest (Imbalanced)…….52

Table 10: Confusion matrix for RapidEye using Support vector machines (Imbalanced)……53

Table 11: Confusion matrix for RapidEye using Random Forest (Balanced)……...…54

Table 12: Confusion matrix for RapidEye using Support vector machines (Balanced)……55

Table 13: OA of RF and SVM for RapidEye and Sentinel2 images of both balanced and imbalanced data set……56

LIST OF ABBREVIATIONS

RS= Remote sensing

MSI= Multispectral instrument

ESA= European Space Agency

SVM= Support Vector machines

RF= Random Forest

GPS=Global Positioning System

SWIR= Short Wave Infrared

USGS= United States Geological Society

CHAPTER ONE

INTRODUCTION

1.1. General Introduction

There has been a tremendous need for land cover maps for the observation and sustenance of the earth's natural resources (Foley et al. 2005; Verburg et al. 2011; Hansen 2012). These maps are used in urban planning, land cover assessment and conservation (Wessels et al. 2003; Gebhardt et al. 2014; Fry et al. 2011). Grasslands are one of the world's most famous types of land cover vegetation (Latham et al. 2014). They play an essential role in plant biodiversity (Bergman et al. 2008; van Swaay, 2002) and spatial heterogeneity (MacFayden et al. 2016).

When mapping grass species, field-based methods have been used in the past at a local scale (Ramoelo et al. 2015). The main advantage of using field based methods is that they are useful when mapping vegetation species of a small area. However, studies have shown the field-based method for grass communities mapping is time-consuming, expensive and sometimes some areas were inaccessible leading to insufficient data (Ling et al. 2014; Kavzoglu and Colkesen, 2009a; Ramoelo et al. 2013). Remote sensing has proven to be a preferred and useful method in mapping vegetation because of its ability in discerning and observing the physical features of an area by assessing its reflected and emitted radiation at an extent from the targeted area (Mutanga, Adam and Cho, 2012). Satellite images help scientists and researchers to understand the earth better as these images allow them to see much more than they would if they were observing the surface from the ground. Remote sensing data offers a more precise alternative to field survey data, especially when dealing with cost and effectiveness. Remote sensing has shown to be very helpful in land cover mapping (Tucker et al. 1985), crop monitoring (Wu et al. 2015) and climate studies (Yang et al. 2013). This rise in interest is predominantly due to the current revolution in data, technologies and conjecture in urban remote sensing (Weng and Quattrochi, 2007; Yang et al. 2013; Salehi et al. 2012).

Multispectral remote sensing (Akasheh et al. 2008; Saatchi et al. 2007) and hyperspectral data (Lawrence et al. 2006; Peerbhay et al. 2013) have both been used in vegetation mapping. Multispectral data, such as SPOT and Landsat TM imagery are limited by their spatial and spectral resolution, which is ineffective in proper vegetation mapping because of its broad bands (Govender et al. 2008). Hyperspectral imagery, on the other hand, has narrow spectral bands

which makes it a more efficient method in mapping vegetation and in land cover use (Koch et al. 2005) as it can identify surface features at a higher spectral resolution. Hyperspectral data come with the challenge of data processing and analysis due to high dimensionality which can lead to inadequate classification and performance of the classification algorithm (Tsai et al. 2007; Kavzoglu and Mather, 2002). The arrival of some new generation sensors such as the Worldview 2 and 3, Sentinel 2 MSI as well as RapidEye have been used more recently for mapping grass communities and vegetation species in a more extensive area (Huang et al. 2017; Sibanda et al. 2017; Drusch et al. 2012).

Classification of remotely sensed images continues to be a difficult task. The set size of the training sample, the spatial resolution of the image, the diversity of vegetation class, attribute of the classification algorithm are some of the factors that have a considerable effect on the classification accuracy (Lu and Weng, 2007; Kavzoglu, 2009). Image classification has proven to be quite essential in remote sensing application. Hence, the importance of using advanced algorithm classifiers. A wrongly classified image can lead to information that is worthless and inadequate. It could also have an unfavourable effect if decisions are based on incorrect classification. Let us say, for example, that image classification was carried out on a satellite image where the grass was incorrectly classified as water. Such a mistake would prove detrimental in urban development or water management. Therefore, image classification plays a vital role in mapping and image interpretation (Li et al. 2014; Ma et al. 2017). Machine learning algorithms are productive and effective because they are not dependent on data scattering assumptions (e.g., Normality) and have positive accuracy (Foody, 1995a; Friedl and Brodley, 1997).

The design of the training samples is of importance. The training sets in many instances determine the quality of supervised classification (Smola and Scholkopf, 2003). In reality, though, classifiers are highly imbalanced or occur in unknown proportions. The spectral characteristics of remote sensing data provide a lot of distinguishing and decisive factors such as near-infrared band or vegetation indices for the plants, forestry and agricultural utilization (Kim and Yeom, 2015). Traditional learning methods are intended principally for balanced samples. A balanced sample has uniformity of classes across the class distribution. When algorithms are

used for imbalanced samples, there tends to be over predicting the appearance of the majority class (Wei and Dunbrack, 2013). A balanced sample is believed to boost overall classification in contrast to an imbalanced sample (He, 2011; Laurikkala, 2001). The characteristics and quality of the training samples are essential in classification which directly impacts classification accuracy (Foody, 1999; Ustuner et al. 2016). Errors, such as interpretation problems and poor quality of training data sets can affect accuracy. The set size of a training sample is essential when classifying minor classes of interest (Ustuner et al. 2016). In some cases, the training sample of one class could differ from another class. This is known as imbalanced training samples. This imbalance leads to low accuracy for minor classes (Foody et al. 2006).

Image classification methods, using remotely sensed data is generally used when mapping grass species. The option of the suitable remotely sensed data in terms of the price and the resolutions and the choice of suitable classification process are critical for valid, accurate vegetation mapping (Adam et al. 2014). There are various types of machine learning algorithm, and the model used is dependent on the user's familiarity with the algorithm and what the user wants to achieve. When it comes to vegetation mapping and remote sensing in general, the two most commonly used are the support vector machine (SVM) and random forest (RF) (Clark et al. 2016; Mountrakis et al. 2011). SVM is a simplified algorithm used when dealing with imbalanced dataset because it handles high dimensionality which is a problem when processing small training samples and the need to achieve high accuracy (Melgani and Bruzzone, 2004; Foody et al. 2006). SVM is greatly reliant on the training sample size (Schohn and Cohn, 2000). Random forest (RF) involves re-sampling the original training samples to increase accuracy and stability. Rodriguez-Galiano et al. (2012), and Pal (2005) is of the opinion that random forest is more robust when it comes to variation in data. Studies have shown that a balanced sample improves overall classification in cases like SVM and RF in comparison to the imbalanced sample (Estabrooks et al. 2004; Weiss and Provost, 2003).

This study looks at the effects, if any, of a balanced and imbalanced dataset on high-resolution images, RapidEye and Sentinel 2, using SVM and RF classifier.

1.2. Problem Statement

Field-based methods are commonly being used to collect data on varying grass species in the Telperion Game Reserve. This is a tedious and time-consuming method that can lead to inaccuracy in classifying the various grass communities, as access to some areas might be a problem. For proper management of the reserve, reliable, current and comprehensive spatial information on the biodiversity of the area is essential (Adam et al. 2010). Remote sensing provides vital information on grass community and its distribution (Darvishzadeh et al. 2008). High-resolution imagery like Sentinel 2 and RapidEye are preferred in research in land cover and vegetation mapping due to global coverage and free access. It is not just the image selected that is important, but also the classification method used as this affects the results of the land cover maps (Lu and Weng, 2007). Of the machine learning-based algorithms, RF and SVM are becoming popular in image classification research (Adam et al. 2014) primarily due to their insensitivity to overtraining and noise, making them better suited to deal with imbalanced data (Breiman, 2001). The design and selection of training samples are significant in the learning stage of a classifier (Ustuner et al. 2016). It is always best to use a balanced sample when dealing with machine learning algorithms (Weiss and Provost, 2003; Japkowicz and Stephen, 2002). Most times the method used to balance the samples depends on the researcher and what they are trying to achieve (Chawla et al. 2002; Chen et al. 2004; Trebar and Steele, 2008). It is believed that a substantial quantity of training samples is vital for image classification and is collected as ground truth data from the field (Hubert-Moy et al. 2002; Mather, 2004). When dealing with high resolution images, a good number of samples are needed because of high sample variation (Tsai et al. 2007; Borges et al. 2007). There is the need to find the optimum number of samples needed for higher spatial resolutions regarding the number of samples and balanced and imbalanced samples across different satellite images and classification algorithm. In the past, studies have tested imbalanced and balanced training sample in individual machine learning classifiers such as RF and SVM (Mellor et al. 2014; Ustuner et al. 2016). Only a finite amount of research has been carried out to compare different classifiers in different high-resolution images using both balanced and imbalanced datasets and its effect if any on the overall accuracy.

1.3. Aims and objectives

This research aims to investigate the impacts of balanced and imbalanced samples on the accuracy of grass community mapping using different machine learning classifiers, and high resolution multispectral remotely sensed data.

The specific objectives are to,

- Map different grass communities in the Telperion Game Reserve using RapidEye and Sentinel 2 images and machine learning classifiers (Support vector machine and Random forest).

- To quantify and analyse the impacts of balanced and imbalanced training data on the performance and the accuracy of Support Vector Machine and Random forest in mapping the grass communities.

- To test the sensitivities of pixel resolution to balanced and imbalance training data in image classification.

CHAPTER TWO

LITERATURE REVIEW

2.1. Mapping grass communities using remote sensing

Vegetation mapping analysis has become predominant in recent years (Cingolani et al. 2004) because they help in differentiating grass species and ecology in an area which leads to valuable information in conservation management (Zhang et al. 2016) and management practices. The traditional field method technique of mapping vegetation is a tedious task used in gaining knowledge about species type and their makeup (Terri and Stowe 1976). This method needs intensive fieldwork and laboratory analysis to measure the biochemical and biophysical properties of the grass species (Mutanga et al. 2003). The intense nature of fieldwork leads to results that are not fully representative of plant population and its distribution, especially in areas of varied diversity (Mutanga et al. 2003). The use of field data alone is insufficient as current and accurate information is required in a proper land cover and vegetation mapping, especially for areas of diverse landscapes (Odindi et al. 2016).

Remote sensing provides an alternative and economical way of analysing grass species as it reduces the field work and the laboratory analysis required by the traditional method. The use of remote sensing has helped in providing information of even the most inaccessible areas at a cost-effective rate (Running et al. 1993; Darvishzadeh et al. 2008). Remotely sensed data has been used in discriminating grassland species (Baldi et al. 2006; Toivonen and Luoto, 2003; Wang et al. 2010). Recent studies in mapping and monitoring vegetation species have incorporated the use of low and medium resolution imagery such as Landsat (Wulder et al. 2008; Vogelman et al. 1998; Giri et al. 2003), SPOT (Kanellopoulos et al. 1992; Chen, Franklin and Spies, 1992) and MODIS (Stefanov and Netzband 2005). The accuracy of using these types of imagery is compromised by their spectral and spatial resolution (Foody 2002). The introduction of multispectral and hyperspectral imaging has dramatically improved the accuracy of vegetation mapping worldwide as they have high spectral and spatial resolution (Mutanga, Adam and Cho, 2012; Akasheh et al. 2008; Harvey and Hill, 2001; Lawrence et al. 2006). New generation imagery such as Worldview 2&3, Sentinel 2 MSI, and RapidEye has emerged recently. Their spectral bands which fall in the electromagnetic spectrum, such as red edge provides a more detailed classification of landscapes (Schuster et al. 2012; Cho et al. 2012; Mutanga, Adam and Cho, 2012). While these new multispectral sensors advantageously provide significant details in

mapping vegetation (Baumstark et al. 2016; Odindi et al. 2014; Omer et al. 2015), acquiring the data is expensive. Image analysis, through the use of vegetation indices, is a standard way in remote sensing for discerning spatial patterns of the distribution of vegetation (Adjorlolo et al. 2012). Remote sensing can also be used to distinguish between local grassland communities, grasslands and frequently co-occurring vegetation species. This is done by comparing classification results from different imagery dataset (Melville et al. 2018). The selection of the appropriate sensor is vital for vegetation mapping and land cover. Low-resolution images are commonly used in the large scale mapping of the identification of a substantial number of vegetation classes while a higher resolution image is used for superior classification of vegetation at a smaller scale. High-quality ground truth data is needed in remote sensing for cross-validation and training algorithm. To this effect, remote sensing is a potent tool when used concurrently with ground truth data (Bredenkamp et al. 1998).

2.1.1. Importance and principle of mapping grass communities

Monitoring land cover is essential for global change investigation (Jung et al. 2006; Lambin et al. 2001). Proper mapping of grass and vegetation species is crucial in managing the earth's natural resources as vegetation supplies a foundation for all living beings (Xiao et al. 2004). Vegetation mapping also includes details about natural and human-made habitat by quantifying vegetation cover at a small or large scale either presently or over an extended period of time (Xie et al. 2008). For proper conservation, it is crucial to obtain new generation cover (Egbert et al. 2002; He et al. 2005). The principle of vegetation mapping using remote sensing, relies on the spectral attribute of the vegetation species and their spectral reflectance and radiance.

2.2. Mapping grass communities using multispectral remote sensing

Multispectral data have been used in vegetation mapping on many occasions (Rignot et al. 1997; Harvey and Hill 2001; Chastain, et al. 2008; Martinez-Lopez et al. 2014). In multispectral imagery, the pixels lead to a mix of vegetation species in varying proportion (Zomer et al. 2009). This mixing is primarily because multispectral sensors give rise to three to six spectral bands spanning from visible to near-infrared of the electromagnetic spectrum (Jensen, 2007). This mixing effect is a significant disadvantage in mapping vegetation. Mansour et al. (2016) used

multispectral remote sensing for mapping grassland degradation. Huang and Siegert (2006) found that SPOT VGT imagery was useful in detecting environmental changes on a larger scale and SPOT images were used to produce vegetation maps in Eastern New Zealand (Mathieu et al. 2006). Zheng et al. (2006) used Landsat TM images to analyse wetland landscape patterns on the Minjiang River. Landsat images are one of the more common types of low to medium resolution images used. Wang et al. (2007) used Quickbird-2 to map aquatic and terrestrial vegetation. Multispectral data were also used for global mapping at a continental scale to map land cover in Central Africa using AVHRR (Mayaux et al. 1998).

Although mapping vegetation using multispectral remote sensing has been promising, there are limitations due to its lower spatial and spectral resolutions, especially when dealing with complex and diverse vegetation types (Adam et al. 2012; Feng et al. 2015).

2.3. Mapping grass community using hyperspectral remote sensing

Hyperspectral remotely sensed data records a large quantity of narrow wavelength bands (over 200) from the visible, near infrared, mid-infrared to the shortwave infrared bands of the electromagnetic spectrum. These bands offer new vegetation index for specific species (Clevers et al. 2007) making it more efficient in vegetation mapping. An advantage of this type of imagery is that the mixed pixel problem seen in multispectral imaging is significantly reduced, providing more information on land cover (Lu and Weng, 2009). Mutanga and Skidmore (2004) using hyperspectral data deduced that the narrowband indices provided better information on grassland biomass. Some researchers have focused on vegetation density (Nichol and Lee, 2005; Small, 2003) while others focused on the creation of land use/land cover maps (Carleer and Wolff 2006; Herold et al. 2003). Vegetation species classified as Invasive species have been successfully mapped using hyperspectral imagery because of its ability in determining the percentage coverage of vegetation species (Mundt al. 2005; Williams and Hunt, 2004; Glenn et al. 2005; Lawrence et al. 2006)

A disadvantage of this though is the problem caused by shadows (Asner and Warner, 2003; Zhou et al. 2008; Lu and Weng, 2009). These shadows can lead to lower accuracy if a suitable

classification algorithm and image processing method is not used (Irons et al. 1985; Cushnie, 1987). This problem was examined recently (Zhou et al. 2008; Mathieu et al. 2007; Walter, 2004; Zhang et al. 2003). High spectral variation is also a problem when dealing with similar land cover types. Object-oriented classification methods have reduced this problem significantly (Mathieu et al. 20007; Zhou et al. 2008; Stow et al. 2007; Jacquin et al. 2008; Laliberte et al. 2004). Another setback of hyperspectral data is that they are expensive (Sibanda et al. 2017).

2.4. Mapping grass communities using new advanced multispectral data

The arrival of new multispectral sensors has been recognized as an improvement from the shortfalls of hyperspectral and multispectral imagery (Mutanga et al. 2012). The higher spatial resolution and extended amount of bands such as the red edge, are preferred for vegetation mapping at higher accuracies (Mansour and Mutanga, 2012; Adelabu, Mutanga and Adam, 2015). RapidEye and WorldView-2&3 imagery is used in various vegetation mapping research (Ustuner et al. 2016; Adam et al. 2014; Luck-Vogel et al. 2016; Adam et al. 2017). Mansour and Mutanga (2012) used WorldView-2 data to map grassland degradation of grass species in South Africa with an overall accuracy of 90%. The addition of a red-edge band helps in discerning variations in vegetation which makes for improved vegetation mapping. Despite these many advantages, these images are expensive. The availability of Sentinel-2 MSI, which also has high spatial and spectral properties, has helped in this aspect as it can be acquired free of charge.

High spectral resolution does not often translate to improved accuracy; thus more advanced and robust classifiers are required (Sesnie et al. 2010; Adelabu et al. 2015; Lawrence et al. 2006). These include classifiers such as SVM and RF.

