Remote sensing of forest health: The detection and mapping of *Pinus patula* trees infested by *Sirex noctilio*

Riyad Ismail

February 2009

Remote sensing of forest health: The detection and mapping of *Pinus patula* trees infested by *Sirex noctilio*

by

Riyad Ismail

Submitted in fulfilment of the academic requirements for the degree of Doctor of Philosophy in the School of Environmental Sciences, Geography, University of KwaZulu-Natal, Pietermaritzburg

December 2008

As the candidate's supervisor I have/have not approved this thesis/dissertation for submission.

Signed: _____ Name: _____ Date: ____

Preface

The work described in this thesis was carried out in the School of Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, from June 2005 to December 2008, under the supervision of Professor Onisimo Mutanga and Professor Urmilla Bob.

These studies represent original work by the author and have not otherwise been submitted in any form for any degree or diploma to any tertiary institution. Where use has been made of the work of others it is duly acknowledged in the text.

Declaration 1: Plagiarism

I Riyad Ismail, declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs, or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   a. Their words have been re-written, but the general information attributed to them has been referenced.
   b. Where their exact words have been used, their writing has been placed in italics and inside quotation marks, and referenced.

5. This thesis does not contain text, graphics, or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed: ……

Declaration 2: Publications

1. Ismail, R., Mutanga, O. and Bob, U., 2006. The use of high resolution airborne imagery for the detection of forest canopy damage by *Sirex noctilio*. In: P.A. Langin and M.C. Antonides (Editors), Precision forestry in plantations, semi-natural areas and natural forest: Proceedings of the international precision forestry symposium. Stellenbosch University, Stellenbosch University, South Africa, pp. 119-134.

2. Ismail, R., Mutanga, O. and Bob, U., 2007. Forest health and vitality: The detection and monitoring of *Pinus patula* trees infected by *Sirex noctilio* using digital multispectral imagery (DMSI). Southern Hemisphere Forestry Journal, 69(1): 39-47.

3. Ismail, R., Mutanga, O., Kumar, L. and Bob, U., 2008. Determining the optimal resolution of remotely sensed data for the detection of *Sirex noctilio* infestations in *Pinus patula* plantations in KwaZulu-Natal, South Africa. The South African Geographical Journal, 90(1): 196-204.

4. Ismail, R., Mutanga, O. and Ahmed, F., 2008. Discriminating *Sirex noctilio* attack in pine forest plantations in South Africa using high spectral resolution data. In: M. Kalacska and A. Sanchez-Azofeifa (Editors), Hyperspectral Remote Sensing of Tropical and Sub-Tropical Forests. Taylor and Francis: CRC Press, pp. 350.

5. Ismail, R. and Mutanga, O., (accepted). Discriminating the early stages *of Sirex noctilio* infestation using random forest and shortwave infrared (SWIR) wavelengths. International Journal of Remote Sensing.

6. Ismail, R. and Mutanga, O., *in review*. A comparison of regression tree based ensemble methods: Predicating *Sirex noctilio* induced water stress. International Journal of Geoinformation and Earth Observation. Special issue: Remote sensing for Africa.

7. Ismail, R., Mutanga, O., Kumar, L., *in review*. Modelling the potential distribution of pine forests that are susceptible to *Sirex noctilio* infestations in Mpumalanga, South Africa. Transactions in GIS.

Signed: ……

Table of contents

Abstract

Sirex noctilio is causing considerable mortality in commercial pine forests in KwaZulu-Natal, South Africa. The ability to remotely detect *S. noctilio* infestations remains crucial for monitoring the spread of the wasp and for the effective deployment of suppression activities. This thesis advocates the development of techniques based on remote sensing technology to accurately detect and map *S. noctilio* infestations. To date, no research has examined the potential of remote sensing technologies for the detection and mapping of *Pinus patula* trees infested by *S. noctilio.*

In the first part of this thesis, the focus was on whether high spatial resolution imagery could characterize *S. noctilio* induced stress in *P. patula* forests. Results showed that, the normalized difference vegetation index derived from high spatial resolution imagery has the potential to accurately detect and map the later stages of *S. noctilio* infestations. Additionally, operational guidelines for the optimal spatial resolutions that are suitable for detecting and mapping varying levels of sustained *S. noctilio* mortality were defined. Results showed that a pixel size of 2.3 m is recommended to detect high (11-15%) infestation levels, and a pixel size of 1.75 m is recommended for detecting low to medium infestation levels (1-10%).

In the second part of this thesis, the focus was on the ability of high spectral resolution (hyperspectral) data to discriminate between healthy trees and the early stages of *S. noctilio* infestation. Results showed that specific wavelengths located in the visible and near infrared region have the greatest potential for discriminating between healthy trees and the early stages of *S. noctilio* infestation. The researcher also evaluated the robustness and accuracy of various machine learning algorithms in identifying spectral parameters that allowed for the successful detection of *S. noctilio* infestations. Results showed that the random forest algorithm simplified the process by identifying the minimum number of spectral parameters that provided the best overall accuracies.

In the final part of this thesis spatial modelling techniques were used to proactively identify pine forests that are highly susceptible to *S. noctilio* infestations. For the first time the random forest algorithm was used in conjunction with geographic information systems for mapping pine forests that are susceptible to *S. noctilio* infestations. Overall, there is a high probability of *S. noctilio* infestation for the majority

(63%) of pine forest plantations located in Mpumalanga, South Africa. Compared to previous studies, the random forest model identified highly susceptible pine forests at a more regional scale and provided an understanding of localized variations of environmental conditions in relation to the distribution of the wasps.

For my parents

Acknowledgments

I completed my PhD research because of the support and knowledge that I received from several individuals and institutions. I would like to thank the University of KwaZulu-Natal for giving me the opportunity to read for a PhD. My gratitude extends to Sappi and the National Research Foundation for providing me with the necessary funding.

I would like to thank my supervisor, Professor Onisimo Mutanga for guiding me every step of the way. Oni, I am indebted to you for teaching me to be more of a scientist and less of a number cruncher. Your determination is unrelenting and your commitment to your students is remarkable.

To Professor Urmilla Bob, thanks for the steadfast guidance and support that you have shown me throughout my academic career. You were always prepared to listen to my academic and social problems. You are truly an inspiration. I would also like to thank Professor Lalit Kumar and Professor Fethi Ahmed for helping me refine my ideas on remote sensing.

The Department of Geography at the University of KwaZulu-Natal supported me throughout my academic career. I would like to thank, Professor Brij Maharaj, Professor Fredric Giraut, Dr Denis Rugege, Dr Vadi Moodley, Craven Naidoo, Johnny Lutchmiah, and Shani Ramsroop. In particular I would like to thank my fellow remote sensor and friend, Michael Gebreslasie for providing me with valuable advice and support.

Thanks to Marcel Verleur and Andre de Wet for critically answering all my questions on *Sirex noctilio.* Based on their hands on approach to pest management, I gained a better understanding of forest health in South Africa. Thanks to Chris Muncaster for assisting with all the computer programming. I would like to also thank Dean Bethell and George Grossi for making it possible for me to complete my PhD while I was employed at Sappi.

With gratitude I remember my family and friends who contributed to my overall well-being. Finally, to my wife Sumaiya and my daughter Zahraa, thanks for the unequivocal love and support.

CHAPTER 1:

General introduction

1.1. Introduction

A key component of forest productivity is forest health and the effect it has on producing a sustainable yield. Emerging evidence suggests that new pests and pathogens are appearing at an increasing rate and that these damaging agents could impact on the future sustainability of the South African forestry industry (Wingfield et al., 2001). For example, the siricid, *Sirex noctilio* (Fabricius), has caused considerable tree mortality in the southern parts of the country with researchers estimating that 35,000 ha of pine forest worth an estimated 45 million dollars are infested and dying (Hurley et al., 2008). To exacerbate matters, *S. noctilio* attacks all commercial pine species with none of the species showing a high resistance to infestation (Tribe and Cillie, 2004). As a result, approximately 72,000 ha of pine forests are potentially susceptible to *S. noctilio* infestation. In an effort to minimize the potential threat of *S. noctilio* to pine forest in South Africa, an integrated management strategy combining detection and monitoring methods, silvicultural treatments, and biological controls has been implemented on an industry wide basis in South Africa. This study is focused on and advocates the development of techniques within remote sensing to accurately detect and map *S. noctilio* infestations. Once developed and tested, remote sensing technology could significantly improve the detection and mapping of *S. noctilio* infested pine forests more accurately and effectively than current methods of assessment.

1.2. Understanding *Sirex noctilio* infestations

In its natural habitat, *S. noctilio* (Figure 1.1a) typically attacks stressed pine trees (Neumann and Minko, 1981). However, as population levels increase, the wasp spreads through mature pine compartments and causes extensive mortality of larger trees (Ciesla, 2003; Haugen, 2000). The female wasp inserts her ovipositor through the bark into the sapwood and deposits up to three separate eggs at a drill site (Figure 1.1b). During the process the female also introduces toxic mucus and the symbiotic fungus *Amylostereum areolatum* into the wood (Slippers et al., 2003). The mucus changes the water balance of the tree, thereby creating conditions that are ideal for the growth and spread of the fungus. Subsequently, the fungus rots and dries the wood, providing a

suitable environment for the survival and development of the insect larvae (Slippers et al., 2003). The combined effects of the mucus and the fungus cause the tree to die. *S. noctilio* eggs hatch and the larvae (Figure 1c) feed on the fungus while tunnelling through the wood towards the centre of the tree. Adult wasps bore their way out of infested trees and leave a characteristic round exit hole (Figure 1d). In the summer rainfall areas of South Africa, the wasp emerges between late October and early January with peak emergence in November (Hurley et al., 2008).

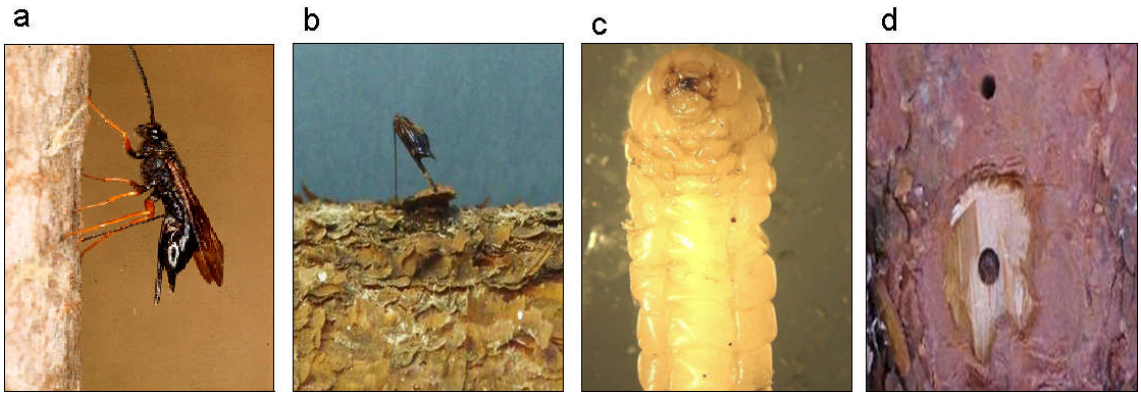


Figure 1.1: (a) *S. noctilio*. (b) A female wasp depositing her eggs in a pine tree. (c) *S. noctilio* larvae. (d) Characteristic round exit holes found in infested trees.

The primary biological control of *S. noctilio* populations is achieved by using the nematode *Deladenus siricidicola* and parasitic wasps such as *Ibalia leucospoides* and *Megarhyssa nortoni*. Silvicultural methods such as thinning are also carried out to improve tree vigour (Haugen, 1990). However, successful implementation of the above control measures depends on the ability to spatially quantify the severity and extent of infestation so that forest managers can adopt the most appropriate course of intervention before the forest reaches a point of non-recovery. Additionally, geographic information systems (GIS) and forest planning systems, which include harvesting schedules, timber volume analysis, and species growth models, have been developed to help foresters manage infested areas. These systems require accurate spatial information on the severity and extent of *S. noctilio* infestations.

1.3. Challenges and opportunities: Remote sensing of *Sirex noctilio* infestations

Current methods used to spatially quantify the severity and extent of *S. noctilio* infestation includes broad scale visual aerial reconnaissance followed by field-based exercises to verify the results. The effectiveness of visual assessments is questionable because the method is qualitative, subjective, and dependent on the skill of the surveyor (McConnell et al., 2000; Stone and Coops, 2004). The ability of remote sensing technology to successfully detect and map forest health has been demonstrated by researchers for a diverse range of forest pests and pathogens, imagery types and modelling techniques (Coops, 2006; Coops et al., 2003; Kelly et al., 2007; Lawrence and Labus, 2003; Leckie et al., 2004; Pontius et al., 2005a; Radeloff et al., 1999). As such, the technology has the potential to ensure that the detection and mapping of *S. noctilio* infestations is an achievable task, provided that the observed symptoms of infestation can be detected using spectral reflectance. However, to date, no research has demonstrated the use of remote sensing technology for the successful detection and mapping of pine forests infested by *S.noctilio.*

Remote sensing has distinct advantages over the current methods of assessment. Digital remote sensing technologies measure the amount of electromagnetic energy reflected from the leaves and canopy of the tree using a number of wavelengths which can range from 350 nm to 2500 nm. Researchers have used this spectral information, in the form of individual bands, band combinations, and vegetation indices to detect and map forest health (Coops et al., 2003; Entcheva et al., 2004; Pontius et al., 2005a). Additionally, remote sensing technology can image large areas and allow for the repetitive monitoring and assessment of tree damage and mortality (Cosmopoulous and King, 2004; Jin and Sader, 2005). Finally, remotely sensed data is usually acquired in a digitized and spatially explicit format that allows for integrated visualization and modelling across a range of operational scales using GIS (Kelly et al., 2007).

However, for remote sensing to be effective and accurate in detecting and mapping *S. noctilio* infestations, a sound understanding of the progression and pattern of symptoms of *S. noctilio* infestations across leaf, canopy, and landscape levels is required. Knowledge of these symptoms allows for the development of algorithms to detect changes in foliar characteristics using remotely sensed data. Consequently,

the challenge would be to assess the observed symptoms of *S. noctilio* infestation (Figure 1.2), and then associate each level of observation with different remote sensing data types in order to provide the appropriate level of detail and accuracy for detection and mapping purposes.

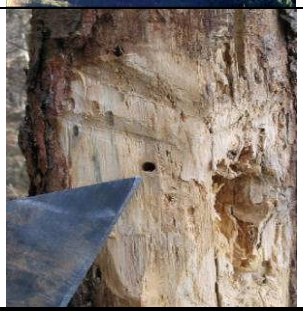
Stages of attack		**Symptoms**
Healthy		No signs of S. *noctilio* infestation.
Green		The appearance of resin droplets and the presence of ovipositors on the bark with a dark fungal stain appearing along the cambium. There is minimal needle loss and the canopy appears green and healthy.
Red		Severe chlorosis results in the canopy of the attacked tree changing colour from green to yellow to reddish brown. Larvae are present in the tree. There is very high needle loss and there is a scattering of dead and dying trees in the plantation.
Grey		The longest part of the *S. noctilio* life span is spent inside the host tree and by chewing round exit holes, the wasps eventually emerge as adults. The canopy is completely defoliated with 100% needle loss.

Figure 1.2: Description of the various stages of *S. noctilio* infestation.

1.3.1. The green stage of infestation

During the initial or green stage of *S. noctilio* infestation, the canopy of the infested tree appears green and visually indistinguishable from healthy trees (Ciesla, 2003). The success in discriminating the initial stages of infestation is dependent on detecting subtle changes in the spectral reflectance of the tree (Ekstrand, 1994). However, slight changes in the spectral reflectance of stressed vegetation, when measured by various broad band sensors, are often masked by the high degree of variation in reflectance caused by factors such as varying view geometry, illumination, and canopy density (Runesson, 1991). Given these limitations, there is strong optimism that high spectral resolution data (hyperspectral) will allow for the effective discrimination of the green stage of *S. noctilio* infestation because the data allow for the detection of detailed features using many narrow bands which would have been otherwise masked by broad band sensors (Schmidt and Skidmore, 2001).

Due to the cost and availability of airborne hyperspectral imagery in South Africa, the focus of this thesis was restricted to hyperspectral data captured in the field and under controlled laboratory conditions. Nevertheless, by using laboratory or field hyperspectral measurements to determine which band or band combinations offer the maximum information content to discriminate the green stage, this study hopes to establish an important prerequisite for the potential upscaling of results to either an airborne or spaceborne platform. This is especially pertinent since it is envisaged that South Africa will soon launch the ZASat-003 satellite that will carry a hyperspectral sensor (van Aardt and Coppin, 2006) thus making airborne hyperspectral data relatively more accessible and available to remote sensing researchers in the country.

However, it should be noted that hyperspectral data, whether captured at laboratory, field, or airborne platforms, tend to be more difficult to process than multispectral data due to the geometrical and statistical properties associated with high dimensional data (Langrebe, 2002). The challenge would be to develop and test robust methods and techniques for the effective processing and classification of hyperspectral data. Additionally, these methods and techniques need to be automated to some level with limited human interaction to allow for critical evaluation (Soh, 1999).

1.3.2. The red stage of infestation

The red stage of attack occurs approximately three months after oviposition, when the canopy of the infested tree wilts and changes colour from green to yellow to reddish brown (Ciesla, 2003). The commercial availability of digital multispectral imagery (DMSI) in South Africa offers a potential data source for the effective collection of spatially accurate, consistent, and timely imagery. Researchers focusing on other forest pests and pathogens have successfully used high spatial resolution DMSI to quantify declining forest health (Leckie et al., 2004). Compared with satellite imagery, high spatial resolution DMSI is capable of achieving higher mapping accuracies by identifying individual crowns. This is particularly useful since pine plantations infested by *S. noctilio* have a scattering of dead and dying trees (Haugen et al., 1990; Haugen and Underdown, 1990), and there is a need to remotely identify small clusters or individual trees. An additional advantage of using airborne DMSI, is its ability to obtain imagery at opportunistic times and at user specified locations (Wulder et al., 2004b).