2.5. Ground and training sampling for mapping grass communities

While remotely sensed data is vital for proper vegetation mapping, ground truth data are equally essential for cross-validation when dealing with remote sensing data (Odindi et al. 2016; Bredenkamp et al. 1998). If the number of samples from the field is not enough, sometimes points are digitized based on proper georeferencing. Sometimes, vectors are manually digitized

(Evans et al. 2012; Al-Mashreki et al. 2010). Millard and Richardson (2015) suggested that the training and testing data should be as numerous as possible.

Hubert-Moy et al. (2001) and Mather, (2004) believed that an adequate amount of training samples is vital for the classification of images and are collected as ground truth data from the field. The training data size is thought to affect the accuracy of classification performance (Mellor et al. 2015; Millard and Richardson, 2015). A substantial amount of training and testing data is believed to be needed to assess the classification accuracy entirely (Jin et al. 2014). Sometimes, the distribution of test sample may be different from that of the training sample, and the actual effects of this miscalculation might not be realized at the learning stage. In recent years, studies have shown the importance of a balanced sample over an imbalanced sample (Japkowicz and Stephen, 2002; Wei and Dunbrack, 2003; Estabrooks et al., 2004; Weiss and Provost, 2003). In a real-world scenario, imbalanced training samples occur due to difficulty obtaining the ground sample for some areas. Mellor (2017) showed that deliberately imbalancing a dataset can improve classification and performance of some classes without undermining overall classification outcome.

Some studies have shown that using an imbalanced sample can lead to low classification accuracy (Kubat and Matwin, 1997; Japkowicz, 2000). Huang et al., (2002) believed that the training data used, affects the classification accuracy more than the classification algorithm used and suggested an increase in the training sample size to improve classification accuracy.

2.6. Machine learning classifiers for mapping grass communities

Machine learning algorithms are a more accurate type of classification algorithm when dealing with extensive data (Muchoney and Williamson, 2001; Kasischke et al. 2004). They can deal with noisy and missing data, especially classification trees (Simand et al. 2000; Hastie et al. 2001). Different researchers in the past few years have compared various classification algorithms in vegetation mapping. These include maximum likelihood (Stuart et al. 2006) decision trees (Wang et al. 2016), random forest (Vanselow and Samimi, 2014), support vector machines (Schwieder et al. 2016) to neural networks (Zhang and Xie, 2012). Most recently

support vector machines and the random forest has been the most commonly used of the classifiers.

Adam et al. (2016) and Pal, (2015) agreed that SVM and RF performed equally well based on high overall accuracy while others showed SVM performed well with a balanced and imbalanced dataset. Ustuner, Sanli and Abdikan (2016) looked at the mapping of diverse vegetation in Aydin, Turkey by classifying a RapidEye imagery using balanced and imbalanced training sample. They used SVM, Maximum Likelihood (ML) and Artificial Neural Network (ANN) classifications for mapping the crop pattern in the area and concluded that SVM was unaffected, showing SVM is an efficient and consistent classifier irrespective of whether it is a balanced or imbalanced training sample. In this case, it is proven that Support vector machine will be a capable and useful classifier. The result further highlighted why the design and choice of training sample into the learning stage of supervised classifiers is so important which is an integral part of this research.

While different algorithms have been used to solve the problem of an imbalanced data sample, random forest and support vector machine have been the most effective. Other methods like weighting and undersampling have been used in the classification of an imbalanced data sample. One such study was that of Anand et al. (2010), who looked at the classification of an imbalanced sample using weighting and undersampling.

There have been a few studies that have been carried out on balanced and imbalanced data samples and the effects of this imbalanced sample set on the overall accuracy and result, but not many have compared this with different high-resolution imagery, hence the focus of this study. The study will look at balanced and imbalanced samples from Sentinel-2 and RapidEye imagery using different algorithms, SVM and RF to determine the factors that affect these sample and if the imaging affects the accuracy of the samples.

2.7. Conclusion

Vegetation mapping of diverse grass community class has been done successfully in prior years using different remote sensing imagery both on a local and global scale. A variety of low-medium-high resolution imagery has been used in the past (Mutanga et al. 2016). The new generation multispectral imagery with higher spectral and spatial resolution are being preferred when it comes to mapping vegetation because they possess the red edge band which best classifies these different species. Classification algorithms such as maximum likelihood, k-means, and minimum distance have been used in the past for classification, but the introduction of newer algorithms such as ANN, SVM, and RF is being used more frequently in the present. Only a few researchers have used these advanced classification algorithms on the new generation multispectral imagery using different (balanced and imbalanced) data set.

CHAPTER THREE

METHODOLOGY

3.1. Study area

The study area is Telperion Nature Reserve (25° 41′ S, 28° 56′ E) depicted in Figure 1 below. The area is approximately 11 000 hectares. Telperion is a section of the more magnificent eZemvelo Nature Reserve, located in Mpumalanga Province of South Africa. The reserve is situated at the border between Gauteng and Mpumalanga Provinces. The Wilge River, which is a tributary of the Oliphant's River, flows northwards through the reserve. The reserve was established in 2008 and is surrounded by farmlands with people practicing agriculture, specifically maize and sunflower, and cattle rearing. Temperatures range from 14^0C and 26^0C during summer and 4^0C and 17^0C during winter. Dry winters are experienced here, which makes it difficult for tree growth and ultimately to the death of grassland. There is a high diversity of flowering plants and grass species. Telperion reserve is characterized by highlands and undulating terrain of ridges and valleys. Only 2% of Telperion are officially under conservation. There is a high diversity of animal and birdlife. The Oppenheimer family has owned Telperion for over 40 years.

Mucina and Rutherford (2006) classified the vegetation type as the Bankenveld and Mixed Bushveld. The dominant grass species identified in the sampled plot are *Eragrostis gummiflua, Hyparrhenia hirta, Cynodon dactylon, mixed grassland, Eragrostis chloromelas, woody vegetation, wetland grass, Aristida congesta* and the *Alien Invasive Species*.

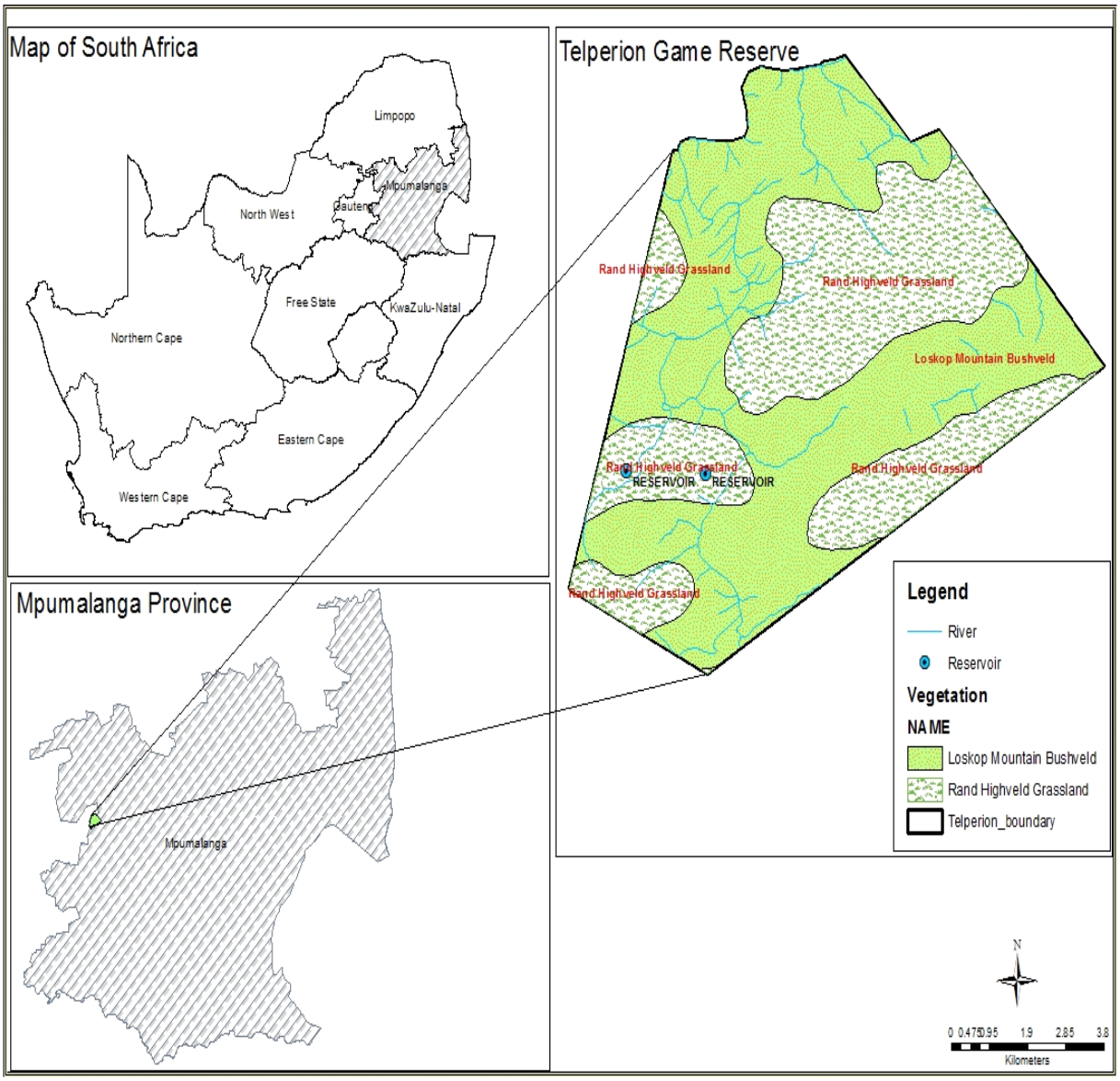


Fig 1: The location of Telperion Nature Reserve

3.2. Remote sensing data acquisition and pre-processing

3.2.1. Sentinel-2 Multispectral Instrument (MSI) image acquisition

The Sentinel-2 MSI of the study area was downloaded from the European Space Agency's website (earthexplorer.usgs.gov) on the 20th of May 2016. Sentinel-2 is a high-resolution multispectral image that was first launched on 23 June 2015. It consists of two similar satellites, Sentinel 2A and Sentinel 2B which helps in frequent revisit every five days under the same viewing angle. Although there can be overlap and some regions will be observed more than once every five days with differing views. It has a spatial resolution of 60m, 20m, and 10m. It consists of thirteen spectral bands with four bands of 10m, six bands of 20m and three bands of 60m (Table 1). It has a 290 kilometre (km) field of view. It was launched to observe natural disaster management, land cover change detection and for other monitoring on the earth's surface.

Table 1. Spectral bands of Sentinel 2A satellite imagery

Sentinel 2 bands	**Centre wavelength(nm)**	**Bandwidth(nm)**	**Spatial resolution(m)**
Band 1	443.9	27	60
Band 2	496.6	98	10
Band 3	560.0	45	10
Band 4	664.5	38	10
Band 5	703.9	19	20
Band 6	740.2	18	20
Band 7	782.5	28	20
Band 8	835.1	145	10
Band 8a	864.8	33	20
Band 9	945.0	26	60

Band 10	1373.5	75	60
Band 11	1613.7	143	20
Band 12	2202.4	242	20

3.2.2. RapidEye image acquisition

A RapidEye imagery map was downloaded from the RapidEye satellite constellation website (rapideye.net/upload) on 21st of May 2016. RapidEye is a high-resolution imagery with multispectral capabilities. It was launched on 29 August 2008 and provides broad area coverage and frequent revisit intervals. RapidEye collects 4 million square kilometres of data per day at 6.5 m ground resolution that can be re-sampled to 5m pixel size. It consists of five satellites equipped with identical sensors located in the same orbital plane. It is capable of daily revisits when off-nadir and revisiting every 5.5days at nadir with a swath width of 77 kilometres (km). It can be used in various fields, including mining, oil and gas exploration, security and emergency, mapping and agriculture. RapidEye imagery is in high demand for land use/land cover maps and mapping vegetation due to its red edge and NIR bands which are sensitive to the chlorophyll content in vegetation. The RapidEye sensors produce imagery in five spectral bands that can be seen in table 2.

Table 2: Spectral bands of RapidEye imagery

Spectral bands	**Wavelength(nm)**
Blue	440-510
Green	520-590
Red	630-685
Red edge	690-730
NIR	760-850

3.3. Remote sensing data pre-processing

Georeferencing is extremely important because of the use of different images with different types of the classification algorithm. For the Sentinel-2 MSI image, this was done by selecting the UTM zone 35S with the 10m spatial resolution used. Atmospheric correction was done on both images using the Sen2cor tool which is available in the Sentinel Application Platform (SNAP) toolbox and performed using a python script. The corrected image was then converted to ENVI format, resulting in 10 bands (2-8, 8a, 11 and 12). The converted bands were displayed on ENVI 5.3. Thereafter, the spectral reflectance from the Sentinel 2 MSI image that corresponds to each GPS sampled point was derived.

For the RapidEye imagery, georeferencing was done by selecting UTM zone 35S with a spatial resolution of 5m. The digitized image was in level 3A (orthoproduct) in which radiometric, sensor and geometric corrections have been implemented for the data. The corrected image was displayed on ENVI 5.3. After this was done, the spectral reflectance from the RapidEye image that corresponds to each GPS sampled point was extracted for further analysis. The image was also used in R studio.

3.4. Field data collection

Field data collection was done to locate the different vegetation species in the game reserve. The field sampling was carried out between the 22nd -24th of May 2016, which was consistent with the window period the images were acquired. The sample plots were randomly fixed and spread evenly across the study area (Ramoelo et al., 2012) with the plots being 10 metres x 10 metres in size to account for the pixel size of the sentinel image (10m). It was the dry season at the time the sample was collected. Global Positioning System (GPS) was used to record the coordinates where each of the samples was obtained and also the coordinate of each sample plot with a total of 80 GPS points recorded. The sample collected from each imagery was split into training and testing data using the typical 70:30 split in R studio and ENVI 5.3 respectively for classification. A look at the samples collected shows that they are imbalanced.

Table 3: Training and test data for the grass species (imbalanced).

Species	**Training samples (70%)**	**Test samples (30%)**	**Total samples**
Alien Invasive Species	48	20	68
Hyparrhenia hirta	41	17	58
Mixed grassland	201	87	288
Cynodon dactylon	35	14	49
Eragrostis chloromelas	52	22	74
Woody vegetation	95	40	135
Wetland grass	42	18	60
Aristida congesta	42	17	59
Eragrostis gummiflua	45	19	64

To balance out the imbalanced dataset, a random undersampling method was carried out to even the distribution by randomly reducing the quantity of majority samples while keeping the total of the lowest minority sample in mind and building a more balanced number of samples from that. The balanced data can be seen in table 4 below.

Table 4: Training and test data for the grass species (balanced).

Species	**Training samples (70%)**	**Test samples (30%)**	**Total samples**
Alien Invasive Species	35	14	49
Hyparrhenia hirta	35	14	49
Mixed grassland	35	14	49
Cynodon dactylon	35	14	49
Eragrostis chloromelas	35	14	49
Woody vegetation	35	14	49
Wetland grass	35	14	49
Aristida congesta	35	14	49
Eragrostis gummiflua	35	14	49

3.5. Image classification

When it comes to remote sensing, the production of land use and land cover maps is an essential function carried out through image classification (Al-doski et al., 2013). Machine learning algorithms such as SVM and ANN has been tested and examined numerous times in remote sensing, from optical to radar data, for image classification in the current years (Pal et al., 2013). Several studies have shown the superiority of SVM and RF in comparison to other types of classification when dealing with remote sensing images and land cover analysis (Adam et al., 2014; Khatami et al., 2016; Qian et al., 2015; Shao and Lunette, 2012). The frequent use of

these two algorithms is why this research focuses on support vector machine and random forest classification algorithms.

3.5.1. Support Vector Machines

Support vector machine is a type of supervised machine learning algorithm that is used for both classification and regression analysis. The concept of SVM is that it creates differing hyperplanes that separate the dataset into a predefined number of classes. The separation is done by using a training sample which is a subset of the dataset. Support vector machines are a powerful kernel-based classification algorithm. Kernel function needs user-defined parameters. Vladimir N. Vapnik invented the original SVM algorithm in 1963. It has since become very popular and has been successful in remote sensing classification. The main reason for SVM's popularity is its high classification accuracy with a small quantity of training data and outperforms other conventional methods like maximum likelihood (Huang et al., 2002). Mountrakis et al. (2011) analysed articles from over a hundred sources and did an overview of the results using SVM as the selected choice of classification and concluded of its high accuracy when dealing with a small training sample and its superiority compared to the other types of classification but its limitations in parameter selection. Camps-Valls et al. (2004) reported SVM's advantage when dealing with hyperspectral remotely sensed data. Although in theory SVM is known for high classification accuracy, it is not as effective when using a significant data because its training difficulty relies heavily on the size of the dataset.

SVM can perform linear classification as well as nonlinear classification, also known as the kernel function. The kernel function transforms the data and then finds the optimal boundary for the outputs. A linear classifier separates points into one of two classes by a straight line, the goal of which is to see a line that passes as far as possible from all aspects to avoid noise (figure 2).

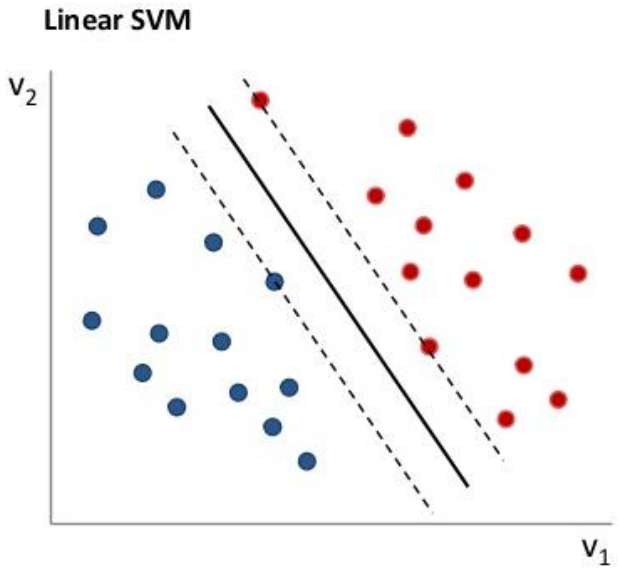


Figure 2: linear classifier (source: Vapnik, 1999 Springer; Scholkopf et al., 2002)

Although in real life, most classification tasks are never really this simple as optimal separation would require a more complex structure than that of a straight line as can be seen in the image below. This type of classifier is known as the nonlinear classifier. Hyperplane classifiers are lines drawn to distinguish and separate objects of a different class. This separation is where SVM thrives. SVM is represented by the formula below:

$$w^T x + b = 0$$

Where w is a weight vector

x is input vector

b is bias

The formula also allows us to write the parallel hyperplane (Burgess, 1998)

$$w^T x + b \geq 0 \text{ for } d_i = +1 \text{ (plus plane)}$$

$w^T x + b < 0$ *for* $d_i = -1$ *(minus plane)*

Where d is the margin of separation (separation between hyperplane and the closest data point for a given w, weight vector and b, bias parameter.

In the figure 3 below, a curve like a backward c, would have to be created to separate the two classes properly.

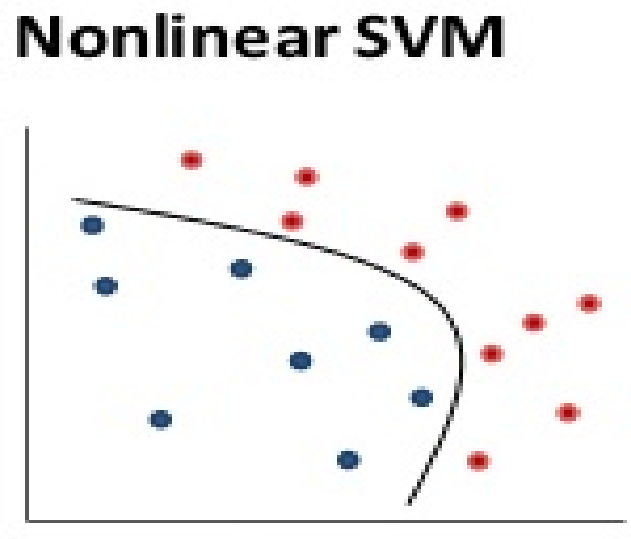


Figure 3: Nonlinear classifier (source: Vapnik, 1999 Springer; Scholkopf et al., 2002)

Figure 4 represents a nonlinear surface where the data would have to be mapped in a higher dimensional feature space through the kernel function, making them linearly separable in this space since there is no possibility to do so in the original area. When it comes to SVM, the most common kernels are linear, Gaussian radial basis function (RBF), polynomial, and sigmoid kennels which were presented by Fletcher (2009) and Haykin (1999). In remote sensing data analysis, the RBF kernel is the most widely used kernel functions due to its high performance (Gomez-Chova et al., 2011).

$$K(x, x_i) = exp\left(-\frac{\|x - x_i\|^2}{\gamma^2}\right), \gamma > 0$$

Where χ is a sample in the data space

χ_i is a corresponding sample in the feature space

γ is the kernel parameter

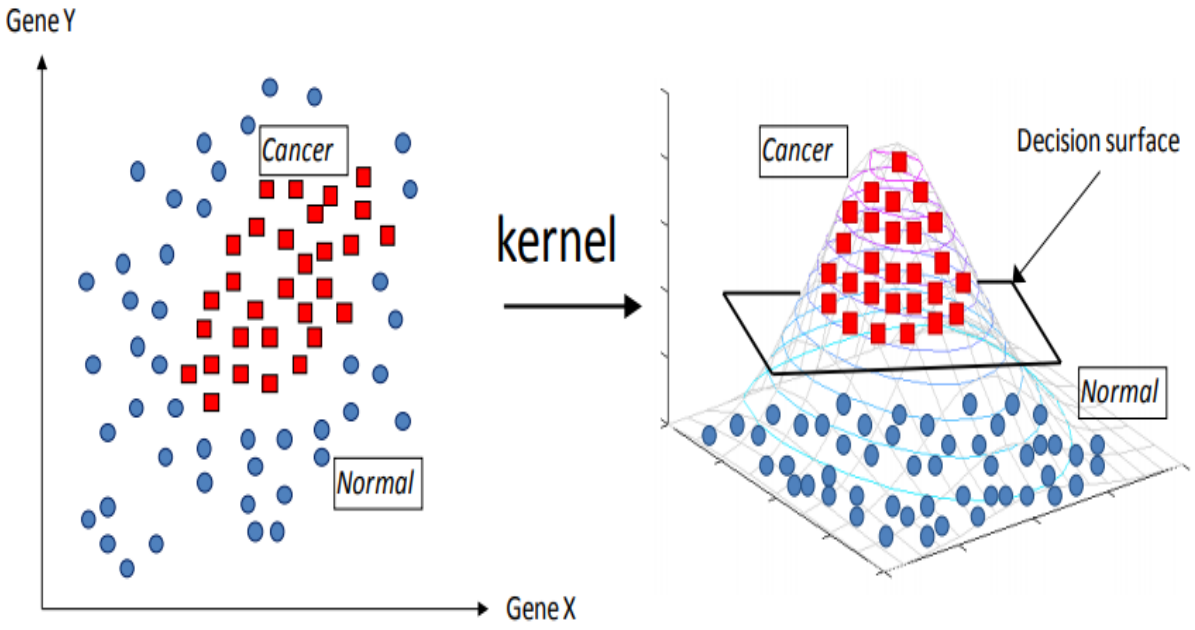


Figure 4: The Main idea of SVM (source: Statnikov et al., 2011)

A disadvantage of SVM is that it will classify all examples as the majority class, a tactic that if the imbalance is severe, can provide the minimal error rate across the data space (Batuwita, R. and Palade, V., 2012). There have been many works of literature that apply different techniques to the SVM framework to overcome problems due to imbalance (Wu and Chang, 2003). There are various ways of mapping non-linear boundary with kernel functions in SVM which includes linear, polynomial, radial basis function and sigmoid kernels. SVM was run using the support

vector classification tool in ENVI 5.3 applying radial basis function (RBF) kernel which is the most commonly used kernel function when dealing with SVM (Pal, Mather 2005; Melgani, Bruzzone 2004; Hermes et al. 1999). The RBF has two tuning parameters- cost (*C*) and gamma (γ), which can affect overall accuracy (Burges, 1998). The ENVI 5.3 software uses the pairwise classification strategy for multiclass classification. The software carries out classification by selecting the highest probability and a threshold is set, with pixels below this threshold deemed unclassified. Support vector is the interval measured between the nearest points of the two classes (Pal and Mather, 2005). The regions of interest (ROIs) were created by overlaying the dataset on the Sentinel 2 and RapidEye images in ENVI. Once the ROIs were generated, the SVM classification began, after which the training dataset (70%) was used for accuracy assessment. SVM was run for both balanced and imbalanced datasets of RapidEye and Sentinel 2 images. SVM was also run on R using a python script to get the parameter tuning and check the results.

3.5.2. Random forest classifiers

Random Forest, developed by Breiman (2001), is a type of supervised classification algorithm. Random forest is based on tree classifiers. In this classifier, the number of decision trees makes the forest. Figure 5 below shows the main idea behind random forest classifiers.

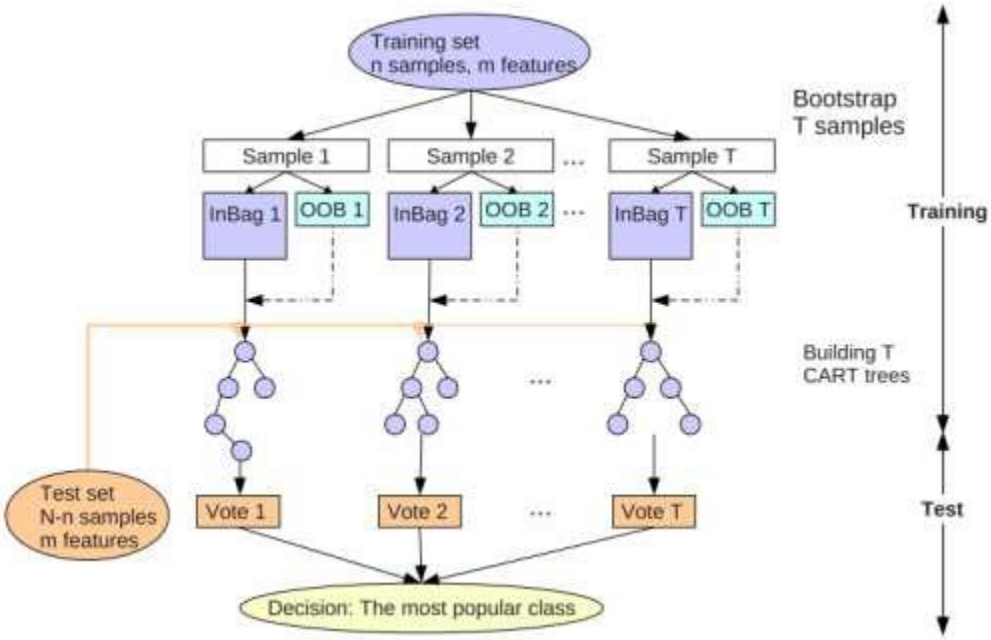


Figure 5: Workflow and main idea of RF (source: Guo et al., 2011)

The random forest classifier uses a set of classification and regression tree, CARTs, to make a prediction (Breiman, 2001). The trees are created through a process known as bagging. Bagging is a method whereby trees are formed by drawing a subset of training samples through replacement. Two-thirds of the samples (referred to as in-bag samples) are used to train the trees while the one third that is left over (referred to as out-of-the-bag samples, OOB) is used for internal cross-validation which helps us estimate how well the RF model performs (Breiman, 2001). There is no pruning when decision trees are produced. The final classification output is created based on a majority vote of the predictions from all individually trained trees (Jin, 2012). The more the trees in the forest, the better the random forest classifier will be. The higher the number of decision trees, the higher the accuracy of the classifier. The decision tree algorithm comes up with a set of rules based on the training data sets. This set of rules is also used on the

testing datasets. Random Forest can be used when handling classification and regression issues. It operates by constructing a high number of decision trees at the time of training and producing the class that is the mode of the classes (classification) or average prediction (regression) of the individual trees. For RF to be implemented, the user-defined number of trees (*ntree*) and the user-defined number of features (*mtry*) must first be set up. The algorithm then creates trees that have a high variance of low bias (Breiman, 2001). The mean average from the trees gives us the predictions of the random forest classification. The RF prediction is described by the formula below

Random forest prediction $s = \frac{l}{k} \sum_{k-1}^{k} k^{th}$

Where the index *k* runs over the individual trees in the forest

Random decision forests correct for the decision trees' habit of overfitting to their training set. It is known for being efficient in its implementation on large datasets and its accuracy among current algorithms. It works well with missing data by replacing missing values. This is done by computing the median of all values in the class. It then uses these average values to substitute all the missing values with rough estimates or by doing a raw filling of the missing values by computing proximity. Its accuracy is not affected by this. This approach provides a way of estimating the importance of the individual variables in classification.

One thing about the RF algorithm is that there are a few assumptions involved which lead to faster results and outputs. These assumptions are based on RF creating many decision trees which help in improving the accuracy. RF can rank variables based on the importance of running the mean decrease accuracy table if the user needs further analysis. As earlier stated, the number of trees (*ntree*) and the number of features in each split (*mtry*) first have to be set up before RF can be carried out. This optimization was carried out four times- one for each imbalanced RapidEye and Sentinel 2 images and one for each balanced RapidEye and Sentinel 2 images. RF was done in RStudio using a python script. RStudio is an open source tool that supports different geospatial analysis of remotely sensed data. Each run produced different *mtry* and *ntree* values.

3.6. Accuracy assessment

When it comes to accuracy of classification performance, overall accuracy and kappa coefficient are the most common. Overall accuracy (OA) is the ratio of the number of correctly classified samples (sum of principal diagonal) and the total number of sample units (Congalton and Green, 2009). Accuracy assessment which is an integral part of any classification will be carried out using 30% of the subset of the referenced data. The assessment was done for both images and the balanced and imbalanced dataset. A confusion matrix was also generated which shows predicted class versus actual class. In other words, it shines a light on the errors made by the classifier. Since the research is not based on looking individually at each grass community species, the OA will be used for comparison.

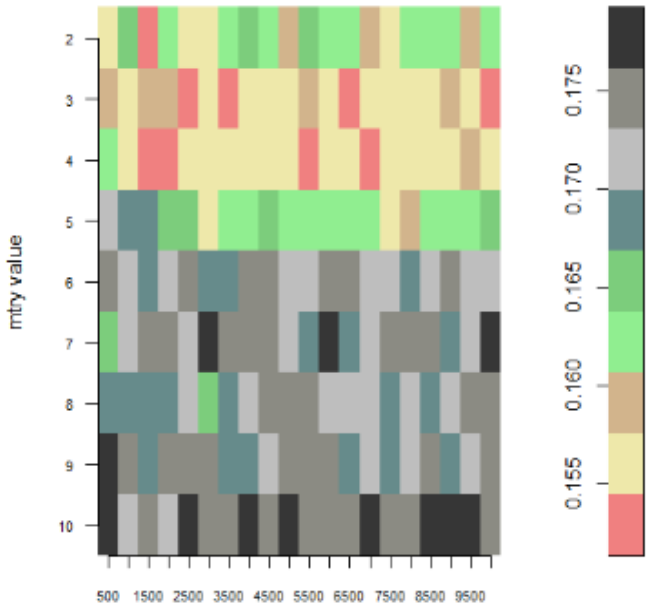
CHAPTER FOUR

RESULTS

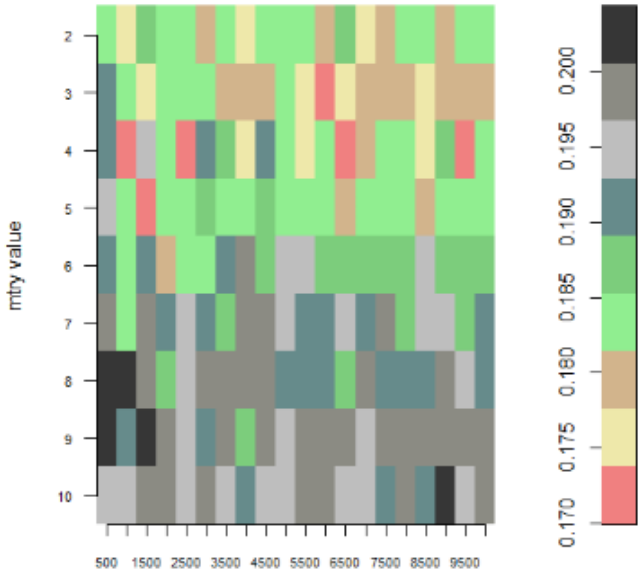
4.1. Optimization of RF parameters

4.1.1. Sentinel-2 MSI imagery

As previously stated, the *ntree* and *mtry* can affect the performance of a RF classifier. The *ntree* and *mtry* values are represented in grids. For the imbalanced dataset, the *mtry* value of 4 and *ntree* value of 2000 created the least OBB error rate of 0.1528962. The highest OBB error rate of 0.177707 was produced by the *mtry* value of 7 and *ntree* value of 3000 (Figure 6a). Subsequently, the *mtry* value of 4 and *ntree* of 2000 was chosen as the input parameters which will be used to train the RF algorithm for the classification of the grass community. For the balanced dataset, an *mtry* of 4 and *ntree* value of 1000 produced the least OBB error rate of 0.171839. The highest OBB error rate of 0.202414 was created with an *mtry* value of 9 and *ntree* value of 500 (Figure 6b). Therefore, the *mtry* of 4 and *ntree* of 1000 was chosen as the input parameters required to train the RF algorithm in order for a classification of the grass community for the balanced dataset.