However, it is unrealistic to expect that a single spatial resolution will be both sufficiently detailed and suitably cost-effective to capture variable infestation levels (%) at a compartment, or even at a broader plantation scale. DMSI is available at a very fine spatial resolution (0.5 m) in South Africa, and there will be a tendency for forestry companies to use the data at the finest spatial resolution available. According to Menges et al. (2001) using remotely sensed data with spatial resolutions finer than the structure of the vegetation may introduce unnecessary variation and this could result in large data volumes and unnecessary costs. The challenge would be to develop methods where each object (that is, infested tree crown) under investigation can be considered at its optimal spatial resolution (Marceau et al., 1994) and where the information content per pixel is maximized (Atkinson, 1997).

1.3.3. Modelling the susceptibility to Sirex noctilio infestations

Thus far, the researcher has attempted to identify the challenges and opportunities regarding the detection and mapping of existing green and red stage *S. noctilio* infestations. However, the strength of detection and mapping methods would be greatly

enhanced if one could proactively identify pine forests that are highly susceptible to *S. noctilio* infestations before any concerted detection and mapping methods are implemented. Maps showing the distribution of susceptible forests will then serve as a spatial guide and allow forest managers to focus their existing detection and mapping efforts to these key areas (hotspots). In this regard, statistical modelling approaches have been increasingly recognized as important tools that improve our understanding of forest pests and pathogens. When used within a GIS framework, these models have the ability to identify areas that are highly susceptible to infestations (Candau and Fleming, 2005; Carnegie et al., 2006; Guo et al., 2005; Kelly and Meentemeyer, 2002; Negron, 1998; Rosso and Hansen, 2003; van Staden et al., 2004). Therefore, the challenge would be to model pine forests that are susceptible to *S. noctilio* infestations in an effort to enhance green and red stage detection and mapping initiatives.

1.4. Aim

Given the above discussion, the aim in this study is to develop and test methods within remote sensing to detect and map *S. noctilio* infested *Pinus patula* trees in KwaZulu-Natal, South Africa.

1.5. Objectives of the study

The main objectives of this study are as follows:

1. To determine if vegetation indices derived from high spatial resolution DMSI can characterize *S. noctilio* induced stress at a canopy level.

2. To define the appropriate spatial resolution that will allow for the accurate detection and mapping of *S. noctilio* infestations at a canopy level.

3. To determine the utility of hyperspectral data in discriminating among the healthy, green and red stages of *S. noctilio* infestation

4. To determine if machine learning algorithms can accurately discriminate between healthy trees and the green stage of *S. noctilio* infestations using resampled HYMAP data.

5. To quantify *S. noctilio* induced water stress in *P. patula* trees using regression tree ensembles and hyperspectral indices.

6. To model the potential distribution of pine forests that are susceptible to *S. noctilio* infestations.

1.6. Description of the study area

S. noctilio infestation levels have reached epidemic proportions in KwaZulu-Natal, with 30% or more of *P. patula* trees being killed in some plantations (Slippers, 2006). The study area includes two commercial forestry plantations located in the southern parts of the province (Figure 1.3a).

The first study area (centroid $30°4'13.83" E$ and $29°38'36.06" S$) is located at the Sappi Pinewoods plantation which is dominated by *P. patula* plantings (Figure 1.3b). The site is located approximately 30 km west of the town of Pietermaritzburg, KwaZulu-Natal. The average altitude for the site is 1,190 m with an average air temperature of 16.1 °C (Macfarlane, 2004). The mean annual rainfall of the area is 916 mm. The terrain consists of low mountains and undulating hills. The geology of the area is a mixture of mudstone, sandstone, tillite, ampholite, and basalt. Soils in the area are mostly sandy-clay and sand-clay loams (Macfarlane, 2004).

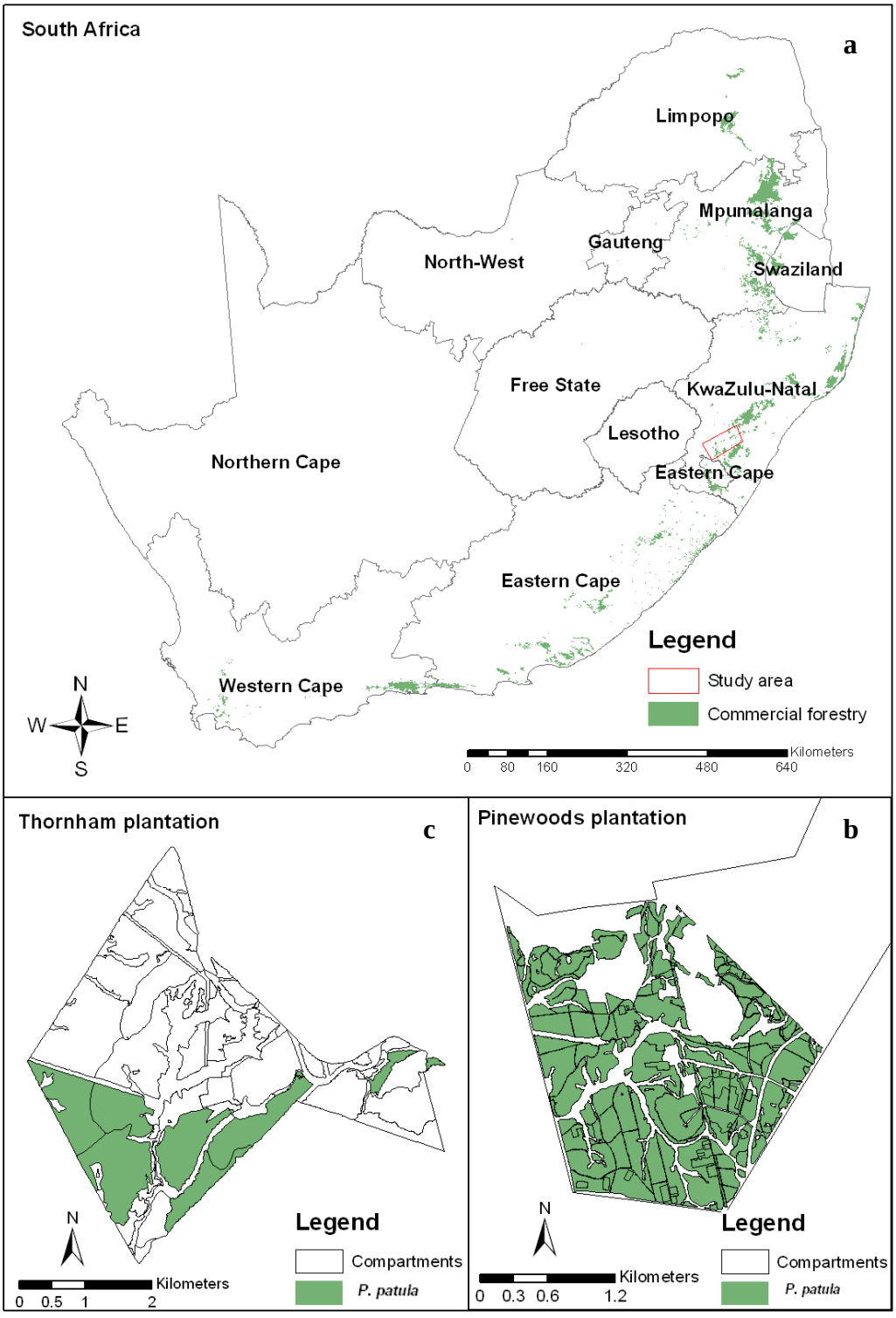


Figure 1.3: Location of the study area. Insert (b) shows the Pinewoods plantation while insert (c) shows the Thornham plantation.

The second study area is part of the Mondi Thornham plantation (Figure1.3c). The area (centroid $29°42'20.02''E$ and $29°52'30.26''S$) lies at an altitude of 1,500 m above sea level with frost occurring in most areas between May and September (Schulze et al., 1997). Rainfall varies between 800 mm and 1,200 mm per year, with high rainfall experienced predominately during the mid-summer months (Schulze et al., 1997). Lithology is predominantly shale, and to a lesser extent dolerite. Soils are characterized by fine sandy clay and humic topsoil which are underlain by yellow or red apedal subsoils (Schulze et al., 1997).

1.7. Outline of thesis

Besides the introduction and the synthesis, each chapter in this thesis is written as an individual paper that was submitted to peer-reviewed journals. The name of each journal as well as the title of the paper is mentioned at the beginning of each chapter.

In Chapter 2 the ability of vegetation indices derived from high spatial resolution DMSI to characterize *S. noctilio* induced stress in *P. patula* compartments is examined. The relative strength of various ratio and linear based vegetation indices are then tested for discriminating the various stages of *S. noctilio* infestations.

In Chapter 3 the necessary spatial resolution guidelines are established for the operational detection and mapping of *S. noctilio* infestations at compartment or plantation scales. The work described in Chapter 2 is extended by using the minimal variance of high spatial resolution DMSI to define appropriate pixel sizes that will capture the spatial variability associated with *S. noctilio* infestations.

In Chapter 4 the ability of hyperspectral data to discriminate between the different stages of *S. noctilio* infestations is investigated. The study determines if there is a significant difference between wavelengths located in the visible-near infrared region (400 nm to 1300 nm) and the healthy, green, and red stages of infestation. For the wavelengths that are significantly different ($p < 0.001$) in this spectral region, it was tested whether some wavelengths have more discriminating power than others.

In Chapter 5 the potential of machine learning algorithms and resampled HYMAP data to accurately discriminate between healthy trees and trees in the green stage of *S. noctilio* infestation is investigated. More specifically, the random forest algorithm and variable selection methods are used to produce the smallest subset of HYMAP wavelengths that will allow for the accurate classification of healthy and green stage spectra.

In Chapter 6 regression tree ensembles are compared for predicting *S. noctilio* induced water stress in *P. patula* trees using several spectral parameters derived from hyperspectral data. Using bagging, boosting, and random forest ensembles, the study examines the ability of simple ratios, normalized ratios, three band ratios and continuum removed wavelengths to assay the water status of *S. noctilio* infested *P. patula* trees.

In Chapter 7 the random forest algorithm is implemented within a spatial framework to determine which pine forests in an unaffected area (that is, Mpumalanga) are highly susceptible to *S. noctilio* infestations. It is assumed that if pine forests in Mpumalanga share similar environmental conditions with those areas with confirmed *S. noctilio* infestations in KwaZulu-Natal, they are more likely to be susceptible to infestation. More specifically, the robustness of random forest algorithm is examined, firstly, in terms of its classification accuracy and secondly, for the empirical selection of explanatory variables.

In Chapter 8 a synthesis is provided of the research carried out and all the findings of individual chapters are brought into perspective.

CHAPTER 2:

The detection and mapping of *Sirex noctilio* infestation using high spatial resolution imagery

* This chapter is based on:

Ismail, R., Mutanga, O. and Bob, U., 2006. The use of high resolution airborne imagery for the detection of forest canopy damage by *Sirex noctilio*. In: P.A. Langin and M.C. Antonides (Editors), Precision forestry in plantations, semi-natural areas and natural forest: proceedings of the international precision forestry symposium. Stellenbosch University, Stellenbosch University, South Africa, pp. 119-134.

Ismail, R., Mutanga, O. and Bob, U., 2007. Forest health and vitality: The detection and monitoring of *Pinus patula* trees infected by *Sirex noctilio* using digital multispectral imagery (DMSI). Southern Hemisphere Forestry Journal, 69(1): 39-47.

Abstract

The Eurasian woodwasp, *Sirex noctilio* is causing considerable tree mortality in commercial pine plantations in southern KwaZulu-Natal. Broad scale visual assessments of infestation provided by forest managers are currently used to measure forest health and vitality. The effectiveness of visual assessments is questionable because they are qualitative, subjective, and dependent on the skill of the surveyor. Remote sensing technology provides a synoptic view of the canopy and thus offers an alternative to the conventional methods of monitoring forest health and vitality. In this study, high spatial resolution (0.5 m x 0.5 m) digital multispectral imagery (DMSI) was acquired over commercial *Pinus patula* trees of varying age classes which had been ground assessed and ranked on an individual tree crown basis using a severity scale. The severity scale was based on a hierarchy of symptoms of decline that are visibly apparent on the infested tree and are represented in this study as the green, red, and grey stages. A series of ratio and linear based vegetation indices were then calculated and compared to the different crown condition classes as determined by the severity scale. Of the vegetation indices derived from the high resolution DMSI, significant differences between the pre-visual (healthy and green stages) and visual (red and grey stages) crown condition classes were obtained. Canonical variate analysis further revealed that the best discriminatory power between the different crown condition classes is obtained when using the normalized difference vegetation index (NDVI) Overall the study demonstrated the potential benefit of using high resolution DMSI to discriminate between healthy trees and trees that were in the visual stage of infestation.

Keywords: *Sirex noctilio*, remote sensing, digital multispectral imagery, vegetation indices

2.1. Introduction

There are approximately 1.5 million hectares of commercial forest in South Africa (Zwolinski et al., 1998) with forest products contributing 1.2% (approximately 1.4 billion US dollars) to the gross domestic product (GDP) of the country (DWAF, 2005). The industry depends almost exclusively on the planting of exotic *Pinus, Eucalyptus,* and *Acacia* species (van Staden et al., 2004). However, emerging evidence suggests that new pests and pathogens are appearing at an increasing rate and could potentially impact on the future sustainability of the industry (Wingfield et al., 2001). *Sirex noctilio* (Fabricius), which was first detected in 1994 in the Western Cape (Tribe, 1995; Tribe and Cillie, 2004), is currently causing considerable tree mortality in commercial forest plantations in southern KwaZulu-Natal. In an effort to minimize the potential threat of *S. noctilio* to commercial pine production in the region, an integrated management strategy combining detection and monitoring methods, silvicultural treatments, and biological controls has been implemented on an industry wide basis in South Africa (Ismail et al., 2005).

The primary control of established *S. noctilio* populations is achieved by biological means using the nematode *Deladenus siricidicola* (Bedding) and parasitic wasps such as *Ibalia leucospoides* (Hochenwarth) and *Megarhyssa nortoni* (Cresson)*;* also silvicultural methods such as thinning are carried out to improve tree vigour and thereby keep damage within acceptable levels (Haugen et al., 1990; Ismail et al., 2005). However, successful implementation of the above control measures depends on the ability to spatially quantify the severity and extent of infestation so that forest managers can adopt the most appropriate course of intervention before the stand reaches a point of non-recovery. Additionally, geographic information systems (GIS) and forest planning systems, which include harvesting schedules, timber volume analysis, and species growth models, have been developed to help foresters manage infested areas. These systems require accurate spatial information on the severity and extent of *S. noctilio* damage. Current methods used to spatially identify the severity and extent of *S. noctilio* infestation includes broad scale visual aerial reconnaissance, followed by field-based exercises to verify the results. Although visual assessments of infestation are widely used to measure forest health (Haara and Nevalainen, 2002), the effectiveness of visual

assessments are questionable because they are qualitative, subjective, and dependent on the skill of the surveyor (McConnell et al., 2000; Stone and Coops, 2004).

Internationally, the use of remote sensing technology to detect, monitor, and map forest health over large areas has been a subject of great interest, resulting in the testing of a variety of airborne remotely sensed data, such as high spatial resolution digital multispectral imagery (DMSI) (Leckie et al., 2005), hyperspectral scanners (Coops et al., 2003), and video recorders (Yuan et al., 1991). The limited potential of satellite based methods is primarily due to the short time available for detection and the different responses at needle, branch, and canopy scales (Radeloff et al., 1999). In South Africa the commercial availability of DMSI offers a potential source for the effective collection of spatially accurate, consistent, and timely imagery regarding the impacts of *S. noctilio* at compartment level. High resolution DMSI (pixel sizes less than 1 m x 1 m) is capable of achieving higher mapping accuracies by identifying individual crowns. This is particularly useful because pine plantations infested by *S. noctilio* have a scattering of dead and dying trees (Haugen et al., 1990; Haugen and Underdown, 1990), and there is a need to remotely identify small clusters or individual trees. Additionally, the advantage of using airborne DMSI is its capacity to mobilize quickly at opportunistic times and at user specified locations (Wulder et al., 2004b). This is an important benefit for the monitoring of forest health and vitality because infection is often linked to other events, such as climate, disturbance, phenology of forest type and infecting agent (Stone and Coops, 2004). As a result, the date for image acquisition is important to maximize the discriminating potential of classification algorithms (Coops et al., 2003).

This study advocates the use of high spatial resolution DMSI and vegetation indices (VI) to provide a quantitative spatial framework for the detection and monitoring of *Pinus patula* trees infected by *S. noctilio*. The reason for using VI includes the removal of variability caused by canopy geometry, soil background, sunview angles and atmospheric conditions (Gilabert et al., 2002). Additionally, a number of VI have been successfully used to assess changes in the reflectance due to the declining health status of the tree (Leckie et al., 2004; Stone and Coops, 2004). For the purpose of this study VI are divided into two categories that is, ratio based indices

and linear based indices. For a complete review of VI see Jackson and Huete (1991) and Thenkabail et al. (2002).