(a)

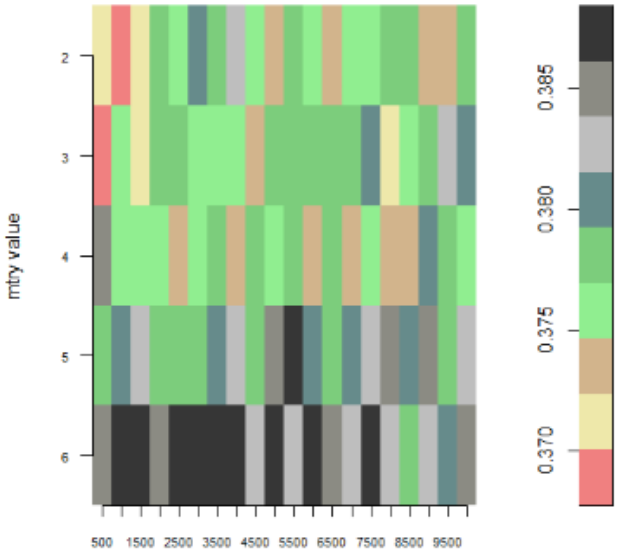


(b)

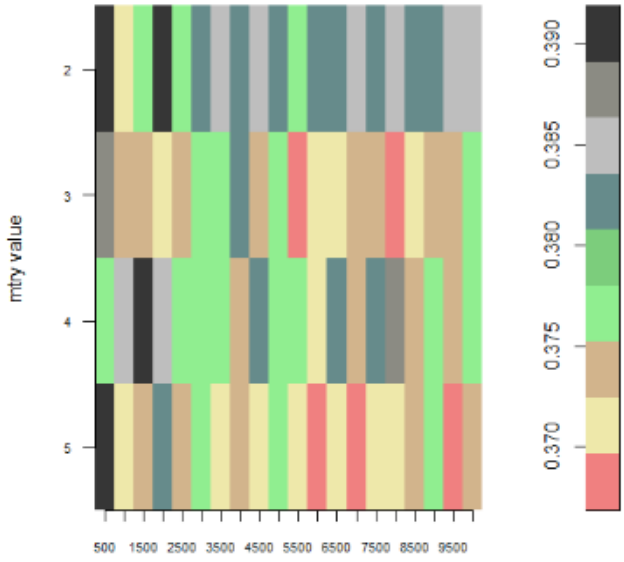
Figure 6: RF parameter optimization (*mtry* and *ntree*) for the imbalanced (a) and balanced (b) data using the grid search procedure.

4.1.2. RapidEye imagery

For the imbalanced RapidEye imagery, the *mtry* value of 2 and *ntree* of 1000 marked the least OBB error rate of 0.3689071. The highest OBB error rate of 0.38724 was presented with *mtry* of 6 in combination with *ntree* of 2500 (Figure 7a). The *mtry* of 2 and *ntree* of 1000 was used as the input parameter to train the Random Forest algorithm. When the dataset was balanced, *mtry* value of 3 and *ntree* of 5500 gave the least OBB error rate of 0.368246. An *mtry* of 4 and *ntree* of 2000 created the highest OBB error rate of 0.390524 (Figure 7b). A mix of *mtry* value of 3 and *ntree* of 5500 was used as the input parameter to train the Random Forest algorithm.



(a)



(b)

Figure 7: RF parameter optimization (*mtry* and *ntree*) for the imbalanced (a) and balanced (b) data using the grid search procedure.

4.2. Parameter tuning of SVM

4.2.1. Sentinel-2 MSI

For the imbalanced data, the cost *C* value of 100 with a gamma γ value of 1 was the best parameters producing the best performance at 0.1495082. These parameters were the input parameters to train the SVM algorithm. For the balanced data, the cost *C* value of 10 with a gamma γ value of 1 was the best parameters producing the best performance at 0.1412644. These values are the input parameters used to train the SVM algorithm.

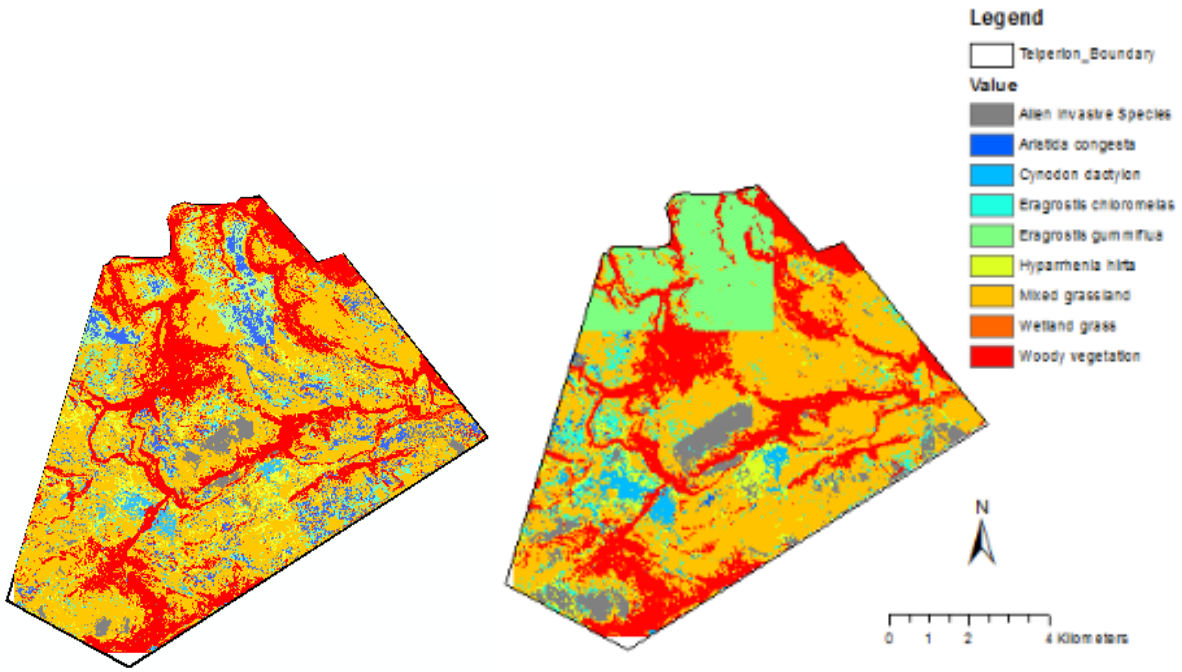
4.2.2. RapidEye imagery

For the imbalanced data, the cost *C* value of 100 with a gamma γ value of 0.1 was the best parameters producing optimal performance at 0.4119672. These parameters were the input parameters to train the SVM algorithm. For the balanced data, the cost *C* value of 100 with a gamma γ value of 0.1 was the best parameters resulting in the best performance at 0.4033266. These values are the input parameters used to train the SVM algorithm.

4.3. RF and SVM performance in mapping grass community

4.3.1. Sentinel-2 MSI imagery (imbalanced training data)

Nine classes were produced using the random forest and a support vector algorithm (Figure 8) on the imbalanced training data. The overall accuracy results show a slight distinction between the vegetation maps produced by the algorithms. As can be seen in the central and southwestern part of the two maps, there is a slight difference in the pixels of alien invasive species (Figure 8). There is also a difference in the northwestern part of the map where there is a significant amount of pixels of *Eragrostis gummiflua* in the SVM map, unlike the RF map. The dominant species on both maps is the *Mixed grasslands.*

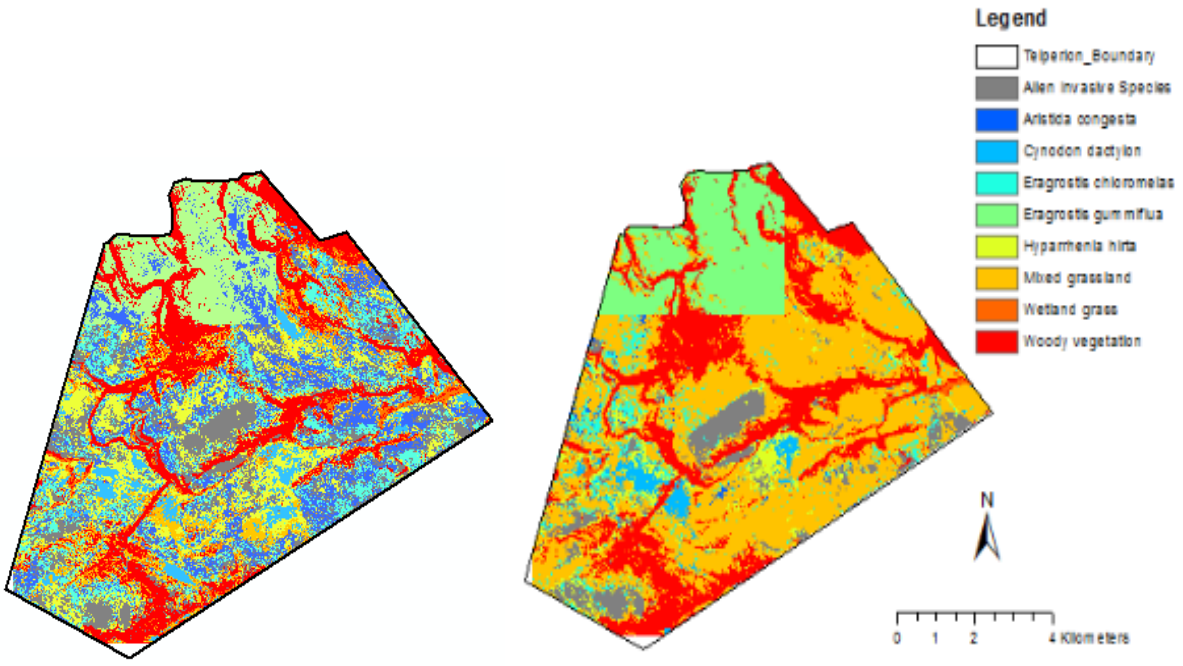


(a) (b)

Figure 8: Vegetation mapping classification using RF (a) and SVM (b) classification algorithm for imbalanced training data.

4.3.2. Sentinel-2 MSI imagery (balanced training data)

A total of nine classes was also obtained using the RF and SVM algorithm on the balanced training data (Figure 9). The overall accuracy results show significant differences between the vegetation maps generated by the two algorithms. The entirety of the two maps is different with the dominant species on the SVM map being *Mixed grasslands* while that of the RF maps is the *Cynodon Dactylon*. As can be seen in the central and southwestern part of the two maps, the pixel of alien invasive species remains relatively the same (Figure 9).

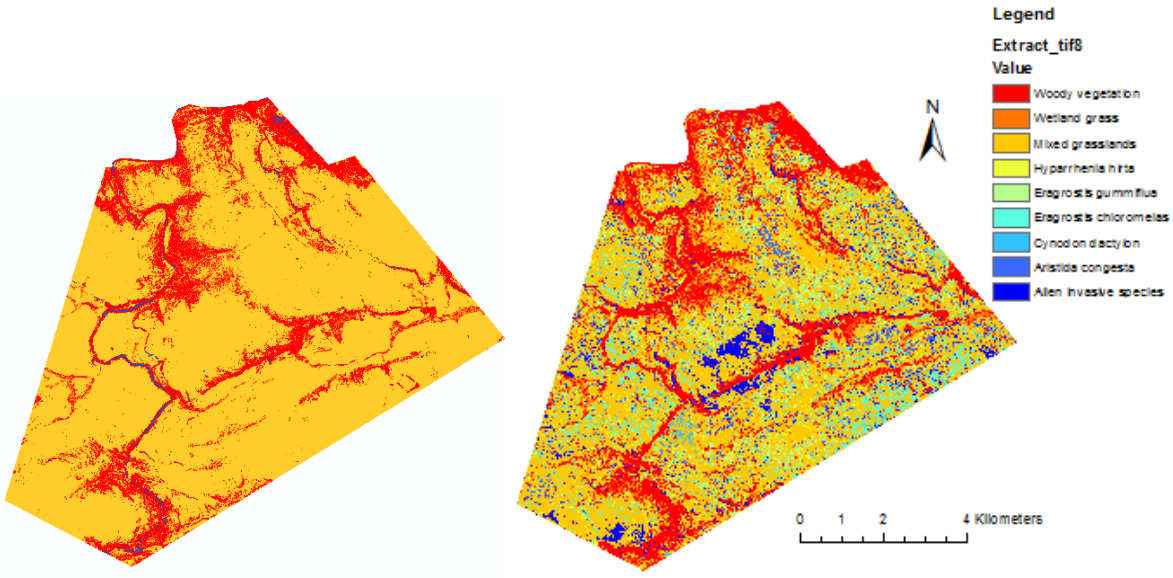


(a) (b)

Figure 9: Vegetation mapping classification using RF (a) and SVM (b) classification algorithm for balanced training data.

4.3.3. RapidEye imagery (imbalanced training data)

Nine classes was also obtained using the RF algorithm and a total of seven for the SVM algorithm on the balanced training data (Figure 10). *Aristida congesta* and *Hyparrhenia hirta* were missing on the SVM classification map. This absence is most likely due to the two classes being misclassified with others. The overall accuracy results show significant differences between the vegetation maps generated by the two algorithms. The entirety of the two maps is different, with the dominant species on the SVM map being *Mixed grasslands* followed by *Woody vegetation*. The RF map shows the dominant species *as Mixed grasslands, Eragrostis gummiflua* and *Alien Invasive species* in the extent of the map. As can be seen in the central and southwestern part of the two maps, the pixel of alien invasive species remains relatively the same (Figure 10).

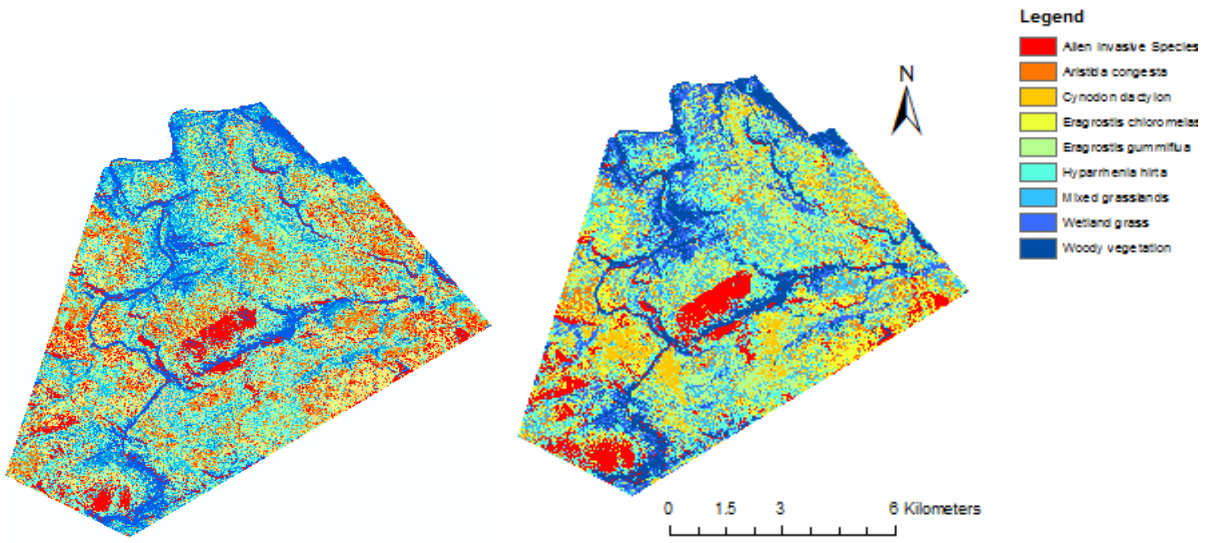


(a) (b)

Figure 10: Vegetation mapping classification using RF (b) and SVM (a) classification algorithm for imbalanced training data.

4.3.4. RapidEye imagery (balanced training data)

A total of nine classes was also obtained using the RF and SVM algorithm on the balanced training data (Figure 11). The overall accuracy results show slight differences between the vegetation maps generated by the two algorithms. In the northeastern and southeastern part of the map differences in the pixel can be seen where *Aristida congesta* is prominent in the RF map and *Eragrostis gummiflua* in the SVM map (Figure 11). It is hard to conclude on which species the most dominant in both maps is.



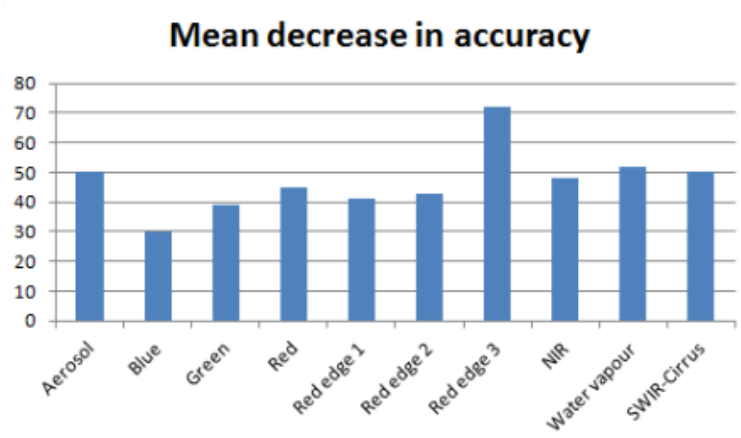
(a) (b)

Figure 11: Vegetation mapping classification using RF (a) and SVM (b) classification algorithm for balanced training data.

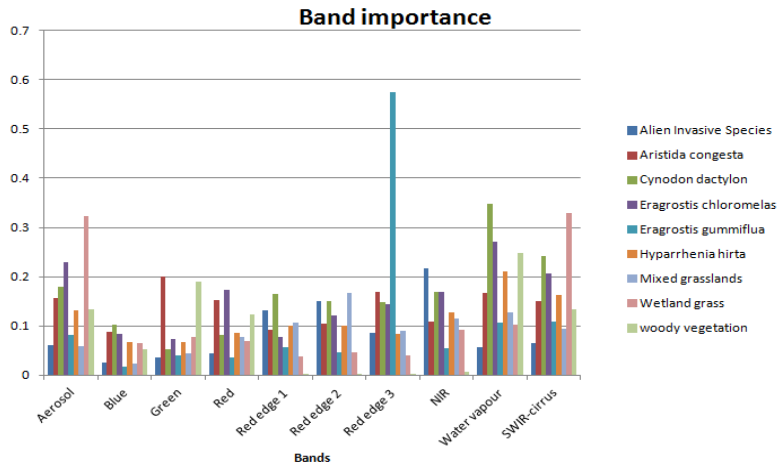
4.4. RapidEye and Sentinel-2 bands significance

4.4.1. Sentinel-2 MSI imagery (balanced training data)

During the classification process of the RF classification algorithm, we are provided with a measure of variable importance. The variable importance provided allowed us to identify the significance of each Sentinel-2 bands in mapping the vegetation (Figure 12). An assessment of the bands shows the Red-edge 3 band as the more dominant in the classification and modelling accuracy. The overall accuracy of the vegetation classification reduces by 70% when the Red-edge 3 band is omitted from the model (Figure 12a). The Red-edge 3 band is shown to be the best for depicting *Eragrostis gummiflua* while the Red-edge bands 1, 2 and 3 is the least important for describing *Woody vegetation* (Figure 12b).