To date, no research has examined the use of remote sensing technology for the detection and monitoring of *P. patula* trees infested by *S.noctilio.* The present study examines if VI derived from high resolution (0.5 m x 0.5 m) DMSI could characterize *S. noctilio* induced stress in *P. patula* compartments. The relative strength of various ratio and linear based vegetation indices is then tested for discriminating the crown condition classes (healthy, green, red and grey) associated with *S. noctilio* infestations. Once operational, these techniques could improve our ability to detect and map infested pine compartments more effectively than the current visual methods of assessment

2.2. Materials and methods

2.2.1. Description of the study area

The study area is approximately 1,750 ha and forms part of the Sappi Pinewoods plantation which is dominated by *P. patula* compartments (Figure 2.1). The site is located approximately 30 km west of the town of Pietermaritzburg, KwaZulu-Natal. The average altitude for the site is 1,190 m with an average air temperature of 16.1 °C (Macfarlane, 2004). The mean annual rainfall of the area is 916 mm. The terrain consists of low mountains and undulating hills. The geology of the area is a mixture of mudstone, sandstone, tillite, ampholite and basalt. Soils in the area are mostly sandy-clay and sand-clay loams (Macfarlane, 2004).

2.2.2. Description of the severity scale

Early evidence of attack, or the green stage, includes the appearance of resin droplets and the presence of ovipositors on the bark with a dark fungal stain appearing along the cambium (Neumann and Minko, 1981; Tribe and Cillie, 2004). There is minimal needle loss, and the canopy appears green and healthy. The red stage occurs later when the canopy of the attacked tree changes colour from green to yellow to reddish brown (Ciesla, 2003). Ultimately, during the grey stage, the tree canopy is completely

defoliated and round exit holes appear on the bark (Neumann and Minko, 1981). A new generation of adult wasps emerge resulting in a compartment of scattered pattern of dead or dying trees (Ciesla, 2003; Haugen et al., 1990; Haugen and Underdown, 1990). During the grey stage of attack the wood is totally desiccated (Haugen and Underdown, 1990) and the timber is not usable, and economic losses are incurred. Figure 2.2 provides a description of the severity classes that were used in this study.

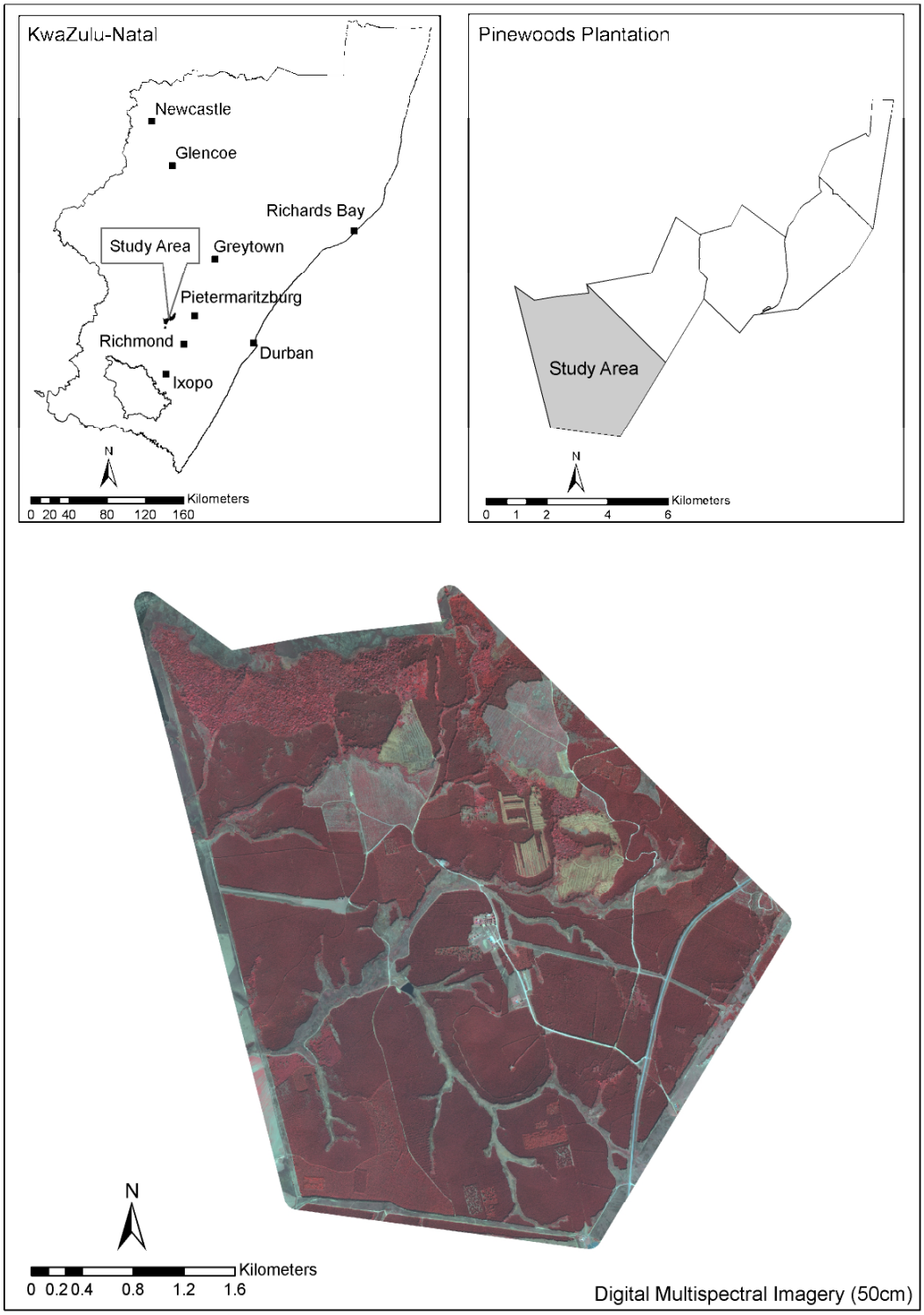


Figure 2.1: Location of the study area. The LrEye imagery was displayed using the near infrared, red and green bands.

Class	Stage	Crown condition		Symptoms
1	Previsual	Healthy		No signs of *S. noctilio* infestation.
2	Previsual	Green		Green crown, presence of resin droplets, cambium stain, ovipositors found on the trunk, and there is no needle loss.
3	Visual	Red		Severe chlorosis, reddish brown canopy, and high needle loss.
4	Visual	Grey		Emergence holes, no canopy, most branches intact, and 100% needle loss.

Figure 2.2: Description of the severity classes used for ground assessment of *S. noctilio* infestations.

2.2.3. Data acquisition

High resolution (0.5 m x 0.5 m) DMSI was acquired on the 9 September 2005 (10:00 GMT) by Land Resources International (LRI) Inc, Pietermaritzburg (South Africa) using the LrEye aerial imaging system. The LrEye sensor is composed of a series of

four monochrome Sony cameras. Each camera collects data for one of the spectral bands shown in Table 2.1. The resulting four bands were registered using Erdas Imagine (Erdas, 2004) to form an image with four co-registered spectral bands that are referenced to the Transverse Mercator projection (Hartebeesthoek datum, central meridian: 31).

Table 2.1: Spectral range of Landsat TM compared to the LrEye sensor.

Band	Color	Landsat TM spectral range (nm)	Landsat TM spatial range (m)	LrEye spectral range (nm)	LrEye spatial range (m)
1	Blue (B)	450 to 520	30	450 to 480	0.5
2	Green (G)	520 to 600	30	550 to 580	0.5
3	Red (R)	630 to 690	30	650 to 680	0.5
4	Near Infrared (NIR)	760 to 900	30	850 to 900	0.5

Field data collection took place one week after the image was acquired. A stratified random sampling technique (Richards and Jia, 1999) was adopted for this study. The strata were based on the age and occurrence of *P. patula* trees. Compartments that were harvested, or that were recently planted, were excluded from the sample. A 50 m x 50 m grid was generated over the study area, and 10 grid cells were randomly selected from each predetermined age stratum (that is, less than 7 years, from 8 to 9 years, 10 to12 years and older than 13 years). This age stratification was adopted because it reflects current *S. noctilio* management guidelines. At the centre point of each grid cell, a 10 m circular plot was created. Tree crowns located within each plot were manually identified on the LrEye imagery and subsequently located in the field using a global positioning system (GPS). In total, 782 trees were assessed for *S. noctilio* infections based on the severity scale that is shown in Figure 2.2.

This process was undertaken with the assistance of Sappi foresters and technical staff who have a detailed understanding of the identification and classification of *S. noctilio* infestations. Additionally, *P. patula* trees that were classified as having red stage infestations were destructively sampled to evaluate the presence or absence of larvae.

2.2.4. Evaluation of vegetation indices

According to Coops et al. (2003), the method used to obtain the spectral values of individual trees when using high resolution imagery is important because significant variations in brightness exist depending on the pixel position within the crown. In a study conducted by Leckie et al. (1992) to account for effects of the variation on individual crown delineation, it was concluded that either the whole tree or the sunlit tree sampling methods were the most suitable methods to derive consistent and representative spectral responses. In this study, the whole crown method was used whereby each of the selected crowns was manually delineated on the LrEye imagery, and the crown spectral response extracted for the ratio and linear based indices.

2.2.4.1. Ratio based indices

It has been reported that plants under stress display a decrease in canopy reflectance in the lower portion of the near infrared, a reduced absorption in the chlorophyll active region, and subsequently a shift in the red edge (Carter and Knapp, 2001). Ratio based indices have been successfully used to assess changes in the reflectance due to the declining health status of a tree (Ekstrand, 1994; Nelson, 1983; Vogelmann, 1990) because they operate by contrasting the intense chlorophyll pigment absorption in the red portion against the high reflectance in the near infrared (NIR) portion of the electromagnetic spectrum (Elvidge and Chen, 1995). The most widely used ratio based indices such as the ratio vegetation index (RVI) (Jordan, 1969), normalized difference vegetation index (NDVI) (Rouse et al., 1973), difference vegetation index (DVI) (Tucker, 1979), and the green normalized difference vegetation index (GNDVI) (Gitelson and Merzlyak, 1998) respond to these differences in the near infrared and visible regions (Lillesand et al., 2004). Table 2.2 shows the various ratio based indices that were used in the study.

Table 2.2: Ratio based vegetation indices used in this study.

	Vegetation Indices	Abbreviation	Equation	Reference
1	Normalized difference vegetation index	NDVI	NDVI = (NIR - red)/ (NIR + red)	Rouse et al. 1973; Jackson, 1983
2	Ratio vegetation index	RVI	RVI = NIR/red	Jordan, 1969
3	Difference vegetation index	DVI	DVI = NIR-red	Tucker, 1979
4	Green normalized difference vegetation index	GNDVI	GNDVI = (NIR – green)/ (NIR + green)	Gitelson and Merzlyak, 1998

2.2.4.2. Linear based indices

The tasseled cap transformation (TCT) converts the original spectral bands of a sensor into linear based indices (Jackson, 1983). Several studies using remotely sensed imagery (Collins and Woodcock, 1996; Healey et al., 2005; Jin and Sader, 2005; Price and Jakubauskas, 1998; Sharma and Murtha, 2001; Skakun et al., 2003) have shown the value of using the linear indices when assessing forest health and vitality. This is largely due to the fact that colour changes (chlorosis) associated with damaged trees is organized along the principal directions of the newly created linear based indices (Skakun et al., 2003).

The Gram-Schmidt orthogonalization process was used to derive the TCT coefficients for the linear based indices (Jackson, 1983). Initially, a soil line and the vector in the brightness direction are determined. Subsequently, from the brightness vector, all other vectors (greenness and yellowness) are orthogonally calculated. Coefficients (Table 2.3) are based on the grey level values of four land cover types (wet soil, dry soil, green vegetation, and senesced vegetation) found on the imagery. Water was used to represent wet soil values because pixels representing wet soils were not found in the imagery (Gong et al., 2003). Dry soil values were collected from dirt roads, and healthy tree crowns represented green vegetation. Dry grass values were used to represent senesced vegetation. Yarbrough et al. (2005) and Jackson (1983) provide a detailed mathematical description for calculating coefficients for *n* space indices using the Gram-Schmidt orthogonalization process.

Table 2.3: The Gram-Schmidt coefficients used in this study.

	B	**G**	**R**	**NIR**
Brightness (TCB)	0.337663	0.586272	0.638220	0.367348
Greenness (TCG)	-0.227113	-0.131965	-0.288569	0.920724
Yellowness(TCY)	0.097931	-0.781721	0.607311	0.102451

The resulting linear equations for brightness, greenness, and yellowness are as follows:

1. Brightness (TCB) = 0.337663 (blue) + 0.586272 (green) + 0.638220 (red) + 0.367348 (NIR)
2. Greenness (TCG) = - 0.227113 (blue) - 0.131965 (green) - 0.288569 (red) + 0.920724 (NIR)
3. Yellowness (TCY) = 0.097931 (blue) - 0.781721 (green) + 0.607311 (red) + 0.102451 (NIR)

2.2.5. Statistical analysis

Firstly, analysis was undertaken to compare the capacity of ratio and linear based indices to discriminate between the crown condition classes (Figure 2.2). This was tested using an analysis of variance (ANOVA) with a Tukey's HSD post hoc analysis (Coops et al., 2003) .

Secondly, canonical variate analysis (CVA) was used to determine which single VI best discriminated among the crown condition classes. CVA is a multivariate statistical technique which discriminates among pre-specified groups of sampling entities based on a suite of characteristics (McGarigal et al., 2000). The technique involves deriving linear combinations (that is, canonical functions) of two or more discriminating variables that will best discriminate among the *a priori* defined groups (Mutanga and Skidmore, 2005). In this study, VI's are entered into the analysis based on their ability to increase group separation (that is, crown condition classes). This reduces the number of indices to a subset that provides the best discrimination among classes. The best linear combination of VI are achieved by the statistical decision rule of maximizing the among group variance, relative to the within group variance (Mutanga and Skidmore, 2005). The first discriminant function provides the best separation

among classes, while the second function separates classes using information not used in the first function and so forth. Additionally, the functions will be independent or orthogonal, that is, their contributions to the discrimination between groups will not overlap (Lawrence and Labus, 2003).

Finally, we used the leave-one-out cross validation technique (n = 782) for estimating the error rate conditioned on the training data (Mutanga and Skidmore, 2005). The advantage of using the leave-one-out cross validation technique is that all the data are used for estimating error. Using this cross validation technique, each observation is systematically removed, the canonical function re-estimated, and the excluded observation classified (Mutanga and Skidmore, 2005). A confusion matrix is then constructed to compare the field (true) crown condition classes with the class assigned by the VI to the sample dataset. It depicts accuracies of the crown condition classes (producer's and user's accuracies). Producer accuracies are calculated by dividing the number of correctly classified trees in each crown condition class by the number of training data used for that class (that is, column total in the confusion matrix). User accuracies are computed by dividing the number of correctly classified trees by the total number of trees that were classified in that crown condition class (that is, row total in the confusion matrix). Additionally, a discrete multivariate technique called kappa analysis that uses the *k* (KHAT) statistic as a measure of agreement with the reference data was used (Congalton and Green, 1999; Skidmore, 1999). This statistic serves as an indicator of the extent to which the percentage correct values of an error matrix are due to true agreement versus chance agreement (Lillesand et al., 2004). If the kappa coefficients are one or close to one then there is perfect agreement between the training and test data.

2.3. Results

The hypothesis that ratio and linear based vegetation indices would discriminate among the various crown condition classes was tested by conducting a one-way ANOVA. Of the vegetation indices calculated, significant differences ($p < 0.001$) were obtained using NDVI, GNDVI, DVI, RVI, TCG, and TCB. A one-way ANOVA shows that there is a significant difference between the vegetation indices and the crown condition

classes, but it does not show which crown condition classes are different. Therefore a Tukey's HSD post hoc test was executed in order to establish differences between each of the crown condition classes (healthy, green, red, and grey). Results with their respective level of significance are shown in the table below.

Table 2.4: Analysis of variance results with a Tukey's HSD post hoc test. Class 1 (healthy), class 2 (green), class 3 (red), and class 4 (grey).

NDVI	**1**	**2**	**3**	**4**
1	..	**	*	*
2	**	..	*	*
3	*	*	..	*
4	*	*	*	..

TCG	**1**	**2**	**3**	**4**
1	..	**	*	*
2	**	..	*	*
3	*	*	..	*
4	*	*	*	..

GNDVI	**1**	**2**	**3**	**4**
1	..	**	*	*
2	**	..	*	*
3	*	*	..	*
4	*	*	*	..

TCB	**1**	**2**	**3**	**4**
1	..	**	**	*
2	**	..	**	*
3	**	**	..	**
4	*	*	**	..

DVI	**1**	**2**	**3**	**4**
1	..	**	*	*
2	**	..	*	*
3	*	*	..	*
4	*	*	*	..

RVI	**1**	**2**	**3**	**4**
1	..	**	*	*
2	**	..	*	*
3	*	*	..	*
4	*	*	*	..

$p < 0.001$ =*, Not Significant =**

Table 2.4 shows that both ratio (NDVI, RVI, DVI, and GNDVI) and linear based indices (TCB and TCG) are poor at discriminating between classes 1 (healthy) and 2 (green stage). However, the VI tested are capable of discriminating between the pre-visual (classes 1 and 2) and visual crown condition classes (classes 3 and 4). The most significant degree of separation occurs between class 1 and classes 3 and 4, and between class 2 and classes 3 and 4. All indices are capable of discriminating between these classes except for TCB which can discriminate only between class 1 and class 4 and between class 2 and class 4. Based on the results from ANOVA, it is difficult to determine which VI has the best discriminatory power. Therefore, a canonical variate analysis was carried out and included all indices (discriminatory variables) except for the TCB component. Additionally, to improve the discriminatory power of the VI, class

2 (green stage) was grouped with class 1 (healthy trees) while the rest of the classes remained the same, that is, class 3 (red stage) and class 4 (grey stage).