(a)



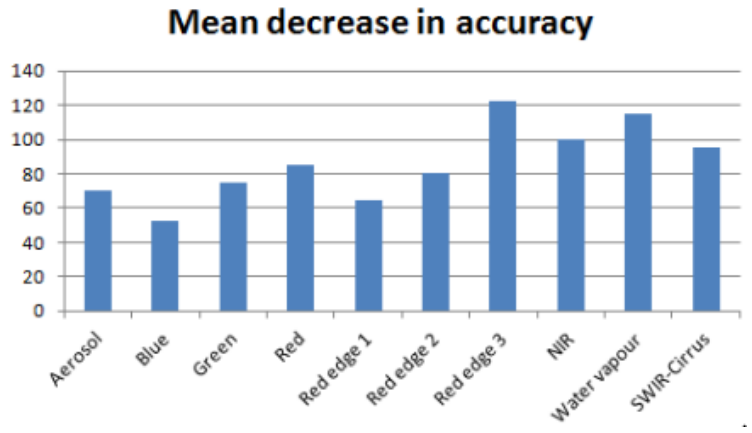
(b)

Figure 12: Sentinel band significance in vegetation classification for all the vegetation species (a) and each vegetation species (b). The most important band is the one with the highest mean decrease in accuracy.

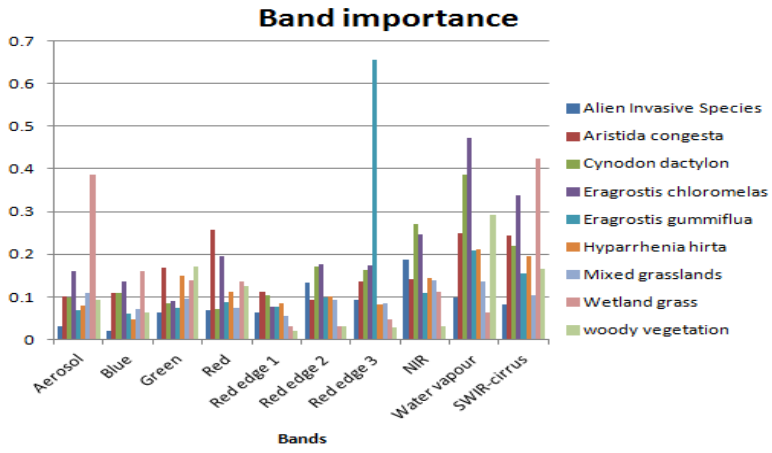
4.4.2. Sentinel-2 MSI imagery (imbalanced training data)

We are provided with a measure of variable importance during the RF classification process which allowed us to identify the significance of each Sentinel-2 band in mapping vegetation

(Figure 13). An assessment of the bands shows the Red-edge 3 band is the dominant band during classification and modelling accuracy (Figure 13a). The Red-edge 3 band is the most significant band for depicting *Eragrostis gummiflua* and the Red-edge is the least for describing *Woody vegetation* (Figure 13b).



(a)

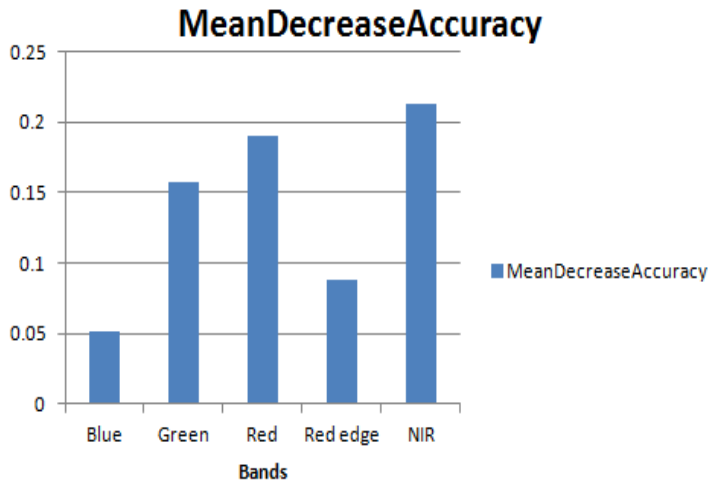


(b)

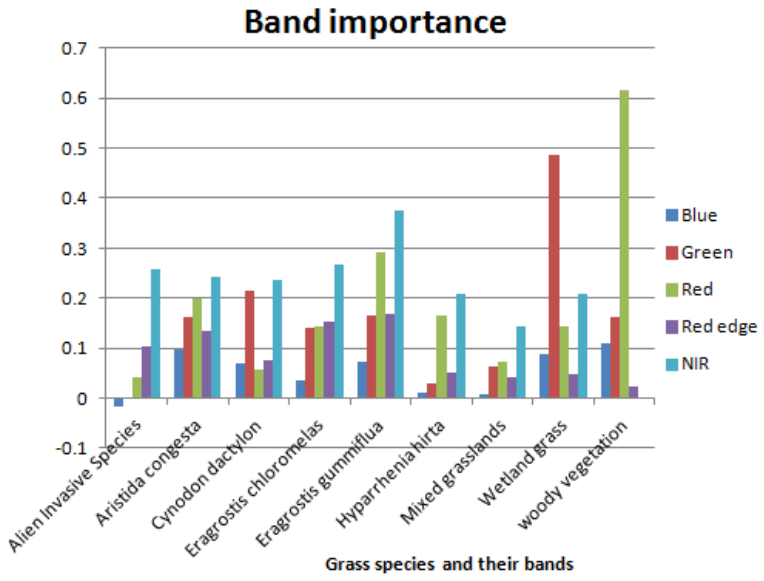
Figure 13: Sentinel band significance in vegetation classification for all the vegetation species (a) and each vegetation species (b). The highest mean decrease in accuracy specifies the most important band.

4.4.3. RapidEye Imagery (balanced training data)

The variable importance provided allowed us to identify the significance of each RapidEye bands in mapping the vegetation (Figure 14). In the classification and modelling accuracy, an assessment of the bands shows the NIR band to be the dominant band (Figure 14a). The Red band is the most valuable for depicting *Woody vegetation* and the blue band is the least relevant for describing *Alien invasive species* (Figure 14b).



(a)

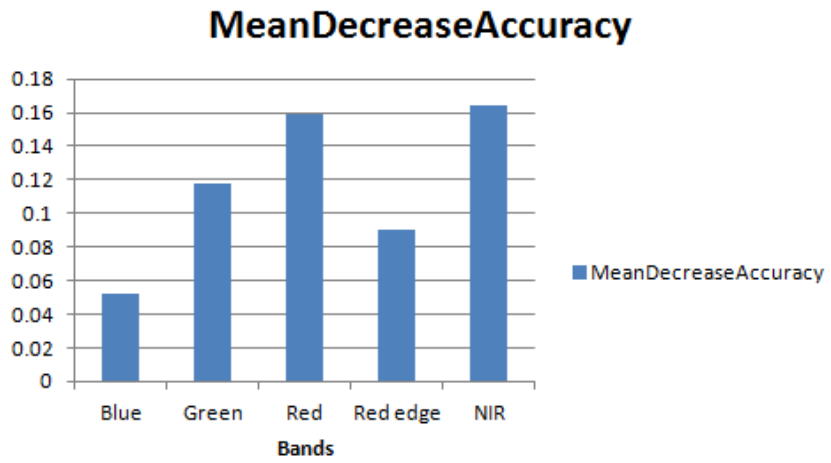


(b)

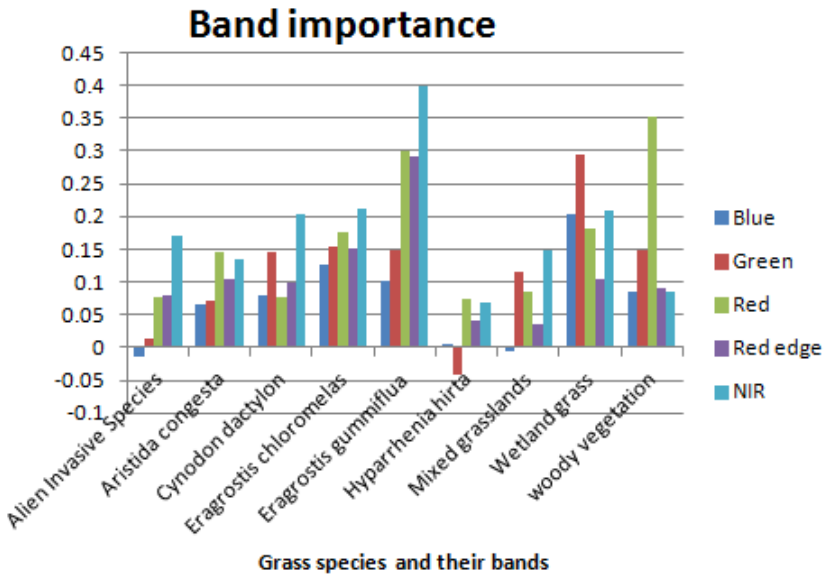
Figure 14: RapidEye band significance in vegetation classification for all the vegetation species (a) and each vegetation species (b). The highest mean decrease in accuracy specifies the most important band.

4.4.4. RapidEye imagery (imbalanced training data)

The variable importance provided allowed us to identify the significance of each RapidEye bands in mapping the vegetation (Figure 15). An assessment of the bands shows the NIR band is the effective band in classification and modelling accuracy (Figure 15a). Meanwhile the NIR band proves the most valuable for depicting *Eragrostis gummiflua* while the green band is least effective for describing *Hyparrhenia hirta invasive species* (Figure 15b).



(a)



(b)

Figure 15: RapidEye band significance in vegetation classification for all the vegetation species (a) and each vegetation species (b). The highest mean decrease in accuracy specifies the most important band.

4.5. Accuracy assessment

4.5.1. Sentinel-2 MSI (balanced dataset)

Accuracy assessment was carried out using the test data of the balanced training data set to enable the performance estimate of the trained models for both random forest and support vector classification algorithms. The overall accuracy of 76.19% and a kappa coefficient of 73.21% were achieved for RF classifier. In general, all vegetation species produced above 60% user's accuracy except for *Mixed grasslands* with an accuracy of 42.86%. RF achieved above 60% producer's accuracy for most vegetation species (Table 5). SVM produced higher accuracy compared to the RF classifier with an overall accuracy of 82.54% and a kappa coefficient of 80.36%. The SVM classifier achieved over 65% of the user's accuracy and over 60% of the producer's accuracy except for *Mixed grassland* with an accuracy of 50% (Table 6). Using this classifier can create some confusion between *Mixed grassland, Aristida congesta* and *Alien invasive species, indicating* that some spectral similarities exist between these grass species.

4.5.2. Sentinel-2 MSI (imbalanced dataset)

Accuracy assessment was carried out using the test data set of the imbalanced data set to enable the performance estimate of the trained models for both random forest and support vector classifiers. The overall accuracy of 79.45% and a kappa coefficient of 74.38% were achieved for RF (Table 7). The user's accuracy was above 65% with possible misclassification between *Alien invasive species, Aristida congesta,* and *Mixed grasslands*. The producer's accuracy was above 70% for the grass species. The overall accuracy of 82.21% and a kappa coefficient of 78.33% for the SVM classifier was produced (Table 8).

Table 5: Confusion matrix using random forest (RF) for vegetation species and the associated accuracies including kappa statistic (KC), overall accuracy (OA), producer's accuracy (PA) and user's accuracy (UA) of the Sentinel-2 image using the test data of the balanced dataset.

Class	**AIS**	**AC**	**CD**	**EC**	**EG**	**HH**	**MG**	**WG**	WV	**TOTAL**	**UA**	**PA**
AIS	9	1	0	0	1	0	3	0	1	15	60.00%	64.29%
AC	1	10	0	0	1	0	1	0	0	13	76.92%	71.43%
CD	0	0	9	0	1	0	2	0	0	12	75.00%	64.29%
EC	2	1	0	12	0	0	0	0	0	15	80.00%	85.71%
EG	0	2	0	0	11	0	0	0	0	13	84.62%	78.57%
HH	1	0	0	0	0	13	1	0	0	15	86.67%	92.86%
MG	1	0	5	2	0	0	6	0	0	14	42.86%	54.54%
WG	0	0	0	0	0	1	1	14	1	17	82.35%	100.00%
WV	0	0	0	0	0	0	0	0	12	12	100%	85.71%
TOTAL	14	14	14	14	14	14	11	14	14	96		
Overall accuracy: 76.19% **Kappa coefficient: 73.21%**												

Table 6: Confusion matrix using support vector machine (SVM) for vegetation species and the associated accuracies including KC, OA, PA and UA of the Sentinel-2 image using the test data of the balanced dataset.

Class	**AIS**	**AC**	**CD**	**EC**	**EG**	**HH**	**MG**	**WG**	**WV**	**TOTAL**	**UA**	**PA**
AIS	10	1	0	0	1	0	3	0	0	15	66.67%	71.43%
AC	3	12	0	0	1	0	1	0	0	17	70.59%	92.31%
CD	0	0	13	0	0	0	0	0	0	13	100.00%	92.86%
EC	1	1	0	12	0	0	0	0	0	14	85.71%	85.71%
EG	0	0	0	0	11	0	2	0	0	13	84.62%	78.57%
HH	0	0	0	0	1	11	0	0	0	12	91.67%	78.57%
MG	0	0	0	2	0	2	7	0	0	11	63.64%	50.00%
WG	0	0	0	0	0	0	1	14	0	15	93.33%	100.0%
WV	0	0	1	0	0	1	0	0	14	16	87.50%	100.0%
TOTAL	14	13	14	14	14	14	14	14	14	104		
Overall accuracy: 82.54% **Kappa coefficient: 80.36%**												

Table 7: Confusion matrix using random forest (RF) for vegetation species and the associated accuracies including OA, UA, PA and KC of the Sentinel-2 image using the test data of the imbalanced dataset.

Class	**AIS**	**AC**	**CD**	**EC**	**EG**	**HH**	**MG**	**WG**	**WV**	**TOTAL**	**UA**	**PA**
AIS	8	0	0	1	0	0	3	0	0	12	66.67%	40.00%
AC	0	9	0	0	0	0	2	0	0	11	81.81%	52.94%
CD	0	0	13	0	0	0	1	0	0	14	92.86%	92.86%
EC	1	0	0	16	0	0	1	0	0	18	88.89%	72.73%
EG	0	2	0	0	16	0	1	0	0	19	84.21%	84.21%
HH	0	0	0	0	0	12	1	0	0	13	92.31%	70.59%
MG	11	5	1	5	2	3	74	1	3	105	70.48%	86.05%
WG	0	0	0	0	0	1	0	17	1	19	89.47%	94.44%
WV	0	1	0	0	1	1	3	0	36	42	85.71%	90.00%
TOTAL	20	17	14	22	19	17	86	18	40	201		
Overall accuracy: 79.45% **Kappa coefficient: 74.38%**												

Table 8: Confusion matrix using support vector machine (SVM) for vegetation species and the associated accuracies including the UA, OA, KC and PA of the Sentinel 2 image using the test data of the imbalanced dataset.

Class	**AIS**	**AC**	**CD**	**EC**	**EG**	**HH**	**MG**	**WG**	**WV**	**TOTAL**	**UA**	**PA**
AIS	14	1	0	2	1	0	2	0	0	20	70.00%	70.00%
AC	0	11	0	0	0	0	2	0	0	13	84.62%	64.71%
CD	0	0	13	0	0	1	0	0	0	14	92.86%	92.86%
EC	0	0	0	17	1	0	1	0	1	22	77.27%	77.27%
EG	0	2	0	0	16	0	2	0	1	21	76.19%	84.21%
HH	0	0	0	0	0	13	2	0	0	15	86.67%	76.47%
MG	6	3	1	3	0	1	71	1	1	87	81.61	82.56%
WG	0	0	0	0	0	0	0	17	1	18	94.44%	94.44%
WV	0	0	0	0	1	2	4	0	36	43	83.72	90.00%
TOTAL	20	17	14	22	19	17	86	18	40	208		
Overall accuracy: 82.21% **Kappa coefficient: 78.33%**												

4.5.3 RapidEye imagery (balanced dataset)

Accuracy assessment was carried out using the test data of the balanced training data set to enable the trained models prediction performance for both random forest and support vector classifiers. The overall accuracy of 57.94% and a kappa coefficient of 52.68% were achieved for RF classifier. In general, all vegetation species produced a low user's accuracy except for *Woody vegetation, Wetland grass and Cynodon dactylon* with accuracies of over 90%. The random forest classifier achieved low producer's accuracy for a good number of the vegetation species except for *Woody vegetation, Eragrostis chloromelas, Wetland grass and Cynodon dactylon* (Table 9). These low values mean that there could be misclassification amongst the other vegetation species. SVM classifier produced a lower accuracy when compared to RF classifier with an overall accuracy of 50.79% and a kappa coefficient of 44.64%. The SVM classifier achieved low accuracy for the user's and producer's accuracy except for *Woody vegetation* and *Wetland grass* (Table 10). Using this classifier can create some confusion among the other grass species.