2.3.1. Canonical variate analysis

We tested the relative strength of various ratio and linear based vegetation indices in detecting *S. noctilio* infestations by carrying out a canonical variate analysis. Table 2.5 shows the eigenvalues as well as the factor structure matrix from the CVA using three crown condition classes (that is, healthy, red, and grey stages). The measure of information contained in the functions is represented by the eigenvalues corresponding to those functions. The eigenvalues are interpreted as the ratio of variances along each function (Richards and Jia, 1999). The largest portion of the explained variance (97.5%) is contained in the first canonical function and the remainder is contained in the second function (2.5%).

Table 2.5: Factor structure matrix representing the correlation between variables and canonical functions (3 classes).

VI	Function 1	Function 2
NDVI	0.633	0.369
GNDVI	0.629	0.605
DVI	0.559	0.550
TCG	0.500	0.669
RVI	0.484	0.463
Eigenvalue	0.961	0.025
% Variance	97.5	2.5

The factor structure coefficients contained in the matrix represent the correlations between the variables and the canonical functions, and are used to interpret the canonical functions (McGarigal et al., 2000). Results indicate that the highest factor structure coefficients are contained in the NDVI (0.633) and the GNDVI (0.629). The second canonical function also shows that one of the largest contributions is contained in the GNDVI (0.605) and to a lesser extent NDVI (0.369) however, the magnitude for the second canonical function is much smaller than that of the first canonical function. The classification accuracy based on the highest factor structure (NDVI) is shown in Table 2.6.

Table 2.6: Confusion matrix showing the NDVI predicted accuracy of *S. noctilio* infestations using a three level classification system: class 1 (healthy and green), class 2 (red), and class 3 (grey).

Class	**1**	**2**	**3**	**User accuracy**
1	695	2	2	99.43
2	8	26	3	70.27
3	2	3	11	68.75
Producer accuracy	98.58	83.87	68.75	
KHAT	0.79			

2.4. Discussion

High resolution DMSI provides a useful and robust tool to improve the ability to detect and monitor *P. patula* trees infected by *S.noctilio.* Ratio (NDVI, RVI, DVI, and GNDVI) and linear based vegetation indices (TCG) derived from high resolution DMSI are able to significantly ($p < 0.001$) discriminate between the pre-visual (healthy and green) and the visual stages of infestations (red and grey). CVA further revealed that best discrimination between the different crown condition classes (Figure 2.2) is obtained when using NDVI. Accuracy assessments (Table 2.6) show that NDVI derived from high resolution DMSI is successful in locating and predicting the condition of tree crowns on the imagery when crown condition classes are reduced to a three classification system, in which case the KHAT value was 0.79. The results obtained from this study are comparable to previous international studies on declining forest health (Leckie et al., 2005; Leckie et al., 2004; Vogelmann, 1990; Wulder et al., 2004a) and emphasize the importance of the visible and NIR bands when studying the effects of declining forest health, especially when infestation results in foliar discoloration (that is, the red stage).

Detecting and monitoring the red stage of infestation is regarded as a priority among forest managers because it gives an accurate indication of the severity and extent of infestation that is taking place that year (current infestation) (Leckie et al., 2005). Additionally, using high resolution DMSI to map out the red stage of infestation provides forest managers with a spatial framework that allows for the repetitive and

cost-effective monitoring over large areas. This improves the ability to quantify the severity and extent of *S. noctilio* infestations, thereby allowing forest managers to design the most appropriate intervention measures. For example, moderate red stage *S. noctilio* infestations (< 10%) would require the inoculation of infested trees with nematodes, whereas heavy infestations (between 10% and 50%) would require sanitization and salvage operations to be implemented (Haugen et al., 1990; Haugen and Underdown, 1990). Figure 2.3 shows the remote sensing work flow that was carried out to spatially quantify red stage *S. noctilio* infestation using high spatial resolution DMSI.

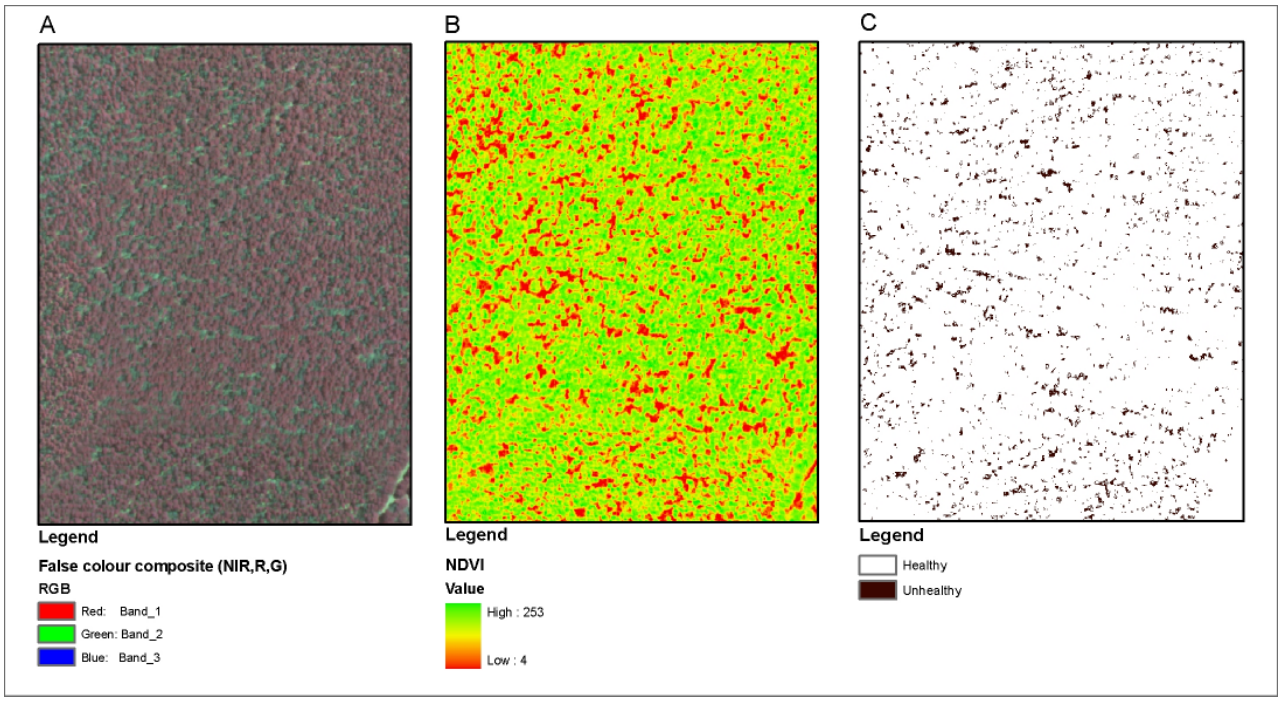


Figure 2.3: Remote sensing work flow showing the operational use of remote sensing technology for the detection and mapping of *S. noctilio* red stage infestations. Insert (a) shows the high spatial resolution (50 cm) DMSI. Insert (b) shows the NDVI image and insert (c) shows the *S. noctilio* infestation map which is based on reclassified NDVI values.

The difficulty in discriminating the green stage of infestation is consistent with other studies that have attempted to classify light to moderate symptoms using high resolution remotely sensed imagery (Leckie et al., 2005; Leckie et al., 2004). The

success of discriminating green stage infestation is dependent on the detection of subtle changes in the spectral reflectance of the tree (Ekstrand, 1994). Slight changes in the spectral reflectance of stressed vegetation, when measured by various broad band sensors, are often masked by the high degree of variation in reflectance caused by factors such as varying view geometry, illumination, and canopy density (Runesson, 1991). Given these limitations, hyperspectral remote sensing offers possibilities to investigate the early stages of infestations based on narrow bands using the entire electromagnetic spectrum. These narrow bands allow for the detection of detailed features which would otherwise have been masked (Schmidt and Skidmore, 2001).

Previous studies (Collins and Woodcock, 1996; Skakun et al., 2003) found changes in the tasseled cap wetness component (TCW) to be a good indicator of conifer mortality and the most consistent indicator of forest change due to the inclusion of the short wave infrared (SWIR) band. In this study the calculations of the tasseled coefficients were limited to four spectral bands found in the visible and NIR parts of the spectrum (400 nm to 900 nm) and therefore only included the TCB, TCG, and TCY and not the TCW. Additionally, spectrometer research conducted by Leckie et al., (1988) regarding discoloration caused by the spruce budworm indicated that the SWIR regions are better than the visible and NIR for discrimination. Similarly, initial attack by *S. noctilio* changes the water balance of the attacked tree (Neumann and Minko, 1981; Slippers et al., 2003), so using a sensor that captures SWIR wavelength could potentially improve the overall classification accuracy as well as the discrimination among the crown condition classes.

2.5. Conclusion

This study has shown that NDVI calculated from high spatial resolution DMSI has the potential to detect and monitor canopy damage caused by *S. noctilio*. Although it was difficult to discriminate between the healthy and green stages of infestation, classification accuracies are improved when using a three class crown condition index that differentiates the healthy, red and grey stages of infestation. Overall the study demonstrated the potential benefit of using high resolution DMSI to discriminate between healthy trees and trees that were in the visual stage of infestation. More

importantly, this has led to the development of a spatial monitoring framework that is capable of replacing traditional detection and monitoring methods.

Acknowledgments

Funding for this research was provided by the National Research Foundation (NRF).The digital multispectral imagery (DMSI) for this research was provided by Sappi Forests. We thank Marcel Verleur, Deane Bethell, James Thorpe, and Nic Tait for advice throughout the project, and Craven Naidoo (Cartographic Unit, University of KwaZulu-Natal) for assisting with the fieldwork. We also thank the referees for valuable comments on the paper.

CHAPTER 3:

Determining the optimal spatial resolution of multispectral remotely sensed imagery for the detection of *Sirex noctilio* infestations

* This chapter is based on:

Ismail, R., Mutanga, O., Kumar, L. and Bob, U., 2008. Determining the optimal resolution of remotely sensed data for the detection of *Sirex noctilio* infestations in *Pinus patula* plantations in KwaZulu-Natal, South Africa. The South African Geographical Journal, 90(1): 196-204.

Abstract

Sirex noctilio is causing considerable mortality in commercial pine plantations in KwaZulu-Natal, South Africa. The ability to remotely detect variable (for example, low, medium, and high) *S. noctilio* infestation levels remains crucial for mapping of the spread of the disease and for the effective deployment of suppression activities. Although high resolution image data can detect and map *S. noctilio* infestations, there are no guidelines that recommend which spatial resolutions are suitable for detection and mapping. This study examines the use of minimum variance to analyze *S. noctilio* infestations in an effort to determine an optimal spatial resolution of remotely sensed data for mapping forest health. High resolution (0.5 m) image data was collected using a four band airborne sensor, and infestation levels were derived using the normalized difference vegetation index (NDVI) and Gaussian maximum likelihood classifier. It was determined that the appropriate spatial resolution for the detection and mapping of *S. noctilio* infestations, as estimated by the minimum variance of sub-samples, narrowly differed based on the level of localized infestations present in the study area. Pixel sizes larger than 2.3 m will not provide adequate information for high infestation levels, while using pixel sizes smaller than the 1.75 m for detecting low to medium infestation levels will yield inappropriate results. The results of this study establish the necessary spatial resolution guidelines needed for the operational detection and mapping of *S. noctilio* infestations.

Keywords: *Sirex noctilio*, remote sensing, spatial resolution, minimum variance

3.1. Introduction

In its natural habitat the Eurasian woodwasp, *Sirex noctilio* typically attacks stressed pine trees (Neumann and Minko, 1981). However, as population levels increase, the wasp spreads through mature pine compartments and causes extensive mortality of larger trees (Ciesla, 2003; Haugen, 2000). *S. noctilio* infestation levels have reached epidemic proportions in KwaZulu-Natal, South Africa, with 30% or more of *Pinus patula* trees being killed in some plantations (Slippers, 2006). In an effort to minimize the economic threat to commercial forestry, management strategies that combine the use of remote sensing, silvicultural treatments, and biological control are currently being implemented (Ismail et al., 2005). The ability to remotely detect variable (for example, low, medium, and high) *S. noctilio* infestation levels remains crucial for the monitoring of the spread of the wasp and for the effective deployment of suppression activities (Ismail et al., 2006). For example, the ability to remotely detect light to medium *S. noctilio* infestations is beneficial because it allows forest managers to adopt a proactive course of remediation (for example, nematode inoculations) before the entire plantation reaches a point of non-recovery. However, it is unrealistic to expect that a single remotely sensed data source will be both sufficiently detailed and suitably cost-effective to capture variable infestation levels at a compartment, or even at a broader plantation scale.

The availability and accessibility of airborne sensors (for example, ArcEagle, LrEye, and Geospace) in South Africa has resulted in the increased acquisition of remotely sensed image data. However, as an increasing number of remotely sensed datasets become commercially available, the factor of spatial resolution plays an important role in the employment of remotely sensed image data (Quattrochi and Goodchild, 1997). Spatial resolution is defined as being the limit on how small an object on the earth's surface can be 'seen' by a sensor (for example, 2.4 m pixel for QuickBird and 4 m for IKONOS) as being separate from its surroundings (Lillesand et al., 2004). The basic information and measurement error contained in a remotely sensed image is strongly dependent on that spatial resolution (Atkinson, 1993; Woodcock and Strahler, 1987). For current *S. noctilio* detection and mapping, forestry companies tend to use the finest resolution (0.5 m) available. However, using remotely sensed data with

spatial resolutions finer than the structure of the vegetation community may introduce irrelevant variation and result in large data volumes and unnecessary cost (Menges et al., 2001). Methods need to be developed where each object under investigation can be considered at its optimal spatial resolution (Marceau et al., 1994), where the information content per pixel is maximized (Atkinson, 1997). Additionally, for remote sensing of forest ecosystems to become operational, spatial resolutions of remotely sensed image data must be appropriate for the specific application (Treitz and Howarth, 2000), and the data should be used with caution because of the potential problems that may arise from mismatches in scale between sensor and the practical requirements of the mapping exercise (Menges et al., 2001). The question then arises: on what basis should the investigator select an appropriate spatial resolution for the detection and mapping of *S. noctilio* infestations?

Previous research (Atkinson, 1993; Atkinson, 1997; Atkinson and Aplin, 2004; Woodcock and Strahler, 1987) has shown that the spatial variation between objects in a scene can be used to select an optimal spatial resolution and method of analysis for a given investigation. In forest environments, this relationship between the spatial variation in the objects of interest and spatial resolution has been described using average local variance (Woodcock and Strahler, 1987), semivariance (Colombo et al., 2004; Treitz and Howarth, 2000), minimal variance (Marceau et al., 1994; Menges et al., 2001) and spatial autocorrelation (Hyppanen, 1996). According to Marceau et al., (1994) the merit of using a minimal variance approach is that it considers different forest classes (for example, infestation levels) as opposed to an entire forest scene, thus making it a more suited indicator of the optimal spatial resolution for each particular class under investigation. Minimal variance is based on the assumption that when a pixel representing an object of interest is considerably larger (L-resolution) or smaller (H-resolution) than the object, the probability of selecting pixels across the image with different digital number (DN) values is high and this leads to a high variance (Marceau et al., 1994). However, when the pixel of the image data delineates the appropriate mixture of ground features composing the object under investigation, the variance is then at the lowest level (information content is maximized) and can be used as an indicator of the optimal spatial resolution required for the investigation (Marceau et al., 1994).

As part of a larger research effort aimed at reducing the effects of *S. noctilio* on pine production patterns in KwaZulu-Natal, Ismail et al. (2006) showed that high resolution image data can be used to detect and map *S. noctilio* infestations. However, there was no guideline to the appropriate spatial resolutions that are suitable for detection and mapping. This study aims to extend the work of Ismail et al. (2006) by using the minimal variance of classified NDVI images to define appropriate pixel sizes to capture the spatial variability of *S. noctilio* infestations. By establishing the spatial limitations of image data under variable infestation levels, this study hopes to contribute useful information and provide the necessary guidelines for the operational detection and mapping of *S. noctilio* at compartment or plantation scales.

3.2. Materials and methods

3.2.1. Study area

The study area is part of the Mondi Thornham plantation (Figure 3.1) and is situated in the Midlands area of KwaZulu-Natal. The area lies at an altitude of 1,500 m above sea level with frost occurring in most areas between May and September (Schulze et al., 1997). Rainfall varies between 800 mm and 1,200 mm per year, with high rainfall experienced predominately during the mid-summer months (Schulze et al., 1997). Lithology is predominantly shale, and to a lesser extent dolerite. Soils are characterized by fine sandy clay and humic topsoil which are underlain by yellow or red apedal subsoils. Dominant soil forms are Inanda and Magwa. Clay contents vary between 25% and 35% in topsoil horizons and attain values of up to 45% in subsoil horizons (Schulze et al., 1997).

3.2.2. Sirex noctilio infestations

As part of the current detection and monitoring framework used by the forestry industry, current *S. noctilio* infestations are determined by identifying the red stage of attack (Ismail et al., 2006). The red stage of attack occurs approximately three months after adult flight and oviposition, when the foliage of infested trees wilts and changes

colour from green to yellow to reddish brown (Haugen et al., 1990; Stone and Coops, 2004). Infestation levels are then categorized into the following damage classes: low (1% to 5%), medium (6% to 10%), high (11% to 15%), and severe (> 16%) (Croft, 2006). Similarly, in this study, infestation levels were calculated as the percentage of red stage trees to the total number of trees.

Table 3.1: Compartments selected for this study (n = 6).