4.5.4. RapidEye imagery (imbalanced dataset)

Accuracy assessment was carried out using the test data of the balanced training data set to enable the trained models performance prediction for both random forest and support vector to be carried out. The overall accuracy of 63.24% and a kappa coefficient of 53.58% were achieved for RF classifier. In general, all vegetation species achieved a low user's accuracy except for *Woody vegetation, Wetland grass, Aristida congesta and Cynodon dactylon* with accuracies of over 90%. The RF classifier also achieved low producer's accuracy for a good number of the vegetation species except for *Woody vegetation, Wetland grass and Mixed grassland* (Table 11). These low values mean that there could be misclassification amongst the other vegetation species. SVM classifier produced a lower accuracy in comparison to the RF classifier with an overall accuracy of 57.31% and a kappa coefficient of 46.11%. The SVM classifier achieved low accuracy for the user's accuracy except for *Woody vegetation* and *Wetland grass* and a low producer's accuracy except for *Eragrostis gummiflua, Mixed grasslands, Woody vegetation* and

Wetland grass (Table 12). Using this classifier can create some confusion among the other grass species.

Table 9: Confusion matrix using random forest for vegetation species and the associated accuracies for RapidEye image using the test data of the balanced train data.

Class	**AIS**	**AC**	**CD**	**EC**	**EG**	**HH**	**MG**	**WG**	**WV**	**TOTAL**	**UA**	**PA**
AIS	6	3	0	0	1	0	3	0	0	13	46.15%	42.86%
AC	1	3	0	0	4	2	0	0	0	10	30.00%	20.00%
CD	2	0	10	1	0	0	0	0	0	13	76.92%	92.86%
EC	5	3	2	11	0	0	1	0	0	22	50.00%	71.43%
EG	0	0	1	0	9	1	3	0	0	14	64.29%	64.29%
HH	0	3	0	1	0	7	4	1	1	17	41.18%	50.00%
MG	0	2	1	1	0	2	2	0	0	8	25.00%	15.38%
WG	0	0	0	0	0	2	0	13	0	15	86.67%	92.86%
WV	0	0	0	0	0	0	0	0	12	12	100.0%	92.31%
TOTAL	14	15	14	14	14	14	13	14	13	73		
Overall accuracy: 57.94% **Kappa coefficient: 52.68%**												

Table 10: Confusion matrix using support vector machines for vegetation species and the associated accuracies for RapidEye imagery using the test data of the balanced train data.

Class	**AIS**	**AC**	**CD**	**EC**	**EG**	**HH**	**MG**	**WG**	**WV**	**TOTAL**	**UA**	**PA**
AIS	5	2	0	2	1	0	4	0	1	15	33.33%	35.71%
AC	1	2	1	1	5	4	0	0	0	14	14.29%	14.29%
CD	1	2	8	0	0	1	2	0	1	14	57.14%	57.14%
EC	7	5	3	10	0	1	1	0	0	27	37.04%	71.43%
EG	0	0	0	1	8	2	3	0	0	14	57.14%	57.14%
HH	0	2	0	0	0	3	0	1	0	6	50.00%	21.43%
MG	0	1	2	0	0	1	3	0	0	7	42.86%	21.43%
WG	0	0	0	0	0	1	1	13	0	15	86.67%	92.86%
WV	0	0	0	0	0	1	0	0	12	13	92.31%	85.71%
TOTAL	14	14	14	14	14	14	14	14	14	64		
Overall accuracy: 50.79% **Kappa coefficient: 44.64%**												

Table 11: Confusion matrix using random forest for vegetation species and the associated accuracies of the RapidEye image using the test data of the imbalanced train data.

Class	**AIS**	**AC**	**CD**	**EC**	**EG**	**HH**	**MG**	**WG**	**WV**	**TOTAL**	**UA**	**PA**
AIS	3	3	0	0	0	0	2	0	2	10	30.00%	15.00%
AC	1	6	0	0	0	0	1	0	0	8	75.00%	35.29%
CD	1	0	7	0	0	0	0	0	1	9	77.78%	50.00%
EC	4	2	3	11	1	0	5	0	1	27	40.74%	52.38%
EG	0	0	0	1	13	1	3	0	1	19	68.42%	68.42%
HH	1	0	0	0	0	2	3	0	0	6	33.33%	11.76%
MG	10	6	4	10	4	11	68	1	0	122	55.74%	79.07%
WG	0	0	0	0	0	1	0	16	1	18	88.89%	88.89%
WV	0	0	0	0	1	2	4	1	34	42	80.95%	85.00%
TOTAL	20	17	14	21	19	17	86	18	40	160		
Overall accuracy: 63.24% **Kappa coefficient: 53.58%**												

Table 12: Confusion matrix using support vector machines for vegetation species and the various accuracies of the RapidEye image using the test data of the imbalanced train data.

Class	**AIS**	**AC**	**CD**	**EC**	**EG**	**HH**	**MG**	**WG**	**WV**	**TOTAL**	**UA**	**PA**
AIS	4	2	0	0	0	0	1	0	2	9	44.44%	20.00%
AC	0	2	0	0	1	0	0	0	0	3	66.67%	11.76%
CD	0	1	3	1	0	0	1	0	1	7	42.86%	21.43%
EC	4	1	5	7	1	2	4	0	0	24	29.17%	31.81%
EG	0	2	0	2	14	2	7	0	1	28	50.00%	73.68%
HH	0	1	0	0	0	4	8	0	0	13	30.77%	23.53%
MG	12	8	6	12	1	7	64	1	4	115	55.65%	74.42%
WG	0	0	0	0	1	0	0	15	0	16	93.75%	83.33%
WV	0	0	0	0	1	2	1	2	32	38	84.21%	80.00%
TOTAL	20	17	14	22	19	17	86	18	40	145		
Overall accuracy: 57.31% **Kappa coefficient: 46.11%**												

Table 13: Overall accuracy of random forest and support vector machines for RapidEye and Sentinel 2 MSI images of both balanced and imbalanced data set.

	Random Forest	Support Vector Machine
RapidEye (balanced data)	57.94%	50.79%
RapidEye (imbalanced data)	63.24%	57.31%
Sentinel 2 (balanced data)	76.19%	82.54%
Sentinel 2 (imbalanced data)	79.45%	82.21%

CHAPTER FIVE

DISCUSSION AND CONCLUSION

5.1. Discussion

The mapping of vegetation plays a significant role in conservation and biodiversity. For proper conservation, up to date information about the spatial distribution of grass community is essential. Over the past few years, the use of various spatial and spectral resolutions of optical sensors has been used in vegetation mapping with varying success. The availability of these remotely sensed data, their different costs, and concerns over their accuracy is still an issue (Lu and Weng, 2007). While the high-resolution imagery is preferred for mapping vegetation, their limited availability due to cost and high dimensionality still pose a problem (Mutanga, Adam and Cho, 2012).

The classification results of this study show the ability of Sentinel-2 in discerning the spectral attributes of different grass species. The red edge bands improved the accuracy of the classification algorithm as has been believed in mapping grass species (Ramoelo et al. 2012; Clevers et al. 2001). The SVM algorithm produced high classification accuracy for the Sentinel-2 image. It is also seen that SVM works well with either a balanced or imbalanced dataset with consistent accuracy (Table 6 and Table 8) with the Sentinel-2 image which is consistent with the study by Shao and Lunetta, 2012. RF classification algorithm produced a drop in accuracy by 3% when using a balanced dataset (Table 5) compared to the imbalanced dataset (Table 7). This drop in accuracy is consistent with Noi and Kappas (2017). The two algorithms performed well using the Sentinel-2 imagery.

The classification results on the RapidEye imagery show a decrease in the overall accuracy of both algorithms when the data set was balanced. RF classification accuracy dropped by 6% (Table 9 and Table 11) while SVM decreased by 7% (Table 10 and Table 12). RF classifier produced higher accuracy than SVM classifier. According to this research, the algorithms worked better on the Sentinel-2 MSI imagery than the RapidEye imagery, although both algorithms for each image, produced accuracies in close range. It is suggested that RF classifier on different satellite imagery with different training sample size generates different accuracy (Noi and Kappas, 2017). Some researchers show that RF accuracy is higher for an imbalanced dataset which is consistent with this study (Mellor et al., 2015; Colditz 2015). This study showed

the difference in performance of SVM and RF which is consistent with Nitze et al. (2012) and Pouteau et al. (2012). However, other studies have shown similarities in the performances of SVM and RF (Pal 2005; Waske et al. 2009).

SVM and RF were unable to deal with the spatial variation problems common in vegetation mapping as some misclassifications were present. Misclassification is a common problem when dealing with high-resolution imagery (Lu and Weng, 2007). The overall classification accuracies might have improved if post classification was done (Duro et al. 2012). This study highlighted the importance of each sentinel-2 (Figure 14 and Figure 15) and RapidEye band (Figure 16 and Figure 17). It showed the improved ability of the new generation multispectral imagery in distinguishing different vegetation species (Cho et al. 2012; Mutanga, Adam and Cho, 2012).

The results and interpretation are however only a prelude to further research into vegetation mapping using high-resolution imagery and the effects of different sizes of training data for different classification algorithm. New analysis can be carried out in this research in a way to deal with the misclassification issues.

5.2. Conclusion

This research assessed and contrasted the classification of RapidEye (five bands) and Sentinel 2 (ten bands) imagery using advanced SVM and RF on balanced and imbalanced training data set. The results provided information about the performances of these two new generation multispectral images using SVM and RF of balanced and imbalanced training dataset in mapping grass communities in the Telperion Nature Game Reserve. This study showed that the performances of RF and SVM classifier is dependent on the type of satellite images used and its accuracy is affected when dealing with a balanced and imbalanced dataset which is in agreement with Noi and Kappa (2017). This claim is in contradiction to other studies that SVM is unaffected when dealing with both balanced and imbalanced datasets (Ustuner et al., 2016; Noi and Kappas, 2017). The importance of each band of both imageries was shown to affect overall accuracy. Misclassification, which was also a problem in other related studies, was attributed to

high spatial variation within an associated vegetation class. Therefore, moving forward, different approaches to solving this problem should be addressed.

References

Adam, E., Mureriwa, N. and Newete, S., 2017. Mapping Prosopis glandulosa (mesquite) in the semi-arid environment of South Africa using high-resolution WorldView-2 imagery and machine learning classifiers. Journal of Arid Environments 145, 43-51.

Adam, E., Mutanga, O., Odindi, J. and Abdel-Rahman, E. M., 2014. Land-use/ cover classification in a heterogeneous coastal landscape using RapidEye imagery: Evaluating the performance of random forest and support vector machines classifiers. *Int. J. Remote Sens.*35, 3440-3458.

Adam, E., Mutanga, O. and Rugege, D., 2010. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: A review. *Wetlands Ecology and Management*, 18 (3), pp. 281–296.

Adam, E., Mutanga, O., Rugege, D. and Ismail, R., 2012. Discriminating the papyrus vegetation (Cyperus papyrus L.) and its co-existent species using random forest and hyperspectral data resampled to HYMAP. *International Journal of Remote Sensing*, 33:2, 552-569.

Adelabu, S., Mutang, O. and Adam, E., 2015. Testing the reliability and stability of the internal accuracy assessment of random forest using different validation methods. *Geocarto International* 30, 810–821.

Adjorlolo, C. et al., 2012. Challenges and opportunities in the use of remote sensing for C-3 and C-4 grass species discrimination and mapping. *African Journal of Range & Forage Science,* 29(2), 47–61.

Akasheh, O. Z., Neale, C. M. U. and Jayanthi, H., 2008. Detailed mapping of riparian vegetation in the middle Rio Grande River using high resolution multispectral airborne remote sensing. Jour of arid environs vol 72(9). 1734-1744

Al-doski, J., Mansor, S., and Shafri, H., 2013. Image classification in remote sensing. *J. Environ. Earth Sci.* 3, 141–147.

Al-Mashreki, M. H., Akhir, J. B. M., Rahim, S. A., Desa, K., and Rahman, Z. A., 2010. Remote Sensing and GIS Application for Assessment of Land Suitability Potential for Agriculture in the IBB Governorate, the Republic of Yemen. *Pakistan Journal of Biological Sciences,* 13: 1116-1128.

Ali, I., Cawkwell, F., Dwyer, E., Barrett, B. and Green, S., 2016. Satellite remote sensing of grasslands: from observation to management, *Journal of Plant Ecology*, Volume 9, Issue 6, 649–671.

Anand, A., Pugalenthi G., Fogel G. and Suganthan P.N., 2010. An approach for classification of highly imbalanced data using weighting and undersampling**,** *Volume 39, Issue 5*, 1385–1391.

Asner, G. P, and Warner, A. S., 2003. Canopy shadow in IKONOS satellite observations of tropical forests and savannas. *Remote Sensing of Environment*; 87: 521–53.

Baldi, G., Guerschman, J. P. and Paruelo, J. M., 2006. Characterizing fragmentation in temperate South America grasslands. Agriculture, Ecosystems and Environment 116(3-4):197-208.

Belgiu, M., and Drăguţ, L., 2016. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing*114: 24–31.

Bergman, K. O., Ask, L., Askling, J. et al., 2008. Importance of boreal grasslands in Sweden for butterfly diversity and effects of local and landscape habitat factors. *Conserv* 17: 139.

Batuwita, R. and Palade, V., 2012. Adjusted geometric-mean: A novel performance measure for imbalanced bioinformatics datasets learning. *Journal of Bioinformatics and Computational Biology*. 10 (4-23).

Baumstark, R., Duffey, R. and Pu, R., 2016. Mapping seagrass and colonized hard bottom in 56 Springs Coast, Florida using WorldView-2 satellite imagery. *Estuarine, Coastal and Shelf Science*, 181, pp.83–92.

Blagus, R. and Lusa, L., 2010. Class prediction for high-dimensional class-imbalanced data, *BMC Bioinformatic*,vol.11;523.

Blaschke, T., 2010. Object-based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing;* 65: 2–16.

Borges, J. S., Marçal, A. R. and Dias, J. M., 2007. Evaluation of feature extraction and reduction methods for hyperspectral images, in *Proceedings of the 26th EARSeL Symposium-New Developments and Challenges in Remote Sensing*, 255–264.

Boyle, S. A., Kennedy, C. M., Torres, J. et al. 2014. High-resolution satellite imagery is an important yet underutilized resource in conservation biology. *PLoS One*;9:e86908.

Bredenkamp, G., Chytery, M., Fischer, H., Neuhauslova, Z. and Van der Maarel, E., 1998. Vegetation mapping: Theory, methods and case studies: Introduction. *Applied Vegetation Science* 1, 161–164.

Breiman, L., 2001. Machine Learning.45: 5-32. https://doi.org/10.1023/A:1010933404324

Burges, C.J. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl.* Discov. 1998, 2, 121–167.

Camps-Valls, G., Chalk, A. M., Serrano-López, A. J. et al. 2004. Profiled support vector machines for antisense oligonucleotide efficacy prediction. *BMC Bioinformatics*, 5: 135–143.

Carleer, A. P., and Wolff, E., 2006. Urban land cover multi-level region-based classification of VHR data by selecting relevant features. *International Journal of Remote Sensing*;27 (5-6):1035–1051.

Chabalala Y. W., 2017. Estimation of nitrogen content across grass communities at Telperion Nature Reserve using Sentinel-2.

Chastain, R.A. Jr., Struckhoff, M. A., He, H.S. and Larsen, D. R., 2008. Mapping vegetation communities using statistical data fusion in the Ozark national scenic river ways, Missouri, USA. *Photogramm Eng Remote Sens*. 74:24–264.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357.

Chen, C., Liaw, A. and Breiman, L., 2004. Using Random Forest to Learn Imbalanced Data. *Statistics Department*. University of California, Berkeley.

Chen, J., Franklin, J. F. and Spies, T. A., 1992. Vegetation responses to edge environments in old-growth Douglas-fir forests. *Ecological Applications*, 2:387-396.

Cho, M. A., Debba, P., Mutanga, O. et al., 2012. Potential utility of the spectral red-edge region of SumbandilaSat imagery for assessing indigenous forest structure and health. *International Journal of Applied Earth Observation and Geoinformation*, 16(1), 85–93.

Cingolani, A. M., Renison, D., Zak, M. R and Cabido. M. R., 2004. Mapping Vegetation in a Heterogeneous Mountain Rangeland Using Landsat Data: An Alternative Method to Define and Classify Land-Cover Units. *Remote Sensing of Environment*, 92 (1): 84–97.

Clark, M. L., and Kilham, N. E., 2016. Mapping of Land Cover in Northern California with Simulated HyspIRI imagery. *ISPRS J. Photogramm. Remote Sens*, *119*, 228–245.

Colditz, R.R., 2015. An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. *Remote Sens.* 7, 9655–9681.

Congalton, R. G., and Green, K., 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Boca Raton: *CRC Press.*

Cushnie, J. L., 1987. The interactive effect of spatial resolution and degree of internal variability within land cover types and classification accuracies. *lnternational Journal of Remote Sensing*, Vol 8, No. 1,15-29.

D'Addabbo, A.and Maglietta R., 2015. Parallel selective sampling method for imbalanced and large data classification. *Pattern Recogn Lett* 62:61–67.