Compartment	Age (years)	Area (ha)	Species	Planted stems per ha (SPH)	Current stems per ha (SPH)
B05	16	21.5	*Pinus patula*	1111	678
B06	16	22.8	*Pinus patula*	1111	732
B07	16	11.8	*Pinus patula*	1111	814
B17	16	23.7	*Pinus patula*	1111	690
B18	16	17.8	*Pinus patula*	1111	1045
B19	16	31.4	*Pinus patula*	1111	701

3.2.3. Selection of Sirex noctilio infested compartments

In order to prevent statistical bias, pine compartments (n = 6) with the same age and species were selected from the study area (Table 3.1), thus reducing the effects of structural parameters on the spatial resolution analysis (Hyppanen, 1996). For our purposes, additional localized sub-samples (50 m x 50 m grids) were then generated over the selected *P. patula* compartments to provide a more representative sample (n = 308) that could be used for further analysis. Previous field visits to the plantation have shown that localized samples consisting of 50 m x 50 m grids are adequate to detect *S. noctilio* infestations. Additionally, other studies examining the effects of spatial resolutions on vegetation mapping have also adopted a localized sub-sample approach in an effort to provide a relatively more representative sample size (Colombo et al., 2004; Murwira, 2003).

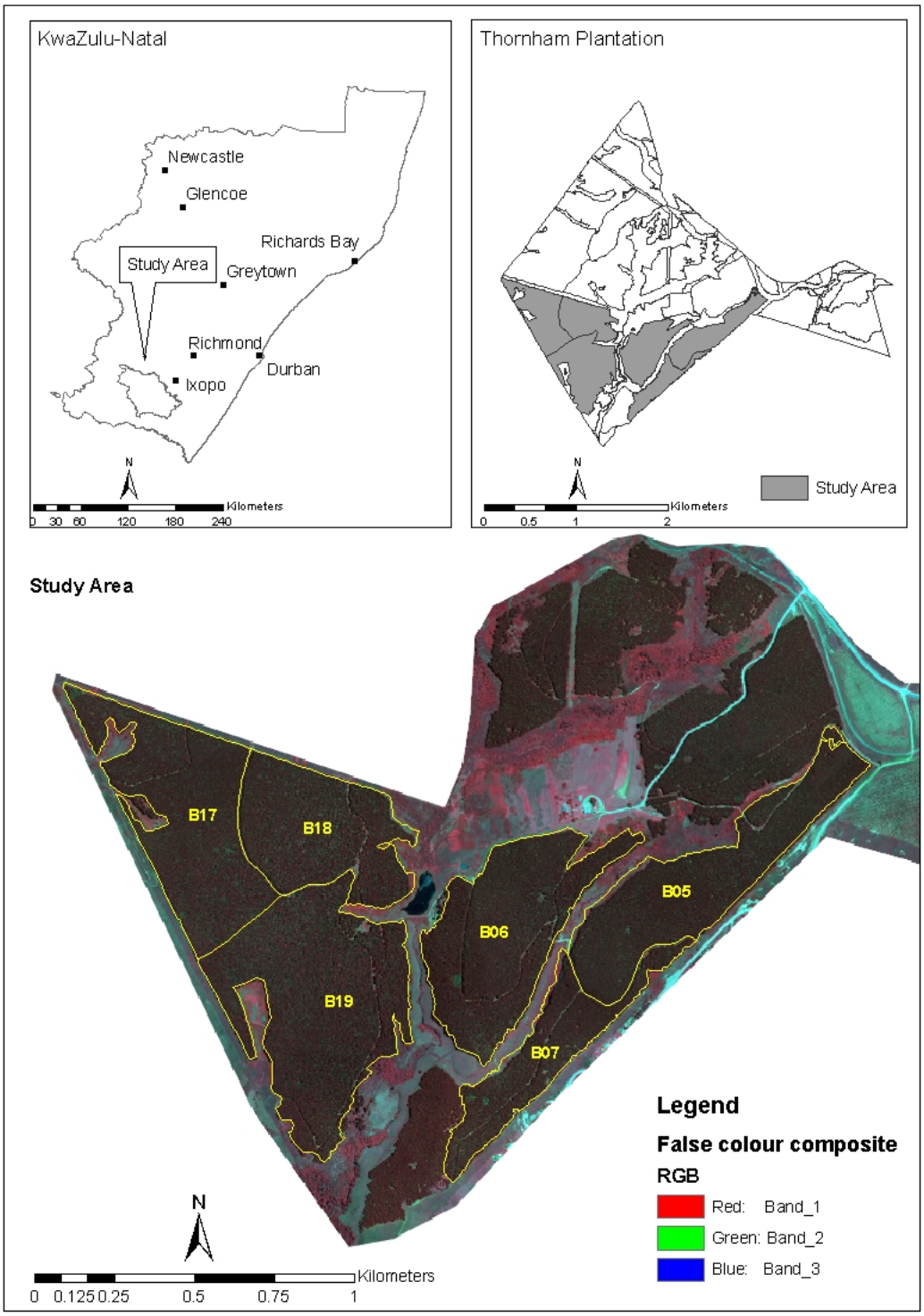


Figure 3.1: Location of the study area. Image data shown is a false colour composite consisting of the NIR, red, and green bands. Compartments selected for the study are indicated in yellow.

3.2.4. Description of image data

High resolution (0.5 m) image data was acquired on the 1 January 2006 by Land Resources International (LRI) Inc, Pietermaritzburg (South Africa) using the LrEye aerial imaging system. The LrEye sensor is composed of a series of four monochrome Sony cameras. Each camera collects data for one of the bands shown in Table 3.2. The resulting four bands are registered to form an image with four co-registered bands that are then referenced to the Transverse Mercator projection (Hartebeesthoek datum, central meridian: 29).

Table 3.2: Spectral and spatial range of the LrEye sensor.

Band	Colour	Spectral range (nm)	Spatial resolution (m)
1	Blue (B)	450 to 480	0.5
2	Green (G)	550 to 580	0.5
3	Red (R)	650 to 680	0.5
4	Near Infrared (NIR)	850 to 900	0.5

3.2.5. Image processing and analysis

A number of vegetation indices (VI) have been successfully used to assess the changes in reflectance due to the declining health status of the trees (Collins and Woodcock, 1996; Coops et al., 2004; Leckie et al., 2004; Vogelmann, 1990). Additionally, the advantage of using VI includes the removal of variability caused by canopy geometry, soil background, sunview angles, and atmospheric conditions (Gilabert et al., 2002). In this study, the normalized difference vegetation index (NDVI) was used to determine *S. noctilio* infestation levels within the study area. Investigators have shown that NDVI calculated from high spatial resolution (0.5 m) image data can successfully (79% classification accuracies) detect *S. noctilio* infestations (Ismail et al., 2006). In the present study, NDVI was derived from the high resolution image data (0.5 m) using Equation (1) and rescaled to the range of 0 to 255 in order to facilitate data handling in the image processing software.

$$NDVI = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}$$ (Rouse et al. 1973; Jackson, 1983) (1)

Where: λ_1, near infrared band (850 nm to 900 nm)

λ_2, red band (650 nm to 680 nm)

To obtain the training signatures, six localized sub-samples were randomly selected from the pine compartments shown in Table 3.1. Tree crowns located within each sub-sample were manually identified on the 0.5 m image data and subsequently located in the field using a global positioning system (GPS). In total, 111 trees were visually assessed for *S. noctilio* red stage of attack. To prevent errors of commission, trees identified as red stage trees (those having a reddish brown canopy) were destructively sampled to check for the presence of *S. noctilio* larvae. Results indicated that all trees identified as red stage trees were positive for *S. noctilio* infestations.

Using the training signature obtained from the study area, the NDVI image was then classified into binary classes of red stage pixels and healthy pixels by means of a Gaussian maximum likelihood (GML) classifier (Erdas, 2004).The GML classifier was used because it is relatively convenient to implement and more robust than other classification rules since it uses variances and covariance of training statistics as opposed to simpler statistics (Chen et al., 2004). Next, using the binary image, infestation levels (%) for the study area were calculated as the ratio of red stage pixels compared to the total number of pixels for each sub-sample (n = 308) generated over the study area. These sub-samples represented the variable infestation levels (%) for which the effects of spatial resolution could then be examined.

3.2.6. Minimum variance

The method for calculating the minimum variance for each localized sub-sample is relatively straightforward and is easily implemented in any image processing software. Firstly, to simulate variable spatial resolutions, the binary image data (0.5 m) as determined by the GML was successively resampled to coarser resolutions. The process involves calculating the average pixel value using odd sized *n* x *n* windows of

increasing dimensions (Table 3.3). According to Marceau et al. (1994), this averaging method is regarded as an efficient and simple way to represent the physical aggregation process of a sensor's instantaneous field of view (IFOV). Additionally, the nearest neighbour and cubic convolution algorithms used for resampling data, induce sharpening or smoothing effects that influences the analytical process (Bian and Butler, 1999).

Table 3.3: Window sizes used during the resampling process.

Window Size	Spatial Resolution (meters)
3 x 3	1.5
5 x 5	2.5
7 x 7	3.5
9 x 9	4.5
11 x 11	5.5
13 x 13	6.5
15 x 15	7.5

Next, the variance (Equation 2) at each sub-sample (n = 308) was calculated for all resampled spatial resolutions (binary images consisting of red stage pixels and healthy pixels). A similar process was adopted by Colombo et al. (2004) who used the semivariance of binary images (forest and non-forest pixels) to determine an appropriate spatial resolution for monitoring tropical forest cover.

$$Variance = \frac{\sum (x_{ij} - M)^2}{n-1} \tag{2}$$

$$Mean = \frac{\sum x_{ij}}{n} \tag{3}$$

Where: x_{ij} = DN value of pixel (i, j)

n = Number of pixels in a window

M = Mean of the moving window

Finally, the spatial resolution at which each sub-sample reaches a minimum variance was observed and tabulated. This spatial resolution was then averaged for each unique infestation level present in the study area and the resulting spatial resolution was defined as the optimal spatial resolution. However, according to Atkinson (1997), where the objective is to map the spatial variation of interest, the spatial resolution chosen should not be the spatial resolution defined in this study as optimal. It was suggested that the spatial resolution used should be much finer than the calculated optimal spatial resolution because the objective for mapping is not to maximize the amount of information per pixel but to ensure that there is sufficient information of interest to be sampled. According to sampling theorems, to effectively sample objects, one must sample at least at one-half the width of the object under investigation (McGrew and Monroe, 2000). Therefore, in this study, a pixel smaller than or equal to half the optimal spatial resolution would be an appropriate resolution for the detection and monitoring of *S. noctilio* infestations.

3.3. Results

3.3.1. Classification results

Figure 3.2a shows the derived NDVI image for the study area. The original NDVI values (-1 to 1) were rescaled to the range of 0 to 255 (Erdas, 2004). The lower limits of the range indicate the absence of vegetation, while the upper limits indicate very healthy vegetation. The derived NDVI values for the study area had a lower limit of 0 and an upper limit of 231. Box plots in figure 3.2b show the spread of NDVI values for the healthy (n = 37) and red stage trees (n = 74). The mean differences between the two groups were tested using a *t* test, and the normality of the data was assessed using a Kolomogrov-Sminov test. The results from the test indicated that NDVI values significantly differ between the red stage and healthy trees ($p < 0.05$). Consequently, the training samples (n = 111) were then used to classify the NDVI image into binary classes of healthy and red stage pixels.

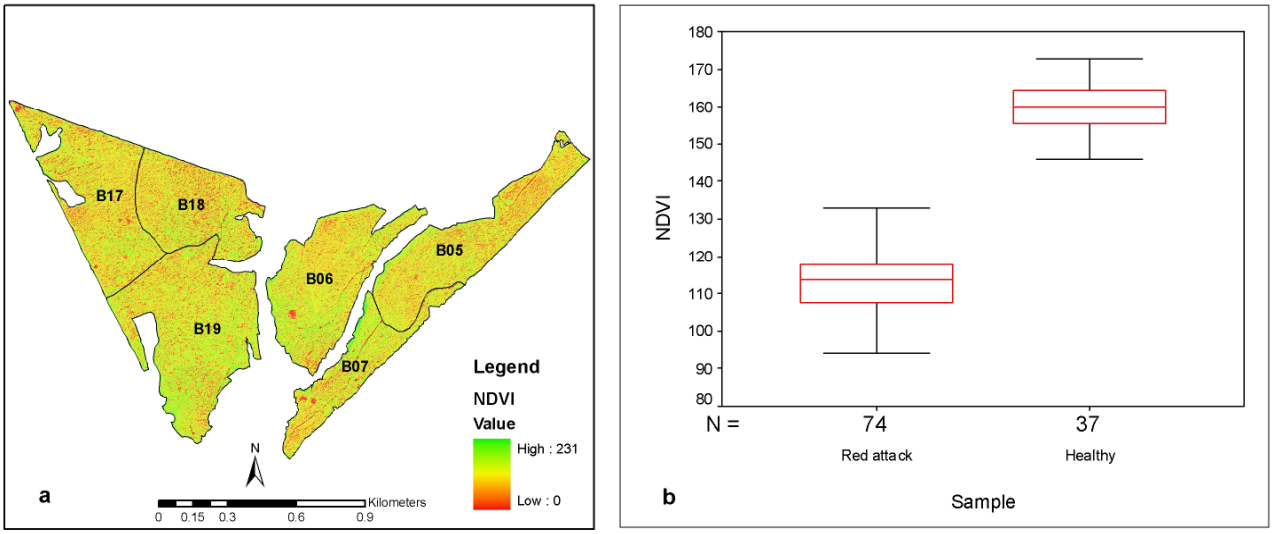


Figure 3.2: NDVI values derived from high spatial resolution (0.5 m) image data. Insert (a) shows the spatial pattern of NDVI values in the study area. Insert (b) shows the spread of NDVI values for healthy (mean = 160; standard deviation = 7.1) and red stage trees (mean = 113; standard deviation = 9.3).

The average *S. noctilio* infestation level for the Thornham plantation was 19.5%. Infestation levels calculated at a compartment level (n = 6) are shown in Table 3.4. For comparative purposes, independent forest health enumerations (field-based) are provided for the selected compartments (Croft, 2006). A Mann-Whitney *U* Test revealed that there was no significant difference ($p > 0.05$) between the field-based and the binary infestation levels. However, in order to determine an appropriate pixel size to capture the spatial variability of *S. noctilio* infestations, the localized sub-samples (50 m x 50 m grids) provided a more representative sample of infestation levels (n = 308) as opposed to infestations levels at a compartment level (n = 6).

Table 3.4: Comparison of *S. noctilio* infestation levels at a compartment scale.

Compartment	Sirex Infestation (%) (NDVI)	Sirex Infestation (%) (Field based results)
B05	24	16
B06	17	17
B07	4	3
B17	25	26
B18	26	35
B19	21	21

Based on the existing damage classes used by foresters, infestation levels were categorized into low (1% to 5%), medium (6% to 10%), high (11% to 15%), and severe (> 16%) classes (Croft, 2006). Figure 3.3 shows the variability of infestation levels throughout the study area as calculated for the sub samples. There are predominately medium to high infestation levels. More specifically, results show that 15.43% of the grid cells have low infestation levels, 51.77% have medium infestation levels, and 32.80% have high infestation levels. As expected there are no sub-samples with severe infestations levels (> 16%). Localized areas having severe infestation levels are easily identifiable by foresters and measures such as clear felling operations would have been implemented in order to salvage usable red stage trees and to prevent *S. noctilio* from spreading to other compartments in the plantation.

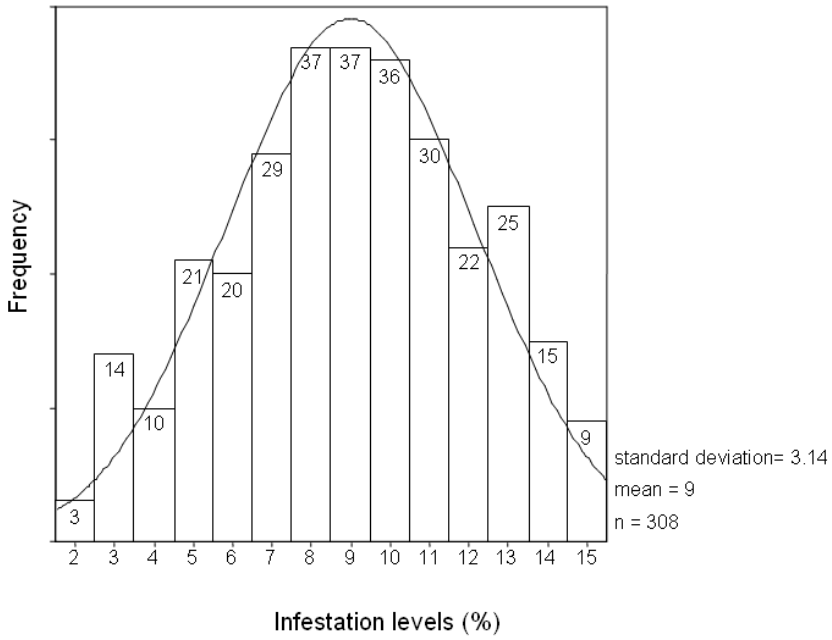


Figure 3.3: Histogram showing the infestation levels within the study area (50 m x 50 m grid).

3.3.2. Minimal variance

Figure 3.4 shows the process of resampling the original 0.5 m binary image (that is, healthy and red stage pixels) by using odd sized windows (3 x 3, 5 x 5, 7 x 7 and so on). One of the effects of the resampling process is that the number of pixels decreases as the resolution becomes coarser (Woodcock and Strahler, 1987). Especially noticeable is that, grids cells with lower infestation levels (1% to 5%) result in a greater loss of pixels during the resampling process. Therefore, there are a limited number of times that the image data can be resampled and still have a reasonable number of pixels to estimate variance.