Dambach, P., Lacaux, J. P., Vignolles, C. et al., 2009. Using high spatial resolution remote sensing for risk mapping of malaria occurrence in the Nouna district, Burkina Faso. doi: 10.3402/gha.v2i0.2094.

Darvishzadeh, R., Skidmore, A., Atzberger, C. and Wieren S., 2008. Estimation of vegetation LAI from hyperspectral reflectance data: effects of soil type and plant architecture. *Int J Appl Earth Obs Geoinform* 10:358–373.

DeFries, R. S., Foley, J. A. and Asner G. P., 2004. Land use choices: Balancing human needs and ecosystem function. *Frontiers in Ecology and the Environment* 2: 249–257.

Drusch, M., Del Bello, U., Carlier, S. et al., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote *Sensing of Environment*, 120: 25-36.

Duro, D. C., Franklin, S. E. and Dubé, M. G., 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment* 118: 259–272.

Egbert, S. L., Park, S., Price, K. P. et al., 2002. Using conservation reserve program maps derived from satellite imagery to characterize landscape structure. *Comput Electron Agric*, vol. 37. 141-56.

Estabrooks, A., Jo, T., Japkowicz, N., 2004. A multiple resampling method for learning from imbalanced data sets. *Comput. Intell.* 20, 18–36.

Evans, R. D., Murray, K. L., Field, S. N. et al., 2012. Digitise This! A quick and easy remote sensing method to monitor the daily extent of dredge plumes.

Fauvel, M., Trabalka, Y., Benediktsson, J. A. et al., 2013. Advances in spectral-spatial classification of hyperspectral images. Proceedings of the IEEE, *Institute of Electrical and Electronics Engineers*, 101 (3), pp.652-675.

Feng, W., Guo, B., Zhang, H. et al., 2015. Remote estimation of above ground nitrogen uptake during vegetative growth in winter wheat using hyperspectral red-edge ratio data. *Field Crops Research*, 180: 197–206.

Filippi, A. M., and Jensen, J. R.,2006. Fuzzy learning vector quantization for hyperspectral coastal vegetation classification. *Remote Sensing of Environment*, vol. 100, no. 4: 512–530.

Fisher, J. R. B., Acosta, E. A., Dennedy-Frank, P. J., Kroeger, T. and Boucher, T. M., 2017. Impact of satellite imagery spatial resolution on land use classification accuracy and modeled water quality. *Remote Sens Ecol Conserv*. doi:10.1002/rse2.61.

Foley, J. A., DeFries, R., Asner, G. P. et al., 2005. Global consequence of land use. 309, 570–574.

Foody, G. M., 2002. Status of Land Cover Classification Accuracy Assessment. *Remote Sensing of Environment* 80 (1): 185–201.

Foody, G. M., 1995a. Cross-entropy for the evaluation of the accuracy of a fuzzy land cover classification with fuzzy ground data. *ISPRS Journal of Photogrammetry and Remote Sensing*, (in press).

Foody, G., 1999. The significance of border training patterns in classification by a feed forward neural network using back propagation learning. *Int. J. Remote Sens.* 20: 37–41.

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., and Boyd, D. S., 2006. Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, 104: 1-14.

Friedl, M. A., and Brodley, C. E., 1997. Decision tree classification of land cover from remotely sensed data. *Remote. Sensing of Environment*, 61: 399–409.

Fry, J. A., Xian, G., Suming, J. et al., 2011. Completion of the 2006 national land cover database for the coterminous United States. *Photogrammetry Engineering Remote Sensing* 77: 858–864.

Gebhardt, S., Wehrmann, T., Ruiz, M. A. M., et al., 2014. MAD-MEX: Automatic wall-to-wall land-cover monitoring for the Mexican REDD-MRV program using all landsat data. *Remote sensing* 6: 3923-3943.

Giri, C., Defourny, P. and Shrestha, S., 2003. Land Cover Characterization and Mapping of Continental Southeast Asia Using Multi-Resolution Satellite Sensor Data. *International Journal of Remote Sensing 24* (21): 4181–4196.

Gong, P. and Howarth, P. J., 1990. An assessment of some factors influencing multispectral land-cover classification, *Photogrammetric Engineering and Remote Sensing*, 56(5):597-603.

Govender, M., Chetty, K., Naiken, V., and Bulcock, H., 2008. A comparison of satellite hyperspectral and multispectral remote sensing imagery for improved classification and mapping of vegetation. *Water SA*, *34*(2), 147-154.

Ghosh, A. and Joshi, P. K., 2014. A comparison of selected classification algorithms for mapping bamboo patches in lower Gangetic plains using very high-resolution WorldView 2 imagery. *Int. J. Appl. Earth Obs. Geoinf*,26, 298–311.

Glenn, N. F., Mundt, J. T., Weber, K. T., Prather, T. S., Lass, L. W. and Pettingill, J., 2005. Hyperspectral data processing for repeat detection of small infestations of leafy spurge. *Remote Sensing of Environment, 9: 399-412.*

Guidici, D. and Clark, M. L., 2017. Land-Cover classification of multi-seasonal hyperspectral imagery in the San Francisco Bay Area, California. *Remote Sensing*, *9*(6): 629.

Guo, L., Chehata, N., Mallet, C. and Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests, *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(1): 56-66.

Hansen, T., 2012. A review of large area monitoring of land cover change using Landsat data. *Remote Sensing of Environment*, 122: 66-74.

Harvey, K, and Hill, G., 2001. Vegetation mapping of a tropical freshwater swamp in the Northern Territory, Australia: a comparison of aerial photography, Landsat TM, and SPOT satellite imagery. *International Journal of Remote Sensing,* 22: 2911–2925.

Hastie, T., Tibshirani, R., and Friedman, J., 2001. The elements of statistical learning: Data mining, inference, and prediction. 536. New York: *Springer–Verlag*.

He, H., 2011. Index, in Self-Adaptive Systems for Machine Intelligence, *John Wiley & Sons, Inc., Hoboken, NJ, USA.* doi: 10.1002/9781118025604.index.

He, C., Li, Z., LI, X. and Shi, P., 2005. Zoning grassland protection area using remote sensing and cellular automata modeling—a case study in Xilingol steppe grassland in northern China, *J Arid Environ*, vol. 63(pg. 814-26).

Hegde, G., Ahamed, J., Hebbar, R. and Raj, U., 2014. Urban land cover classification using hyperspectral data. *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. XL-8.

Hermes, L., Frieauff, D., Puzicha, J. and Buhmann, J. M., 1999. Support vector machines for land usage classification in Landsat TM imagery, Geoscience and Remote Sensing Symposium, 1999. IGARSS'99 Proceedings. *IEEE 1999 International*, 348.

Herold, M., Liu, X. H. and Clarke, K.C., 2003. Spatial metrics and image texture for mapping urban land use. *Photogrammetric Engineering and Remote Sensing* 69(9): 991–1001.

Herold, M., Gardner, M. E., Roberts, D. A., 2003. Spectral resolution requirements for mapping urban areas. *IEEE Transactions on Geoscience and Remote Sensing* 41 (9 PART I), 1907-1919.

Hill, T., and Lewicki, P., 2007. Statistics: Methods and applications: a comprehensive reference for science, industry and data mining. *StatSoft*; Tulsa, Okla.

http://web.mit.edu Accessed 23rd February 2018.

http://www.harrisgeospatial.com/docs/BackgroundSVM.html. Accessed 25th January 2018.

Huang, C., Davis, L. S. and Townshend, J. R. G., 2002. An Assessment of Support Vector Machines for Land Cover Classification. *International Journal of Remote Sensing* 23 (4): 725–749.

Huang, S. et al., 2017. Potential of RapidEye and WorldView-2 Satellite Data for Improving Rice Nitrogen Status Monitoring at Different Growth Stages. *Remote Sensing* 9(3): 227.

Huang, S., and Siegert, F., 2006. Land cover classification optimized to detect areas at risk of desertification in North China based on SPOT VEGETATION imagery. *Journal of Arid Environments*. 67; 308- 327.

Hubert-Moy, L., Cotonnec, A., Le Du, L., Chardin, A. and Perez, P., 2001. A comparison of parametric classification procedures of remotely sensed data applied on different landscape units. *Remote Sensing of Environment*, 75: 174–187.

Hubert-Moy, L., Corgne, S., Mercier, G. and Solaiman, B., 2002. Land use and land cover change prediction with the theory of evidence : a study case in an intensive agricultural region in France. *Proceedings Fusion Conference*, Washington D.C., pp. 114-122.

Humboldt state university website 2015 (second edition 1993).

Irons, J. R., Markham, B. L., Nelson, R. F., et al., 2007. The effects of spatial resolution on the classification of Thematic Mapper data. *International Journal of Remote sensing*, 6(8): 1385-1403.

Jacquin, A., Misakova, L. and Gay, M., 2008. A hybrid object-based classification approach for mapping urban sprawl in periurban environment. *Landscape and Urban Planning* 84: 152–165.

Japkowicz, N., 2000. The Class Imbalance Problem: Significance and Strategies. In: Proceedings of the 2000 International Conference on Artificial Intelligence: Special Track on Inductive Learning, Las Vegas, Nevada.

Japkowicz, N. and Stephen, S., 2002. The class imbalance problem: a systematic study. *Intell. Data Anal*. 6: 429–450.

Jensen, J. R., 2007. Introductory to Digital Image Processing: A Remote Sensing Perspective. Prentice Hall Series in Geographic Information Science.

Jin, H., Stehman, S. V. and Mountrakis, G., 2014. Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado, *International Journal of Remote Sensing*, 35:6, 2067-2081.

Jin, J., 2012. A Random Forest Based Method for Urban Land Cover Classification using LiDAR Data and Aerial Imagery. *UWSpace.*

Jung, M., Herold, M., Henkel, K., and Churkina, G., 2006. Exploiting synergies of land cover products for carbon cycle modelling, Remote Sens. Environ., 101, 534–553.

Kandala, U.M., Gopi, K 2016. Fire monitoring and assessment in forest using remote sensing and GIS. *SSRG International Journal of Industrial Engineering (SSRG-IJIE)* – Vol 3 Issue 5: 26.

Kanellopoulos, I., Varfis, A., Wilkinson, G. and Mégier, J., 1992. Land-cover discrimination in SPOT HRV imagery using an Artificial Neural Network—A 20-Class Experiment. *International Journal of Remote Sensing* 13 (5): 917–924.

Kasischke, E. S., Goetz, S., Hansen, M. C. et al., 2004. Temperate and boreal forests. In S. Ustin (Ed.), *Manual of Remote Sensing* Volume 4: Remote Sensing for Natural Resource Management and Environmental Monitoring: John Wiley and Sons 848.

Khatami, R., Mountrakis, G. and Stehman, S. V., 2016. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment* 177: 89–100.

Kavzoglu, T., 2009. Increasing the accuracy of neural network classification using refined training data. Environmental Modelling & Software, 24(7): 850-858.

Kavzoglu, T., and Colkesen, I., 2009a. 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. and Mather, P. M., 2002. The role of feature selection in artificial neural network applications. *International journal of Remote sensing* 23: 2787-2803.

Kim, H. and Yeom, J., 2015. Comparison of NDVIs from GOCI and MODIS data towards improved assessment of crop temporal dynamics in the case of paddy rice. *Remote Sens*. *7*(9): 11326-11343.

Koch, M., Inzana, J. and El Baz, F., 2005. Applications of Hyperion hyperspectral and aster multispectral data in characterizing vegetation for water resource studies in arid lands. Paper No. 44-5. *Proc. Geol. Remote Sens. Annual Meeting*. Salt Lake City, USA

Kubat, M. and Matwin, S., 1997. Addressing the curse of imbalanced training sets: One-sided selection in Machine Learning-International Workshop then Conference, (Nashville, TN, USA), 179–186.

Kulkarni A. D. and Barrett Lowe, 2016. Random Forest Algorithm for land cover classification. *International Journal on Recent and Innovation Trends in Computing and Communication* 4 (3): 58–63.

Laliberte, A. S., Rango, A., Havstad, K. M. et al., 2004. Object-oriented image analysis for mapping shrub encroachment from 1937 to 2003 in southern New Mexico. *Remote Sensing of Environment* 93: 198–210.

Lambin, E. F., Turner, B. L., Geist, H. J. et al., 2001. The causes of land-use and land-cover change: Moving beyond the myths' *Global Environmental Change*, vol 11,(4): 261-269.

Latham, J., Cumani, R., Rosati, I. and Bloise, M., 2014. FAO Global Land Cover (GLC-SHARE) Beta-Release 1.0 Database.

Laurikkala, J., 2001. Improving identification of difficult small classes by balancing class distribution. 63–66. *Proc. Conf. AI in Medicine in Europe: Artificial Intelligence Medicine,* 63-66.

Lawrence, R. L., Wood, S. D. and Sheley, R. L., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing Environment 100*: 356–362.

Li, M., Zang, S., Zhang, B. et al., 2014. A review of remote sensing image classification techniques: The role of spatio-contextual information. *European Journal of Remote Sensing* 47: 389-411.

Ling, B., Goodin, D. G., Mohler, R. L. et al., 2014. Estimating canopy nitrogen content in a heterogeneous grassland with varying fire and grazing treatments: Konza Prairie, Kansas, USA. *Remote Sensing* 6(5),

Lowther, S. A, Curriero, F.C., Shields, .T. et al., 2009. Feasibility of satellite image-based sampling for a health survey among urban townships of Lusaka, Zambia. *Trop Med Int Health;* 14(1): 70-78.

Lu, D. and Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing,* 28 (5): 23-870.

Lu, D. and Weng, Q., 2009. Extraction of urban impervious surface from an IKONOS image. *International Journal of Remote Sensing*;30(5): 1297–1311.

Lussem, U., Hütt, C. and Waldhoff, G., 2016. Combined analysis of Sentinel-1 and rapideye data for improved crop type classification: an early season approach for rapeseed and cereals**.** *Int. Arch. Photogramm. Remote Sens*. Spatial Inf. Sci. XLI-B8; 959-963.

Ma, L., Lia, M., Ma, X., Chenga, L., Du, P. and Liu, Y., 2017. A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 130: 277–293.

Ma, Z., Liu, Q., Sun, K. and Zhan, S., 2016. A syncretic representation for image classification and face recognition. *CAAI Trans Intell Technol* 1(2):173–178.

MacFayden, S., Hui, C., Verberg, P.H. and Van Teeffelen, A. J. A., 2016. Quantifying spatiotemporal drivers of environmental heterogeneity in Kruger National Park, South Africa, *Landscape Ecology* 31: 2013-2029.

Mansour, K and Mutanga, O., 2012. Classifying increaser species as an indicator of different levels of rangeland degradation using WorldView-2 imagery. *J Appl Remote Sens*. 6:1–18.

Mansour, K., Mutanga, O., Adam, E. and Abdel-Rahman, E., 2016. Multispectral remote sensing for mapping grassland degradation using the key indicators of grass species and edaphic factors, *Geocarto International*, 31(5): 477-491.

Massetti, A., Sequeira, M. M., Pupo, A. et al., 2016. Assessing the effectiveness of RapidEye multispectral imagery for vegetation mapping in Madeira Island (Portugal*), European Journal of Remote Sensing*, 49(1): 643-672.

Mather, P. M. 2004. Computer Processing of Remotely-Sensed Images: An introduction, 3rd Edition. John Wiley & Sons. ISBN:978-0-470-02101-9.

Mathieu, R., Aryal, J., and Chong, A. K., 2007. Object-Based Classification of Ikonos Imagery for Mapping Large-Scale Vegetation Communities in Urban Areas. *Sensors (Basel, Switzerland) 7*(11): 2860–2880.

Mathieu, R., Seddon, P., and Leiendecker, J., 2006. Predicting the distribution of raptors using remote sensing techniques and Geographic Information Systems: a case study with the Eastern New Zealand falcon (Falco novaeseelandiae), *N.Z. J Zool*, 33:73-84.

Martínez-López, J., Carreño, M. F., Palazón-Ferrando, J. A. et al., 2014. Remote sensing of plant communities as a tool for assessing the condition of semiarid Mediterranean saline wetlands in agricultural catchments. *Int J Appl Earth Obs Geoinformation.* 26: 193–204.

Maxwell A. E., Warner T.A. and Fang, F., 2018. Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing* Vol. 39, Iss. 9.

Mayaux, P., Achard, F. and Malingreau, J. P., 1998. Global tropical forest area measurements from coarse resolution satellite imagery: a comparison with other methodologies. Environmental Conservation, 25, 37–52.

Melgani, F. and Bruzzone, L. 2004. Classification of hyperspectral remote sensing images with support vector machines. *Geoscience and Remote Sensing, IEEE* Transactions on, vol. 42 (8): 1778-1790.

Mellor A., Boukir S., Haywood A. and Jones S., 2015. Exploring issues of training data imbalance and mislabeling on random forest performance for large area land cover classification using the ensemble margin. *ISPRS J. Photogramm. Remote Sens 105*: 155–168.

Melville, B., Lucieer, A.and Aryal, J., 2008. Object-based random forest classification of Landsat ETM+ and WorldView-2 satellite imagery for mapping lowland native grassland communities in Tasmania, Australia. *Int. J. Appl. Earth Obs. Geoinform*, 66: 46–55.

Melville, B., Lucieer, A. and Aryal, J., 2018. Assessing the impact of spectral resolution on classification of lowland native grassland communities based on field spectroscopy in Tasmania, Australia. *Remote Sensing 10*(2): 308.

Millard, K. and Richardson, M., 2015. On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote Sens 7*: 8489–8515.

Möckel, T., 2015. Hyperspectral and multispectral remote sensing for mapping grassland vegetation. *Department of Physical Geography and Ecosystem Science*, Lund University.