Figure 3.5 shows the variance plotted as a function of the resampled spatial resolutions. Two dominant trends are observed from the data. Firstly, for low to medium infestation levels (1% to 10%) the variance is relatively high at 1.5 m, decreases to reach a minimum at an intermediate resolution, and reaches a value of zero in the coarser resolutions (Figure 3.5a). As mentioned earlier, the zero variance values obtained at coarser resolutions are due to the resampling process.

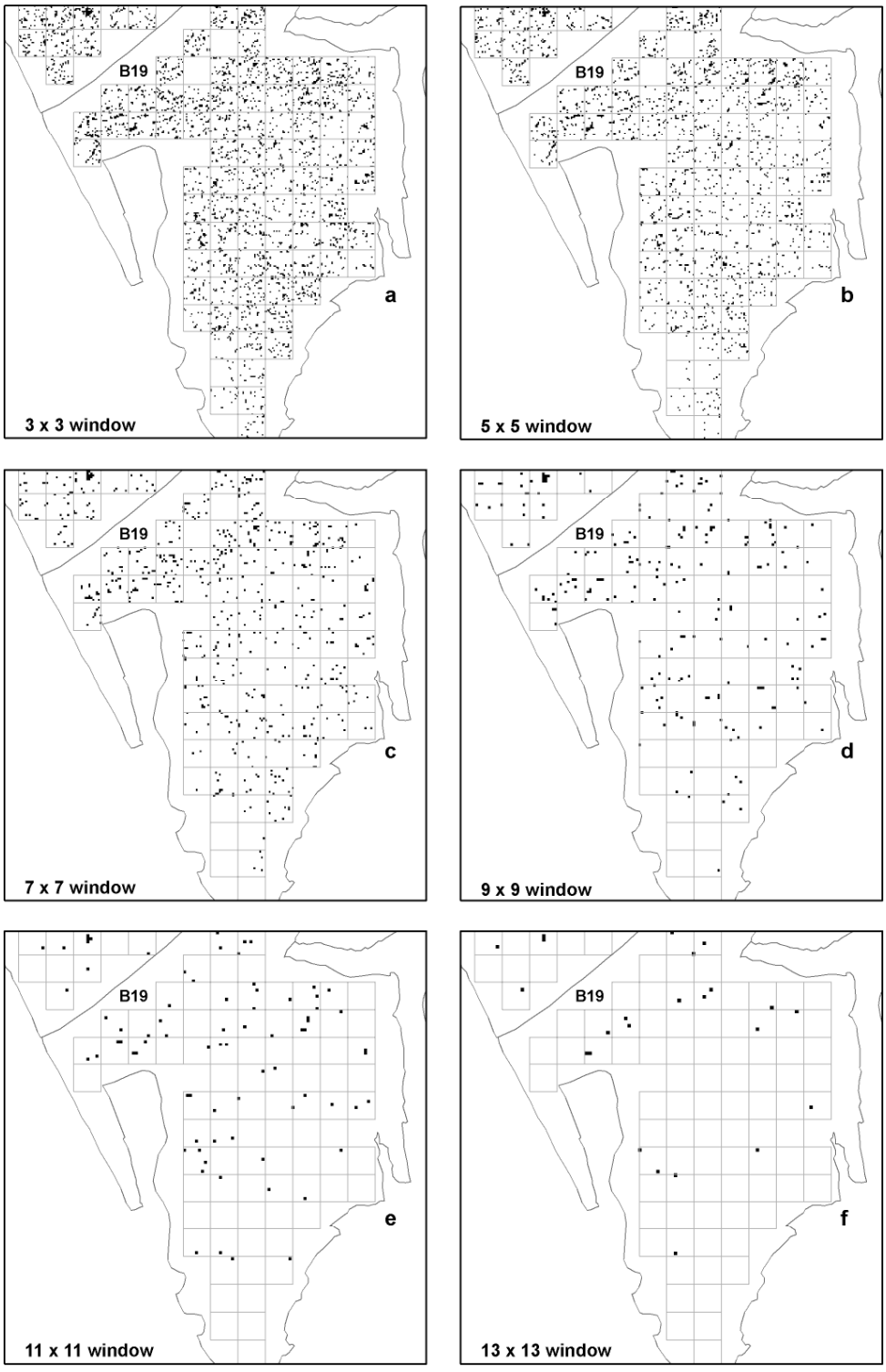


Figure 3.4: An example of the resampled binary images that were used to determine the minimal variance. The spatial pattern of *S. noctilio* infestations for compartment B19 are shown at 1.5 m (a), 2.5 m (b), 3.5 m (c), 4.5 m (d), 6.5 m (e), and 7.5 m (f) spatial resolutions.

Secondly, for the high infestation levels (11% to 15%) the variance is relatively high at 1.5 m, decreases towards the intermediate resolution, and stabilizes in the coarser resolutions (Figure 3.5b). In both cases, the minimum variance is observed when the variance of the pixels for each sub-sample is at the lowest level. This drop in variance (minimal variance) is then used as a measure of the optimal resolution that takes into account the inherent spatial properties of varying infestation levels.

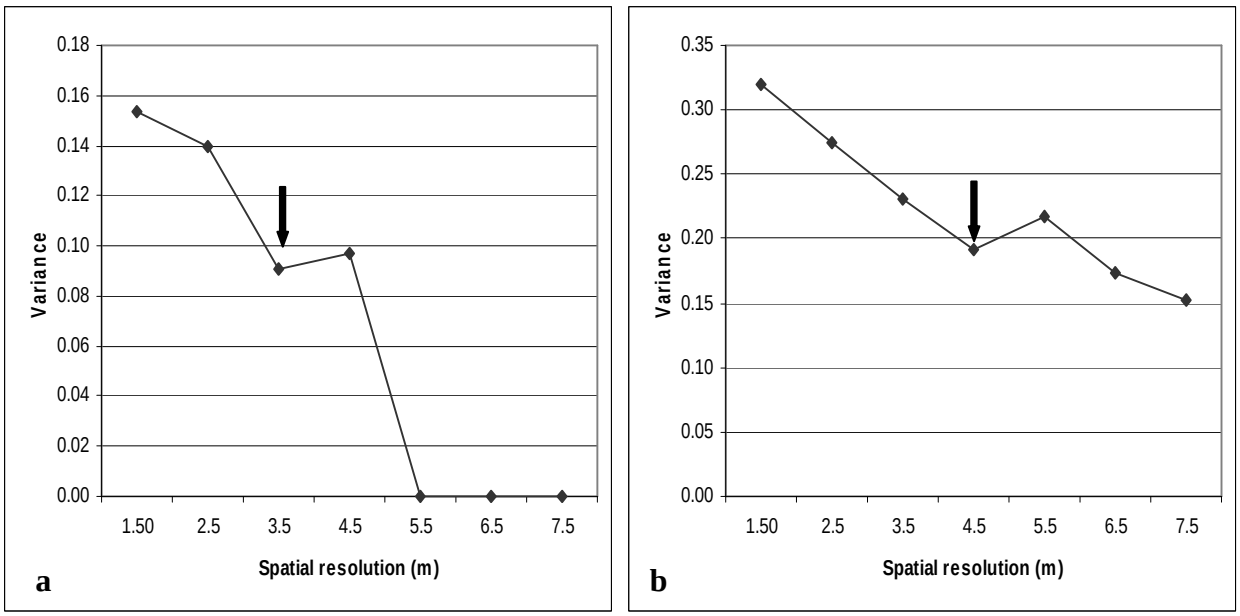


Figure 3.5: Graphs indicating trends of the calculated minimal variance. The spatial resolution at which each sub-sample reaches a minimum variance is shown with an arrow. Insert (a) shows trends prevalent in low to medium infestation levels (1% to 10%), while insert (b) shows the trend in variance for high infestation levels (11% to 15%).

The optimal spatial resolutions as determined by minimum variance was then averaged for each unique infestation level (n = 14), and the resulting spatial resolutions are shown in Table 3.5. Since our aim was to define an appropriate pixel size to capture the spatial variability of *S. noctilio* infestation and following the sampling theorem (McGrew and Monroe, 2000), results show that the appropriate resolutions for low to medium infestation levels (1% to 10%) range between 1.75 m and 1.93 m, while the appropriate resolution for higher infestation levels (11% to 15%) are between 1.99 m

and 2.31 m. Furthermore, correlation analysis was undertaken to examine the relationship between the appropriate spatial resolutions and *S. noctilio* infestation levels. The correlation coefficient ($r = 0.87$, $p < 0.001$) indicated that there is a strong correlation between the appropriate spatial resolutions and *S. noctilio* infestation levels. Results indicate that areas with high infestation levels can be detected using coarser resolution remotely sensed data, and areas with low infestation levels can be detected using finer remotely sensed data.

Table 3.5: Infestation levels and spatial resolution.

Infestation Rate (%)	**Resolution (Minimum Variance)**	**Infestation Level**	**Resolution (Sampling Theorem) (m)**
2	3.50	Low	1.75
3	3.50	Low	1.75
4	3.50	Low	1.75
5	3.64	Low	1.82
6	3.50	Medium	1.75
7	3.64	Medium	1.82
8	3.64	Medium	1.82
9	3.92	Medium	1.96
10	3.86	Medium	1.93
11	3.98	High	1.99
12	3.95	High	1.98
13	4.18	High	2.09
14	4.37	High	2.19
15	4.61	High	2.31

3.4. Discussion

For remote sensing technologies to be widely accepted by forest managers, and for these tools to be used on an operational basis, methods must allow for the efficient and cost-effective mapping of *S. noctilio* infestations. In this context, minimal variance calculated for localized sub-samples has proven to be a useful tool in determining an appropriate spatial resolution for the detection and mapping of *S. noctilio* infestation levels. Although the range of appropriate spatial resolutions is narrow (< 0.5 m), there would be a significant reduction in the cost of acquiring image data, since costs are primarily dependant on pixel size. The results obtained are consistent with the

hypothesis that each object mapped using remotely sensed data has a scale or a narrow range of scales associated with it, which provides its best representation (Marceau et al., 1994).

The results from this study provide the following guidelines: (i) for areas that have known *S. noctilio* infestations (medium to high infestation levels) pixels sizes between 1.75 m and 2.3 m would be sufficiently detailed to capture infestation rates, (ii) for newly colonized areas or areas that have low infestation levels, a pixel size of 1.75 m would be appropriate. Using pixel sizes larger than 2.3 m may not provide adequate information for high infestation levels (11% to 15%), while using pixel sizes smaller than the 1.75 m for detecting low to medium infestation levels (1% to 10%) could mean an unnecessarily large volume and cost of data.

However, determining the appropriate resolution for an investigation is a function of the type of environment, the kind of information desired, and the techniques used to extract the information (Chen et al., 2004; Garrigues et al., 2006). For example, studies on optimal resolution have shown that using different vegetation indices produce different results (Menges et al., 2001; Rahman et al., 2003). According to Menges et al. (2001), these differences are related to the suppression or enhancement of certain features on the image. Furthermore, the inclusion or exclusion of certain wavelengths might have implications for users wishing to select an appropriate resolution (Atkinson and Aplin, 2004).

To summarize, defining appropriate pixel size for an application is a complex task and depends mainly on the objectives of the study and the techniques used to retrieve the required information. Firstly the pixel size should be large enough to be consistent with the object (tree crowns) targeted and fine enough to capture the spatial variability of the data and minimize intra-pixel variability. The appropriate pixel sizes proposed in this study provide an indication of the upper (2.3 m) and lower (1.75 m) limit of the appropriate pixel sizes that are suitable for detection and mapping of *S. noctilio* infestations.

3.5. Conclusion

In this study, the effects of spatial resolution on detecting *S. noctilio* infestation levels were examined at a sub-sample level using classified NDVI images. This procedure established the appropriate spatial resolution guidelines necessary for the operational detection and mapping of *S. noctilio*. The appropriate pixel size should be chosen between the upper and lower limits proposed in this study but additional factors such as economic and technical constraints should be considered. Some of the major findings from the study are as follows:

1. Minimum variance calculated for localized sub-samples is a useful tool for identifying the appropriate spatial resolution needed for a particular investigation.
2. When using a spectral classifier (for example, NDVI) to detect infestation levels, pixel sizes larger than 2.3 m will not provide adequate information for high infestation levels (11% to 15 %), while using pixel sizes smaller than the 1.75 m for detecting low to medium infestation levels (1% to 10 %) could mean an unnecessarily large volume and cost of data.
3. Although the identified range of appropriate spatial resolutions is narrow (< 0.5 m), using the appropriate spatial resolutions as determined by this study would result in reduced costs of future image data acquisitions.

Acknowledgements

This research is the result of a joint collaboration between SAPPI, Mondi, NCT and the ICFR. We thank Mondi for allowing us access to the image data. The early contributions of Philip Croft and Mark Norris-Rogers are gratefully acknowledged. Funding for this research was provided by the National Research Foundation (NRF). We also thank the referees for valuable comments on the paper.

CHAPTER 4:

Discriminating *Sirex noctilio* infestations using high spectral resolution data

* This chapter is based on:

Ismail, R., Mutanga, O. and Ahmed, F., 2008. Discriminating *Sirex noctilio* attack in pine forest plantations in South Africa using high spectral resolution data. In: M. Kalacska and A. Sanchez-Azofeifa (Editors), Hyperspectral Remote Sensing of Tropical and Sub-Tropical Forests. Taylor and Francis: CRC Press, pp. 350.

Abstract

High spectral resolution data was used to identify diagnostic spectral features of *Pinus patula* needles showing varying degrees of attack by the wood boring pest, *Sirex noctilio*. The pest attacks all commercial pine species in South Africa, and the symptoms on infested trees can be represented on a severity scale, as the green, red, and grey stages of attack The objective of this study was to determine whether there is a significant difference between the mean reflectance (%) at each measured wavelength (from 400 nm to 1300 nm) for the green, red, and grey stages of attack. Next, for the wavelengths that were significantly different ($p < 0.001$) in this spectral region, the Jeffries–Matusita (J-M) distance was used to test whether some bands had more discriminating power than others. Using a field spectrometer, ninety reflectance measurements were obtained from several infested *P. patula* trees in KwaZulu-Natal, South Africa. Results indicate that spectral bands located in the visible portion (350 nm to 700 nm) and some spectral bands in the red edge (670 nm to 737 nm) of the electromagnetic spectrum could spectrally discriminate the different levels of *S. noctilio* attack. Although no single band is capable of total separability, results of the J-M analysis indicate that an acceptable separability of 99.22% (J-M value of 1403) for green, red, and grey stages was reached when using a four band combination comprising bands located at 500 nm, 521 nm, 685 nm, and 760 nm. The results encourage canopy scale detection and mapping of *S. noctilio* attack in pine forest plantations using airborne or spaceborne sensors.

Keywords: *Sirex noctilio*, high spectral resolution, Jeffries-Matusita distance

4.1. Introduction

The wood boring pest *Sirex noctilio* Fabricius (Hymenoptera:Siricidae) (Tribe and Cillie, 2004) is causing mortality along the heavily afforested eastern regions of South Africa, with recent reports indicating that mortality might be as high as 30% in some forestry compartments (Slippers, 2006). *S. noctilio* affects all commercial pine species in South Africa, with none of the species showing a high resistance to attack (Anon, 2004). Based on recent bioclimatic studies (Carnegie et al., 2006), the wasp is likely to spread further north, where the majority of South Africa's commercial pine forests are located (Figure 4.1).

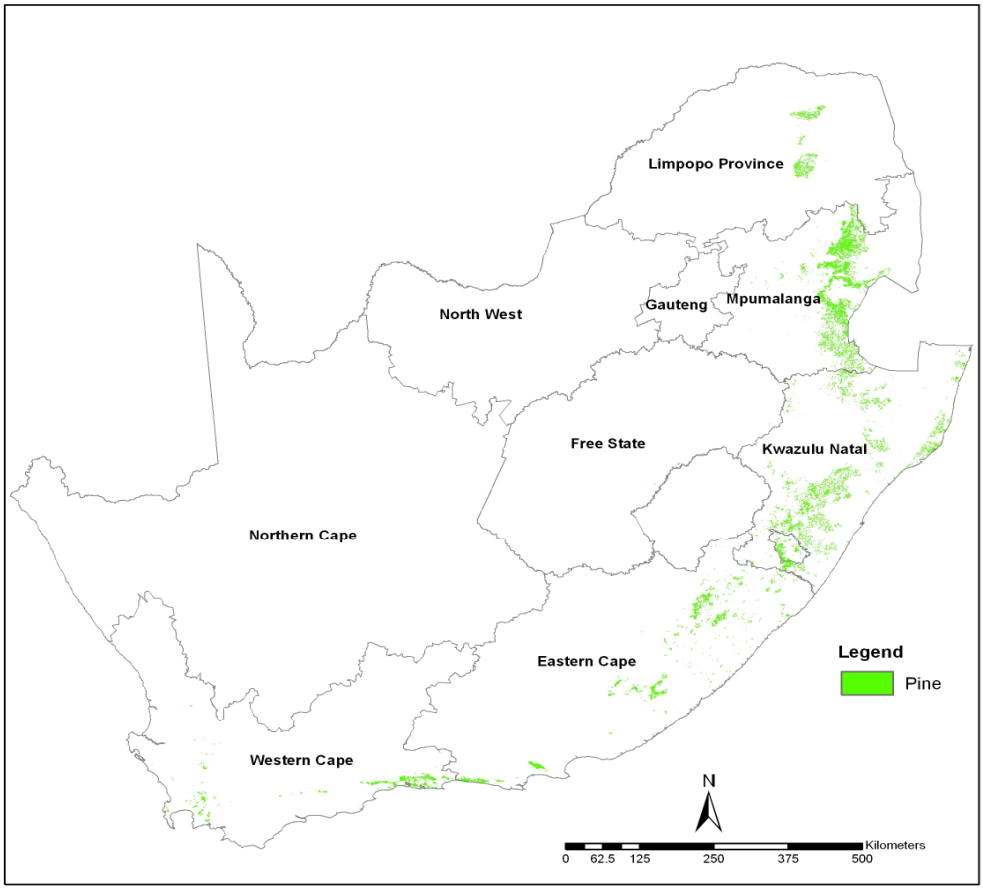


Figure 4.1: Map showing the distribution of pine (all commercial species) in South Africa. There are approximately 721,000 hectares of commercial pine plantations, with plantations concentrated in Limpopo Province (3.9%), Western Cape (12.1%), Eastern Cape (14.6%), KwaZulu-Natal (25.4%), and Mpumalanga (44%) (DWAF, 2005).

Management strategies by South African forest companies now focus on the combined use of remote sensing, silvicultural treatments, and biological control to reduce *S. noctilio* population numbers and to minimize the potential economic threat to the industry (Ismail et al., 2006). Remote sensing is a key component of the integrated management strategy and remains crucial for the detection and monitoring of the wasps and for the effective deployment of appropriate suppression activities (Ismail et al., 2006).

However, there are operational limitations that restrict the successful implementation of remote sensing by South African forestry companies. These limitations are primarily due to classification errors that arise because of the inability of broad band multispectral sensors to discriminate the different damage classes associated with *S. noctilio* attack. *S. noctilio* symptoms can be represented on a damage scale, as the green, red, and grey stages of attack (Figure 4.2).

Stages of attack		**Symptoms**
Green		Green healthy crown, presence of resin droplets, cambium stain, ovipositors found on the trunk, and there is no substantial needle loss
Red		Severe chlorosis, reddish brown canopy, and high needle loss.
Grey		Emergence holes, no canopy, and 100% needle loss.

Figure 4.2: Description of the damage symptoms due to *S. noctilio* attack.

Initial evidence of attack or the green stage of attack includes the appearance of resin droplets and the presence of ovipositors on the bark with a dark fungal stain appearing along the cambium. There is minimal needle loss and the canopy appears green and healthy. The red stage of attack occurs approximately three months later when the canopy of the attacked tree wilts and changes colour from green to yellow to reddish brown (Ciesla, 2003). Ultimately, during the grey stage of attack, the canopy defoliates and round exit holes appear on the bark. A new generation of adult wasps emerges resulting in a compartment of scattered pattern of dead or dying trees (Ciesla, 2003; Haugen et al., 1990; Haugen and Underdown, 1990). During the grey stage of attack the wood is totally desiccated and the timber is not usable and economic losses are incurred. It has been shown that the spectral separability among these different classes of damage remains elusive when using high spatial resolution broad band multispectral sensors (Ismail et al., 2006).

The limitations associated with the use of broad band multispectral channels are primarily due to classification errors that arise because of the inability of multispectral sensors to accurately discriminate among healthy, green, and red stage trees. For example, it has been shown that the spectral separability between healthy-green and green-red trees remains problematic when using high spatial resolution (0.5 m) multispectral sensors (Ismail et al., 2006). Given these limitations, there is strong optimism that with the new generation of hyperspectral sensors, significantly higher quality data would be available (Kumar et al., 2001). Hyperspectral data allows for the detection of detailed features which would have been otherwise masked by broad band sensors (Schmidt and Skidmore, 2001).

Internationally, the number of airborne and spaceborne hyperspectral sensors has increased (Lucas et al., 2004; Wulder, 1998). However, due to cost, availability, and accessibility of hyperspectral imagery in South Africa, only a few studies have investigated the potential of using high spectral resolution data. For example, Mutanga and Skidmore (2005) and Mutanga and Skidmore (2004a) successfully used HYMAP imagery to map grass quantity and quality in savanna ecosystems. Using the same imagery, Ferwerda (2005) mapped the total content of polyphenols and condensed tannins in *Colophospermum mopane* trees in an effort to understand herbivore distributions in the Kruger National Park. Studies focusing on the use of hyperspectral imagery in forestry

are limited, with the exception of Ahmed and Mthembu (2006) who calculated the leaf area index (LAI) of *Eucalyptus grandis* trees using spaceborne hyperspectral imagery (HYPERION) and narrow band vegetation indices.

However, there should be an increased interest in using high spectral resolution data for a wide variety of environmental applications due to the future availability and accessibility of hyperspectral sensors in the southern African region. It is envisaged that the South African satellite, Sumbandila (ZASat-002) is due for launch in 2007 from a Russian submarine, followed by ZASat-003 to be launched in 2008 (Scholes and Annamalai, 2006). ZASat-003 will carry a full multisensor microsatellite imager (MSMI) instrument as well as a hyperspectral sensor with a 14.9 km swath and 14.5 m ground sampling distance. This hyperspectral sensor will slice the spectrum between 400 nm and 2350 nm into 200 bands, each 10 nm wide (Scholes and Annamalai, 2006; van Aardt and Coppin, 2006). The question that then arises is, with the future availability and accessibility of high spectral resolution data in South Africa, is there potential to successfully discriminate between healthy and *S. noctilio* attacked pine trees?

Therefore, the preliminary aim of this study was to use high spectral resolution data to identify diagnostic spectral features of *Pinus patula* needles showing varying degrees of *S. noctilio* attack. The results obtained from this study could thus form the basis of future algorithms or spectral indices capable of discriminating *S. noctilio* attacked trees from healthy trees at either airborne or spaceborne platforms. More specifically, the objectives of this paper were: (i) to determine whether there is a significant difference in the mean reflectance (%) at each measured band (from 400 nm to 1300 nm) for the three stages of *S. noctilio* attack (green, red, and grey) and (ii) for the wavelengths that are spectrally different in this region, to test whether some bands have more discriminating power than others in the detection of *S. noctilio* induced stress. Reflectance measurements were taken from three groups of *P. patula* trees, classified according to the severity of attack by *S. noctilio* using a field spectrometer.

4.2. Materials and methods

4.2.1. Foliar samples

During April 2006, needle samples from healthy, green, and red stage *P. patula* trees were collected from a known *S. noctilio* attacked compartment at the Pinewoods plantation (centroid $30°4'13.83"E$ and $29°38'36.06"S$) KwaZulu-Natal, South Africa. All pine species are susceptible to attack (Anon, 2004). However, only *P. patula* trees have been attacked in KwaZulu-Natal. Before any spectral measurements were taken, the trees were carefully examined with the assistance of experienced foresters and classified into mutually exclusive classes (that is, healthy, green, or red). Green stage trees were checked for the presence of ovipositors, resin droplets, and cambium staining. Additionally, trees that were classified as red stage were destructively sampled to evaluate the presence or absence of *S. noctilio* larvae. The grey stage of attack was excluded from this study since grey stage trees are completely defoliated and therefore no needle samples were collected. Five trees from each stage, including the healthy trees were then sampled. Pine needles were collected from three branches (upper, middle, and lower crowns) with two needle samples from the same branch. Sampled needles used for spectral measurements were from trees of the same age, and no other damaging agents were observed on the pine needles. The samples (n = 90) were immediately sealed in zip lock plastic bags and placed in coolers on ice and taken within four hours of collection to the laboratory at the University of KwaZulu-Natal for spectral measurements.

4.2.2. Spectral data acquisition

Spectral measurements of pine needles were acquired using analytical spectral devices (ASD) FieldSpec Pro FR spectroradiometer, which senses in the spectral range from 350 nm to 2500 nm at a spectral bandwidth of 1.4 nm to 2.0 nm and a spectral resolution of 3 nm to 10 nm (Analytical Spectral Devices, 2002). This spectral range incorporates the visible (400 nm to 700 nm), NIR (700 nm to 1200 nm) and the short wave infrared (1200 nm to 2500 nm). The ASD spectroradiometer, equipped with a

field of view of 25°, was mounted on a tripod and positioned 0.5 m above the needle sample at the nadir position. A 150 Watt halogen bulb was used to illuminate the pine needle leaves. Reflectance spectra were obtained by calibrating the radiance of the target pine needles with the radiance of a standard (white reference panel, spectralon) of known spectral characteristics. Needle samples for the different classes (healthy, green, and red) were randomly stacked on a target platform. The needles were shuffled and reflectance measurements were taken. This process was carried out for the healthy (n =30), green (n = 30), and red (n = 30) classes. The entire experiment was conducted under controlled laboratory conditions (that is, dark room, 25 °C) in order to avoid ambient light sources unrelated to the true spectral signal of the needles (Vaiphasa et al., 2005). In total, 900 reflectance values were acquired for each class but these were later averaged (n = 90) in order to reduce within class variability.

4.2.3. Data analysis

The hypothesis that the mean reflectance of the healthy, green, and red stages was significantly different at each measured wavelength in the 350 nm to 1300 nm region was tested using one-way analysis of variance (ANOVA) with a Tukey's HSD post hoc test. The ANOVA with the post hoc test was calculated at each measured wavelength for the respective class pair (that is, healthy-green, green-red, and healthy-red, hereafter referred to as H-G, G-R, and H-R) and then summarized using a histogram. The histogram was calculated by counting the number of significant bands at each wavelength for all class pairs. The histogram then indicates the frequency of significant wavelengths and which wavelengths are relatively more important for discriminating all classes (that is, healthy, green, and red, hereafter referred to as H-G-R). Additionally, to identify wavelengths most responsive to *S. noctilio* attack, sensitivity analysis were undertaken following the procedure described by (Carter, 1994; Cibula and Carter, 1992).

$$R_{\lambda} = (R_{\lambda i} - R_{\overline{\lambda h}}) / R_{\overline{\lambda h}} \quad (1)$$

Where $(R_{\lambda i} - R_{\overline{\lambda h}})$ is the spectral reflectance difference curves calculated by subtracting the mean reflectance of the healthy pine needles at each wavelength from the mean reflectance of the green or red stage needles. When the reflectance of healthy needles are subtracted from the reflectance of the green or red stage needles, the resulting difference curves indicate the wavelengths in which reflectance changed greatly with stress (Carter and Knapp, 2001). When this difference curve is divided by the mean reflectance of the healthy needles, the results yield the relative change in reflectance or reflectance sensitivity (Stone et al., 2003). Sensitivity analysis therefore shows the wavelengths at which reflectance was strongly affected by *S. noctilio* attack. The coefficient of variation (CoV) was also calculated at each wavelength for reflectance values across all three classes. The CoV is the ratio of the standard deviation to the mean reflectance (Stone et al., 2003). Higher CoV values result in greater variability and hence discriminatory power of the respective wavelength.

Finally, we tested the hypothesis that some bands have more discriminatory power than others by calculating the Jeffries-Matusita (J-M) distance (Richards and Jia, 1999). The J-M distance calculation delivers a value between 0 and √2 (≈ 1.414), with higher values representing better separability of class pairs (Richards and Jia, 1999). Therefore the band or band combinations producing the highest J-M distance averaged over each class pair can be considered to be the best for discriminating *S. noctilio* attack. Although previous studies used J-M distance thresholds of ≥ 95 % (Vaiphasa et al., 2005) to indicate separability, the present study uses higher separability values ≥ 99 % largely due to the potential of upscaling from leaf to tree canopy and the variability associated with canopy reflectance when compared to leaf samples (Stone et al., 2001).

The Jeffries-Matusita distance is formally stated as

$$\alpha = \frac{1}{8}\left(\mu_i - \mu_j\right)^T \left(\frac{C_i + C_j}{2}\right)^{-1} \left(\mu_i - \mu_j\right) + \frac{1}{2}\ln\left(\frac{\left(\left|C_i + C_j / 2\right|\right)}{\sqrt{\left|C_i\right| * \left|C_j\right|}}\right) \qquad (2)$$

For normally distributed classes this distance becomes the Bhattacharyya (BH) distance (Richards and Jia, 1999) and is stated as

$$JM_{ij} = \sqrt{2\left(1 - e^{\alpha}\right)} \quad (3)$$

Where:

i and j = the two classes being compared

C_i = the covariance matrix of signature i

μ_i = the mean vector of signature i

ln = the natural logarithm function

$|C_i|$ = the determinant of C_i (matrix algebra)

4.3. Results

The results of the ANOVA for individual class combinations (H-G, G-R, and H-R) are shown in Figure 4.3 (a, b, and c). The shaded areas indicate the reflectance wavelengths where the class pairs show a significant statistical difference in reflectance ($p < 0.001$). Table 4.1 shows the frequency of statically significant wavelengths for spectral regions as defined by Gong et al. (2002). These spectral regions were selected to simplify the spectral interpretation of individual class pairs. There are no significant wavelengths located in the blue range. However, the blue edge is very responsive to the G-R and H-R class pairs. Additionally, the H-R combination has more significant wavelengths (n = 60) located in the green range than any other class pairs (H-G, n = 40 and G-R, n = 25), whereas the H-G combination has more significant wavelengths (n = 33) located in the yellow edge when compared to the G-R (n = 21) and H-R (n = 19) class pairs. For all class pairs, there are more statistically significant wavelengths located in the red region (600 nm to 700 nm) of the electromagnetic spectrum (H-G, n = 100; H-R, n = 100 and G-R, n = 93).

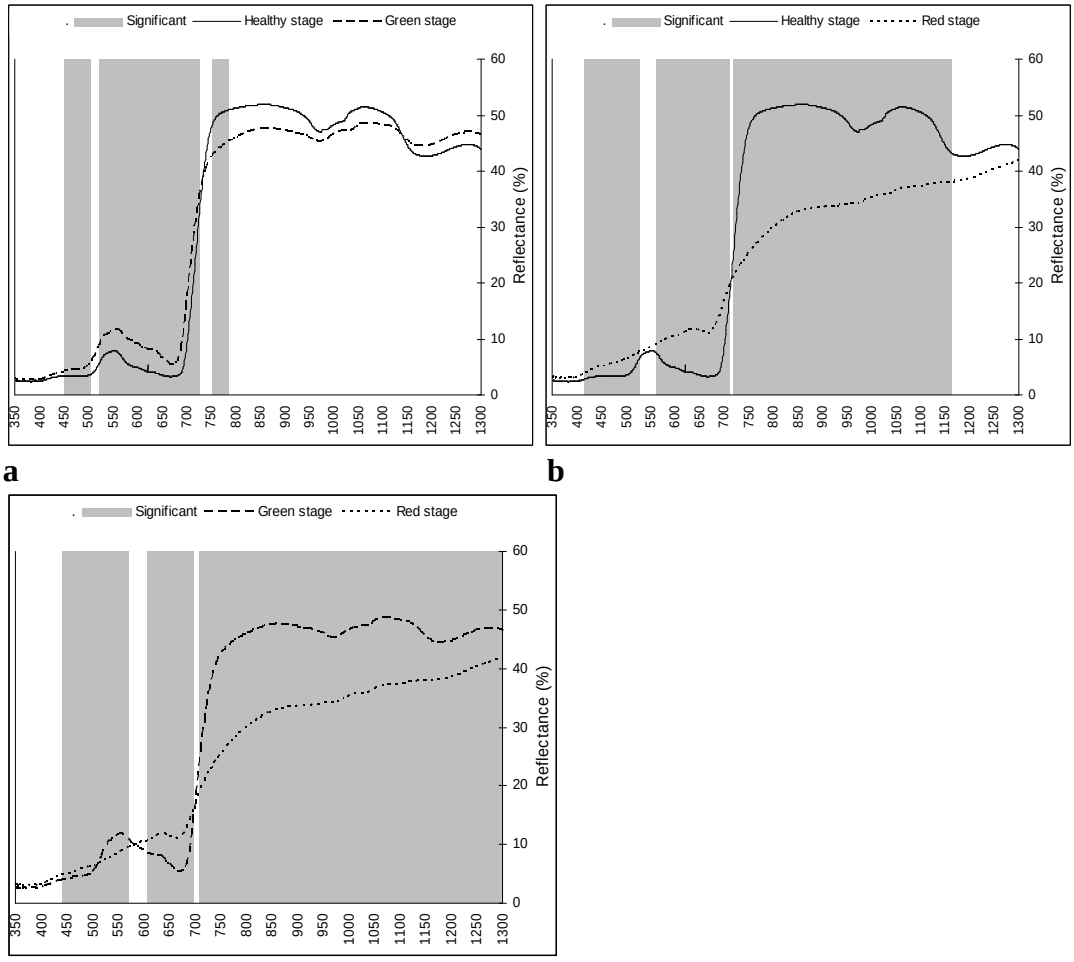


Figure 4.3: ANOVA results for individual class pairs (a) healthy-green, (b) healthy-red, and (c) green-red. The grey shades indicate regions of electromagnetic spectrum where there were significant differences.

However, there are far more significant wavelengths located in the near infrared (NIR) for the G-R (n = 594) and H-R (n = 450) class pairs when compared with the H-G (n = 57) class pair. An examinations of all wavelengths (from 350 nm to 1300 nm) used in this study reveal that there are proportionally more significant wavelengths in this region for the H-R (n = 814) and G-R (n = 699) class pairs. For the H-R combination, 62.68% of these significant wavelengths are located in the visible region, and for the G-R combination, 73.50% of the significant wavelengths are located in the visible region.

Table 4.1: Frequency table of statically significant bands for the spectral regions defined by Gong et al. (2002).The table shows the results for individual class pairs.