Moran, E. F., 2010. Land Cover Classification in a Complex Urban-Rural Landscape with Quickbird Imagery. *Photogrammetric Engineering and Remote Sensing, 76(10), 1159–1168.*

Mountrakis, G., Im, J. and Ogole, C., 2011. Support Vector Machines in Remote Sensing: A Review. *ISPRS J. Photogramm. Remote Sens 66:* 247–259.

Muchoney, D. and Williamson, J., 2001. A Gaussian adaptive resonance theory neural network classification algorithm applied to supervised land-cover mapping using multitemporal vegetation index data. *IEEE Transactions on Geoscience and Remote Sensing*, 39(9): 1969−1977.

Mundt, J. T., Glenn, N. F., Wever, K. T. et al., 2005. Discrimination of hoary cress and determination of its detection limits via hyperspectral image processing and accuracy assessment techniques. *Remote Sensing of Environment* 96: 509 – 517.

Mutanga, O. and Skidmore, A. K., 2004. Hyperspectral band depth analysis for a better estimation of grass biomass (Cenchrus ciliaris) measured under controlled laboratory conditions. *International Journal of Applied Earth Observation and Geoinformation*, 5(2): 87–96.

Mutanga, O., Adam, E. and Cho, M. A., 2012. High Density Biomass Estimation for Wetland Vegetation Using WorldView-2 Imagery and Random Forest Regression Algorithm. *International Journal of Applied Earth Observation and Geoinformation* 18: 399–406.

Mutanga, O., Skidmore, A. K. and Van Wieren, S., 2003. Discriminating tropical grass (Cenchrus ciliaris) canopies grown under different nitrogen treatments using

spectroradiometry. *ISPRS Journal of Photogrammetry and Remote Sensing* 57(4): 263–272.

Nichol, J. and Lee, C. M., 2005. Urban vegetation monitoring in Hong Kong using high resolution multi-spectral images. *International Journal of Remote Sensing* 26(5): 903–918.

Nitze, I., Schulthess, U. and Asche, H., 2012. Comparison of Machine Learning Algorithm Random Forest, Artificial Neural Network and Support Vector Machine to Maximum Likelihood for Supervised Crop Type Classification. *In Fourth International Conference on Geographic Object-Based Image Analysis (GEOBIA)*, 035, Rio de Janeiro.

Nkonyana T.N., 2016. Image classification using machine learning techniques.

Noi, P. T., and Kappas, M., 2017. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors*,18.

Odindi, J., Elhadi, A., Zinhle, N. et al., 2014. Comparison between {WorldView}-2 and {SPOT}-5 images in mapping the bracken fern using the random forest algorithm. *Journal of Applied Remote Sensing*, 8(1): 83527.

Odindi, J., Mutanga, O., Rouget, M. and Hlanguza, N., 2016. Mapping alien and indigenous vegetation in the KwaZulu-Natal Sandstone Sourveld using remotely sensed data, *Bothalia*46(2): a2103.

Omer, G., Mutnga, O., Abdel-Rahman, E., and Adam, E., 2015. Performance of Support Vector Machines and Artifical Neural Network for Mapping Endangered Tree Species Using WorldView-2 Data in Dukuduku Forest, South Africa. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(10): 4825–4840.

Padma S., Kumar S.S. and Manavalan R., 2011. Performance analysis for classification in balanced and unbalanced data set. *Industrial and Information Systems (ICIIS), 6th IEEE International Conference*.300-304.

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

Pal, M. and Mather, P. 2005. Support vector machines for classification in remote sensing. *International Journal of Remote Sensing* 26(5): 1007-1011.

Pal, M., Maxwell, A. E., and Warner, T. A., 2013. Kernel-based extreme learning machine for remote-sensing image classification. *Remote Sensing Letters* 4: 853-862.

Pidwirny, M., 2006. Introduction to Geographic Information Systems. *Fundamentals of Physical Geography*, 2nd Edition.

Peerbhay, K. Y., Mutanga, O. and Ismail, R.,2013. Commercial tree species discrimination using airborne AISA Eagle hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) in KwaZulu–Natal, South Africa. *ISPRS Journal of Photogrammetry and Remote Sensing* 79: 19–28.

Pouteau, R., Meyer, J., Taputuarai, R.and Stoll, B., 2012. Support vector machines to map rare and endangered native plants in Pacific Islands Forests. *Ecological Informatics* 9: 37–46.

Qian, Y. G., Zhou, W. Q., Yan, J. L. et al., 2015. Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sensing* 7 (1): 153–168.

Quattrochi, D. A., and Goodchild, M. F. (Eds). 1997. *Scale in Remote Sensing and GIS*, New York*: Lewis Publishers.*

RapidEye. 2010. "RapidEye Satellite Constellation, Germany." Accessed May 2016. http://www.rapideye.net/upload/RE_Constellation.pdf.

Ramoelo, A., Skidmore, A. K., Cho, M. A. et al., 2013. Non-linear partial least square regression increases the estimation accuracy of grass nitrogen and phosphorus using in situ hyperspectral and environmental data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 82: 27–40.

Ramoelo, A., Cho, M., Mathieu, R. et al., 2015. Potential of Sentinel-2 spectral configuration to assess rangeland quality. *Journal of Applied Remote Sensing*, 9(1): 94096.

Ramoelo, A., Skidmore, A., Cho, M. et al., 2012. Regional estimation of savanna grass Nitrogen using the red-edge band of the spaceborne RapidEye sensor. *International Journal of Applied Earth Observation and Geoinformation* 19: 151–162.

Richards, J. A., 2006. Remote sensing digital image analysis: An introduction 5th Edition *Springer-Verlag. New York, Inc*. Secaucus, NJ, USA.

Rignot, E., Salas, W. A., and Skole, D. L., 1997. Mapping deforestation and secondary growth in Rondonia, Brazil, using imaging radar and thematic mapper data. *Remote Sensing Environ*ment 59: 167–179.

Rodriguez-Galiano, V. F., Chica-Olmo, M., Abarca-Hernandez, F. et al., 2012. Random forest classification of mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment* 121: 93–107.

Running, S. W., Heinsch, F. A., Zhao, M. et al., 2004. A continuous satellite derived measure of global terrestrial production. *BioScience,* 54: 547-560.

Saatchi, S. S., Houghton, R. A., Dos Santos Alvala, R. C. et al., 2007. Distribution of aboveground live biomass in the Amazon basin. *Global Change Biol* 13: 816–837.

Salehi,B., Zhang, Y., Zhong, M. and Dey, V., 2012. Object based classification of urban areas using VHR imagery and height points ancillary data. *Remote sens*. 4(8): 2256-2276.

Schohn, G. and Cohn, D., 2000. Less is more: Active learning with support vector machines. *In Proc. 17th Int. Conf. Machine Learning* 839-846.

Sch¨olkopf, B. and Smola, A. J., 2002. Learning with Kernels. *MIT Press*.

Schumacher, P., Mislimshoeva, B., Brenning, A. et al., 2016. Do red edge and texture attributes from high-resolution satellite data improve wood volume estimation in a semi-arid mountainous region? *Remote Sens* 8: 540.

Schuster, C., Förster, M. and Kleinschmit, B., 2012. Testing the red edge channel for improving land-use classifications based on high-resolution multi-spectral satellite data. *International Journal of Remote Sensing* 33 (17): 5583–5599.

Schwieder, M., Leitão, P. J., Bustamante, M. M. et al., 2016. Mapping Brazilian savanna vegetation gradients with Landsat time series. *International Journal of Applied Earth Observation and Geoinformation* 52: 361–370.

Sesnie, S. E., Finegan, B., Gessler, P. E. et al., 2010. The multispectral separability of Costa Rican rainforest types with support vector machines and random forest decision trees. *International Journal of Remote Sensing* 31(11): 2885–2909.

Shao, Y. and Lunetta, R.S., 2012. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. ISPRS *J. Photogramm. Remote Sensing*, 70: 78–87.

Shi, D. and Yang, X., 2015. Support vector machines for land cover mapping from remote sensor imagery. In Monitoring and Modeling of Global Changes: A Geomatics Perspective. *Springer:* Dordrecht, The Netherlands 265–279.

Shields, T., Pinchoff, J., Lubinda, J. et al., 2016. Spatial and temporal changes in household structure locations using high-resolution satellite imagery for population assessment: an analysis of household locations in southern Zambia between 2006 and 2011. *Geospat Health* 11(2): 410.

Sibanda, M., Mutanga, O. and Rouget, M., 2017. Testing the capabilities of the new WorldView-3 space-borne sensor's red-edge spectral band in discriminating and mapping complex grassland management treatments. *International Journal of Remote Sensing*, 38(1), pp.1–22.

Simand, C., Chillà, F. and Pinton, J. F., 2000. Inhomogeneous turbulence in the vicinity of a large-scale coherent vortex: erratum. *Europhys. Lett.* 49: 821.

Simard, M., Saatchi, S. S., and De Grandi, G., 2000. The use of a decision tree and multiscale texture for classification of JERS-1 SAR data over tropical forest. *IEEE Transactions on Geoscience and Remote Sensing*, 38(5), 2310−2321.

Small, C., 2003. High spatial resolution spectral mixture analysis of urban reflectance. *Remote Sensing of Environment* 88(1-2): 170–186.

Statnikov, A., Hardin, D., Guyon, I., Aliferis, C. F., 2011. A gentle introduction to support vector machines in biomedicine. *World Scientific Publishing Co., Inc.*

Stefanov, W. L., and Netzband, M., 2005. Assessment of ASTER land cover and MODIS NDVI Data at multiple scales for ecological characterization of an arid urban center. *Remote Sensing of Environment* 99 (1–2): 31–43.

Stow, D., Lopez, A., Lippitt, C. et al., *2007.* Object-based classification ofresidential land use within Accra, Ghana based on QuickBird satellite data. *Int J Remote Sens 28(22): 5167-5173.*

Stuart, N., Barratt, T. and Place C., 2006. Classifying the neotropical savannas of Belize using remote sensing and ground survey. *J Biogeogr* 33: 476–90.

Terri, J. A and Stowe, L. G., 1976. Climatic patterns and the distribution of C4 grasses in North America. *Oecologia* 23: 1-16.

Toivonen, T. and Luoto, M., 2003. Landsat TM images in mapping of semi-natural grasslands and analysing of habitat pattern in an agricultural landscape in south-west Finland. *Fennia - International Journal of Geography*, *181*(1): 49-67.

Torbick, N., Ledoux, L., Salas, W. and Zhao, M., 2016. Regional mapping of plantation extent using multisensor imagery. *Remote Sensing* 8(3): 236.

Trebar M. and Steele N., 2008. Applications of distributed SVM architectures in classifying forest data cover types**.** *Computers and Electronics in Agriculture* 63(2): 119-130.

Tsai, F., Lin, E. K., and Yoshino, K., 2007. Spectrally segmented principal component analysis of hyperspectral imagery for mapping invasive plant species. *International Journal of Remote Sensing*, 28(5): 1023-1039.

Tucker, C. J., Townshend, J. R. G. and Goff, T. E., 1985. African land-cover classification using satellite data. Science 227, 369–375.

Tucker, C. J., Townshend, J. R. G., Goff, T. E. and Holben B. N., 1986. Continental and Global Scale Remote Sensing of Land Cover. In: Trabalka J.R., Reichle D.E. (eds) The Changing Carbon Cycle. *Springer, New York, NY.*

U.S. Geological Survey. Available online: https://earthexplorer.usgs.gov/ (accessed on 20th May 2016).

Ustuner, M., Sanli F. B and Abdikan, S., 2016. Balanced Vs. Imbalanced training data: Classifying RapidEye data with Support Vector Machines. *Volume XLI-B7, 2016 XXIII ISPRS Congress.*

van Swaay, C. A. M., 2002. The importance of calcareous grasslands for butterflies in Europe. *Biological Conservation* 104: 315–318.

Vanselow, K. A., and Samimi, C., 2014. Predictive mapping of dwarf shrub vegetation in an arid high mountain ecosystem using remote sensing and random forests. *Remote Sensing* 6(7): 6709–6726.

Verburg, P. H., Neuman, K, and Nol, L., 2011. Challenges in using Land use and land cover data for global change studies. *Glob. Chang. Biol*, 17: 974-989.

Vogelmann, J., T. Sohl, P. Campbell, and D. Shaw., 1998. Regional land cover characterization using Landsat Thematic Mapper Data and ancillary data sources. *Environmental Monitoring and Assessment 51* (1/2): 415–428.

Walter, V., 2004. Object-based classification of remote sensing data for change detection. ISPRS *Journal of Photogrammetry and Remote Sensing*. 58: 225–238.

Wang, X., Mannaerts, C. M., Yang, S. et al., 2010. Evaluation of soil nitrogen emissions from riparian zones coupling simple process-oriented models with remote sensing data. *Science of the Total Environment* 408(16): 3310–3318.

Wang, Y., Traber, M., Milstead, B. and Stevens,S., 2007. Terrestrial and submerged aquatic vegetation mapping in fire island national seashore using high spatial resolution remote sensing data. *Marine Geodesy* 30(1–2): 77–95.

Wang, Z. W., Wang, Q., Zhao, L., et al., 2016. Mapping the vegetation distribution of the permafrost zone on the Qinghai-Tibet Plateau. *Journal of Mountain Science* 13(6): 1035–1046.

Waske, B., Benediktsson, J. A., Árnason, K. and Sveinsson, J. R., 2009. Mapping of hyperspectral AVIRIS data using machine-learning algorithms. *Canadian Journal of Remote Sensing* 35 (S1): S106–S116.

Wei, Q. and Dunbrack Jr, R. L., 2013. The role of balanced training and testing data sets for binary classifiers in bioinformatics . *PLoS ONE, 8*(7): e67863.

Weiss, G. M., and Provost, F., 2003. Learning when training data are costly: The effect of class distribution on tree induction. *J. Artif. Intell*. Res. 19: 315–354.

Weiss, G. M., and Provost, F. 2001. The effect of class distribution on classifier learning: An empirical study. *Department of Computer Science*. Rutgers Univ.

Weng, Q. and Quattrochi, D. A., 2007. Urban remote sensing. *Landscape and Urban Planning* 99(s 3–4): 259–260

Wessels, K. J., Prince,S. D., Frost, P. E. and VanZyl, D., 2004. Assessing the effects of human induced land degradation in the former homelands of northern South Africa with a 1-km AVHRR NDVI time-series. *Remote Sensing and Environment* 91: 47-67

Williams, A. P. and Hunt Jr., E. R., 2004. Accuracy assessment of detection of leafy spurge with hyperspectral imagery. *Journal of Range Management* 57: 106 – 112.

Wu, G., and Chang, E., 2003. Adaptive feature-space conformal transformation for imbalanced data learning. *To appear in Proc. of the 20th International Conference on Machine Learning*.

Wu, B., Gommes, R., Zhang, M., Zeng, H., Yan, N., Zou, W., Zheng, Y., Zhang, N., Chang, S., Xing, Q., Van Heijden, A., 2015. Global crop monitoring: A satellite-based hierarchical approach. *Remote Sensing 7*: 3907–3933.

Wulder, M. A., White, J. C., Goward, S. N. et al., 2008. Landsat Continuity: Issues and Opportunities for Land Cover Monitoring. *Remote Sensing of Environment* 112 (3): 955–969.

Xiao, X., Zhang, Q., Braswell, B. et al., 2004. Modeling gross primary production of temperate deciduous broadleaf forest using satellite images and climate data. *Remote sensing of environment* 91: 256-70.

Xie, Y., Sha, Z. and Yu, M., 2008. Remote sensing imagery in vegetation mapping: a review. *Journal of Plant Ecology* 1(1): 9–23.

Yang, C., Everitt, J. H., Du, Q. et al., 2013. Using High-Resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture. , *IEEE* 101(3).

Yang, J., Gong, P., Fu, R. and Dickinson, R., 2013. The role of satellite remote sensing in climatechange studies. *Nature Climate Change volume* 3: 875–883.

Zhang, C., Cai, D., Guo, S. et al., 2016. Spatial-Temporal Dynamics of China's Terrestrial Biodiversity: A Dynamic Habitat Index Diagnostic. *Remote Sens.* *8*(3): 227.

Zhang, C. and Xie, Z., 2012. Combining object-based texture measures with a neural network for vegetation mapping in the Everglades from hyperspectral imagery. *Remote Sensing of Environment* 124: 310–320.

Zhang, Q., Wang, J., Gong, P. and Shi, P., 2003. Study of urban spatial patterns from SPOT panchromatic imagery using textural analysis. *International Journal of Remote Sensing* 24: 4137–4160.

Zheng, C. H., Zeng C. S., Chen, Z. Q., et al., 2006. A study on the changes of landscape pattern of estuary wetlands of the Minjiang River. *Wetland Sci* 4: 29-35.

Zhou, W., Troy, A. and Grove, J. M., 2008. Object-based land cover classification and change analysis in the Baltimore metropolitan area using multi-temporal high resolution remote sensing data. *Sensors* 8: 1613–1636.

Zomer, R. J., Trabucco, A., and Ustin, S. L., 2009. Building spectral libraries for wetlands land cover classification and hyperspectral remote sensing. *Journal of Environmental Management* 90: 2170–2177.

Zomer, R. J., Trabucco, A., Coe, R. et al., 2014. Trees on Farms: An update and reanalysis of agroforestry's global extent and Socio-Ecological characteristics. ICRAF Working Papers, No. 179. World Agroforestry Centre (ICRAF). Bogor, Indonesia.