Wavelength region (nm)	Description	Number of bands	H-G	%	G-R	%	H-R	%
350-400	Blue range	51	0	0.00	0	0.00	0	0.00
490-530	Blue edge	41	21	51.22	38	92.68	41	100.00
501-560	Green range	60	40	66.67	25	41.67	60	100.00
550-582	Yellow edge	33	33	100.00	21	63.64	19	57.58
640-680	Red well	41	41	100.00	41	100.00	41	100.00
670-737	Red edge	68	54	79.41	58	85.29	58	85.29
700-900	NIR	201	58	28.86	191	95.02	194	96.52
350-700	Visible	351	231	65.81	248	70.66	220	62.68
350-1300	All wavelengths	951	288	30.28	699	73.50	814	85.59

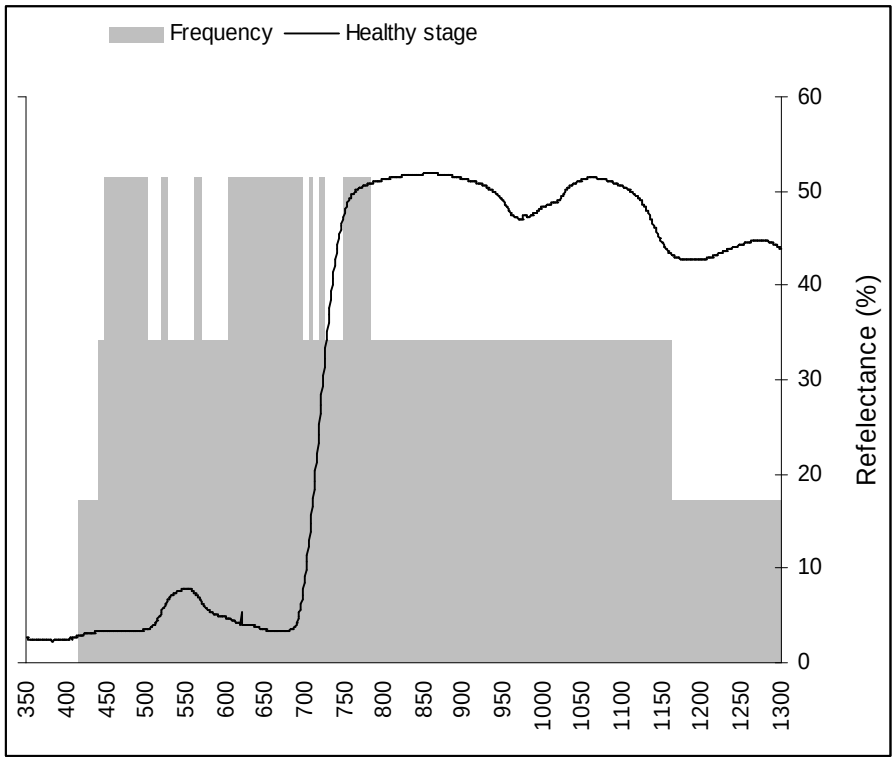


Figure 4.4: Frequency of statistically significant differences for all classes (healthy, green, and red). The maximum grey shadings indicate the wavelengths that could discriminate between all class combinations of damaged trees. The spectral signature for the healthy needle is included for comparative purposes.

The histogram in Figure 4.4 summarizes the results of the frequency table (Table 4.1). The histogram then indicates which wavelengths can potentially discriminate all three classes (H-G-R). These responsive wavelengths are defined by regions with the maximum grey shading shown in Figure 4.4. Wavelengths most responsive are located in the regions from 450 nm to 500 nm (n = 51), 521 nm to 525 nm (n = 5), 562 nm to 568 nm (n = 7), and 604 nm to 696 nm (n = 93) and 750 nm to 783 nm (n = 34). The majority of the significant bands (82.10%) are located in the visible part (300 nm to 700 nm) of the electromagnetic spectrum, and the remainder of the significant wavelengths are located in the shoulder region of the NIR.

4.3.1. Sensitivity analysis

Sensitivity analysis was used to reduce the number of significant bands prior to the J-M calculation. Therefore, sensitivity analysis was restricted to regions that were identified as being significant for all three classes. These significant areas are shaded in grey in Figure 4.5. It was not possible to calculate the J-M distance by using all significant bands (n = 190) because of the singularity problem of matrix inversion (Vaiphasa et al., 2005). Additionally, spectral information needed to discriminate among healthy, green, and red stages could be extracted from only a few optimal bands since many bands are highly correlated and often redundant (Lucas et al., 2004).

Figure 4.5 shows the results of the sensitivity analysis. The wavelengths most responsive to discriminating *S. noctilio* attack are shown using black arrows which indicate wavelengths that were selected for further analysis. Sensitivity maxima located at 500 nm and 695 nm and a wavelength located at 720 nm were identified from the sensitivity curves (Figure 4.5b). Wavelengths located at 521 nm, 565 nm and 707 nm were identified from the spectral difference curves (Figure 4.5a). Minima located at 685 nm and 760 nm were also identified from these difference curves. Table 4.2 shows the results of the CoV calculated for the wavelengths. Noticeable are the high CoV values for the 500 nm, 685 nm, 690 nm, and 695 nm wavelengths located in the red edge region of the electromagnetic spectrum.

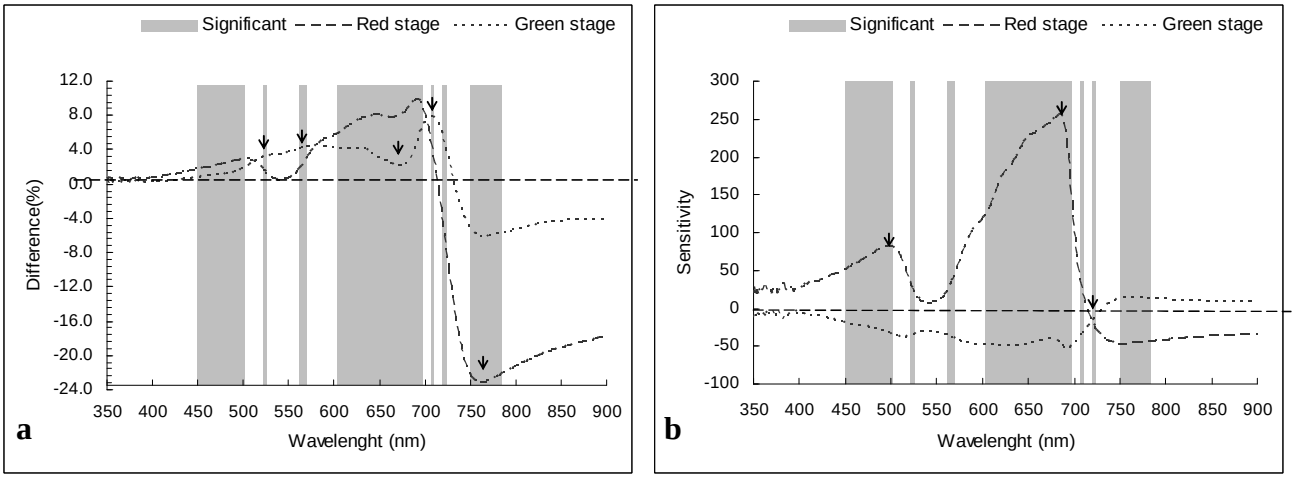


Figure 4.5: Sensitivity analysis. Reflectance difference curves for the red and green stages of attack are shown in insert (a). The curves were calculated by subtracting the mean reflectance of the red as well as the green stage needles from the reflectance of healthy needles. Insert (b) shows the result of the sensitivity analysis. The curves were calculated by dividing the reflectance difference curves of the red and green stage needles by the reflectance of healthy needles.

Table 4.2: Coefficient of variation results for the nine selected bands

Band	nm	Description	CoV
1	500	Blue edge	30.10
2	521	Blue edge	24.21
3	565	Yellow edge	25.12
4	685	Red edge	56.26
5	690	Red edge	51.79
6	695	Red edge	42.86
7	707	Red edge	24.59
8	720	Red edge	19.38
9	760	NIR	25.01

4.3.2. Distance analysis

Based on the results from the sensitivity and CoV analysis, nine spectral bands (Table 4.2) were selected for further analysis. Jeffries–Matusita distances (Richards and Jia, 1999) were then calculated to determine the best combinations of bands for separating the classes (that is, healthy, green, and red) from each other. J-M distances range from 0 to 1.414, with higher values representing better separability of all class pairs (H-G, G-R,

and H-R). Therefore, the bands or band combinations producing the highest J-M distance averaged over all class pairs can be considered to be optimal for discrimination.

Table 4.3 shows the results of the J-M distance averaged for all combined class pairs. The best separability was obtained when using a single band located at 685 nm which produces an unacceptable J-M value of 1.186. However, using more bands considerably improves the separability of the classes. The best average J-M values (1.413) are reached when using a seven band combination, with individual bands located at 500 nm, 521 nm, 565 nm, 685 nm, 690 nm, 695 nm, and 760 nm. However, using a four band combination, comprising of bands located at 500 nm, 521 nm, 685 nm and 760 nm produces an acceptable J-M value of 1.404 or 99.29% separability. Additionally, the frequencies of these bands in the best combinations as shown in Table 4.3 are relatively high when compared to other band combinations.

Table 4.3: Results of the average Jeffries–Matusita distance analysis for all the three classes (healthy, green, and red). The symbol (*) indicates the optimal bands that were selected in each band combination.

Best combination	**Wavelength(nm)**									**J-M value**	
	500	521	565	685	690	695	707	720	760		%
Single band				*						1.186	83.88
Two band							*	*		1.354	95.76
Three band	*	*		*						1.392	98.44
Four band	*	*		*					*	1.404	99.29
Five band	*	*		*	*				*	1.410	99.72
Six band	*	*	*	*	*				*	1.412	99.86
Seven band	*	*	*	*	*	*			*	1.413	99.93
Eight band	*	*	*	*	*		*	*	*	1.413	99.93
Nine band	*	*	*	*	*	*	*	*	*	1.413	99.93

Table 4.4 shows the J-M distances calculated individually for H-G, H-R, and the G-R class pairs. For the H-R pair, two band combinations located at 707 nm and 720 nm, or at 707 nm and 760 nm produce a saturated J-M value of 1.414 and hence total separability (100%). However, a single band located at either 685 nm or 690 nm is capable of 99.29% separability.

Table 4.4: J-M values for individual class pairs using all possible band combinations.

Best combination	H-G	%	H-R	%	G-R	%
Single band	1.240	87.69	1.404	99.29	1.224	86.56
Two band	1.339	94.70	1.414	100.00	1.368	96.75
Three band	1.370	96.89	1.414	100.00	1.395	98.66
Four band	1.400	99.01	1.414	100.00	1.403	99.22
Five band	1.408	99.58	1.414	100.00	1.408	99.58
Six band	1.410	99.72	1.414	100.00	1.411	99.79
Seven band	1.411	99.79	1.414	100.00	1.413	99.93
Eight band	1.412	99.86	1.414	100.00	1.413	99.93
Nine band	1.413	99.93	1.414	100.00	1.414	100.00

The total separability for the G-R pair is reached when using all nine bands. However, an acceptable separability of 99.22% (J-M value of 1403) is reached when using a four band combination comprising bands located at 500 nm, 521 nm, 685 nm, with the fourth band located at 760 nm. The H-G pair does not reach total separability, but the best separability (1.413) is obtained when using all nine bands. However, using a four band combination produces an acceptable separability of 99.01 %, with bands located at 500 nm, 521 nm, 685 nm, and 760 nm.

4.5. Discussion

Numerous researchers have examined the ability of hyperspectral data to discriminate the damage caused by forest pests and pathogens (Ahern, 1988; Carter et al., 1996; Coops et al., 2002; Coops et al., 2003; Pontius et al., 2005a; Ruth et al., 1991; Stone et al., 2001; Stone et al., 2003). However, the present study is the first to report on *S. noctilio* attacked *P. patula* foliage, with results indicating that hyperspectral reflectance data measured in a laboratory environment can successfully discriminate varying levels of needle damage. Results show that there is a significant difference between the mean reflectance for all three classes with a large number of significant wavelengths located in the visible region of the electromagnetic spectrum. More specifically, results show that at least 76% of significant wavelengths are located from 450 nm to 500 nm (n = 51) and from 604 nm to 696 nm (n = 93). The results obtained in this study corroborate previous studies at leaf level that have shown that reflectance

measured at visible wavelengths (400 nm to 700 nm) are generally the most consistent response to stress (Carter, 1994; Carter et al., 1996; Carter and Knapp, 2001).

Although chlorophyll content was not directly measured in this study, it is plausible that the prevalence of significant wavelengths in the visible region, especially in the red region, is due to the effects of chlorophyll. Wavelengths within the 690 nm to 700 nm regions are particularly sensitive to decreases in the content of leaf chlorophyll and represent the blue shift of the red edge that frequently accompanies stress (Carter, 1994; Cibula and Carter, 1992). There is evidence that the combined effects of a phytotoxic mucus and the wood decaying fungus *Amylostereum areolatum* result in the breakdown of chlorophyll during initial *S. noctilio* attack (Neumann and Minko, 1981; Tribe and Cillie, 2004). It is therefore assumed that the chlorophyll concentrations will vary with the different stages of *S. noctilio* attack. This hypothesis is still speculative and would have to be further tested using foliar biochemistry analysis. While pigments may control the spectral responses of leaves in the visible wavelengths, it is the cellular structure and water content of leaves that are the main determinants in the near and mid infrared regions of the electromagnetic spectrum (Coops et al., 2002). NIR spectral responses are prevalent during the later stages of attack (that is, red stage). The red stage is characterized by the collapse of vascular tissue due to the growth of the fungus, *A. areolatum* (Neumann and Minko, 1981; Slippers et al., 2003). More specifically results indicate that there are more significant bands located in the NIR region for the G-R (n = 594) and H-R (n = 450) class pair when compared with the H-G (n = 57) combination.

Prior to the J-M calculation, sensitivity analysis was used to reduce the number of significant bands, and nine spectral bands were selected for further analysis. Two bands from the blue edge, one band from the yellow edge, five bands from the red edge, and one band from the NIR were selected. J-M distances indicate only a few optimal bands are needed to produce acceptable separability results. Although no single band is capable of total separability, spectral separability of all the classes (healthy, green, and red) is possible when using a four band combination. Bands located at 500 nm, 521 nm, 685 nm and 760 nm provided the best average separability (99.01%) for all stages of *S. noctilio* attack. Similar band combinations (500 nm, 521 nm, 685 nm, and 760 nm) also produce the best separability results for individual class pairs. The results are

consistent with previous studies that state that only three or four well placed bands were needed for a good classification (Leckie et al., 2005). The results therefore encourage further investigation into the capability of using airborne and satellite hyperspectral sensors for mapping *S. noctilio* attack.

4.6. Conclusion

To summarize, a better understanding has been gained about specific regions of the electromagnetic spectrum that offer the maximum information content for discriminating *S. noctilio* attack. It has been shown that there is a significant difference between the mean reflectance for all three classes with a large number of significant wavelengths located in the visible region of the electromagnetic spectrum. Therefore, an important prerequisite (that is, band selection) for the potential upscaling of results to either an airborne or spaceborne platform was established. Although no single band can discriminate among all the stages of *S. noctilio* attack, bands located at 500 nm, 521 nm, 685 nm, and 760 nm show the greatest potential for discrimination. Overall, the results provide evidence that encourages canopy scale investigation into the capability of high spectral resolution data for discriminating *Sirex noctilio* attack in pine forest plantations in South Africa

Acknowledgements

We thank Sappi for allowing us access to the Pinewoods plantations. The contributions of Marcel Verleur in identifying *Sirex noctilio* infestations are gratefully acknowledged. We thank Wayne Jones for assisting with the sampling of pine needles. Eric Economon from the Agricultural Research Centre (ARC) of South Africa provided assistance with the ASD spectroradiometer. Funding for this research was provided by the National Research Foundation (NRF) South Africa.

CHAPTER 5:

Discriminating the early stages of *Sirex noctilio* infestation using random forest and shortwave infrared (SWIR) wavelengths

* This chapter is based on:

Ismail, R. and Mutanga, O., *in review b*. Discriminating the early stages of *Sirex noctilio* infestation using random forest and shortwave infrared (SWIR) wavelengths. International Journal of Remote Sensing.

Abstract

In this study we evaluated whether the random forest algorithm can accurately discriminate between healthy trees and the early stages of *Sirex noctilio* infestation using reflectance measurements in the shortwave infrared (SWIR) wavelengths. Results show that the random forest algorithm produces slightly better classification results than a competing boosting tree algorithm for all three variable selection methods used in the study. Additionally, results indicate that wavelength noise is less harmful than class noise on the performance of the random forest algorithm. The ability of wavelengths located at 1990 nm, 2009 nm, 2028 nm, 2047 nm, and 2065 nm to discriminate between healthy and green stage spectra could be explained by the rapid physiological changes that occur as a result of toxic mucus and a fungus that is injected into the tree during the early stages of *S. noctilio* infestations. Overall the results are encouraging and show that there is a link between the selected SWIR wavelengths and existing physiological knowledge thereby improving the chances of detecting the early stages of *S. noctilio* infestation at a canopy or landscape level.

Keywords: Random forest, *Sirex noctilio*, variable selection, shortwave infrared (SWIR), noise

5.1. Introduction

Sirex noctilio is currently the most important pest of conifers in South Africa causing an estimated 45 million US dollars of damage in the summer rainfall areas of the country (Hurley et al., 2008). Multispectral remotely sensed data can detect the later, more visible stages of *S. noctilio* infestations when the canopy of the attacked tree changes colour from green to yellow to reddish brown (Ismail et al., 2007; Ismail et al., 2008b). However, a primary limitation remains on the effective discrimination between healthy trees and the early (or green) stage of *S. noctilio* infestations (Ismail et al., 2008a). High spectral resolution data (hyperspectral) has the ability to discriminate the early stages of insect infestations (Lawrence and Labus, 2003; Pontius et al., 2005b; Pontius et al., 2008) because the wavelengths are narrow (10 nm or less), and small spectral differences can be distinguished (Schmidt and Skidmore, 2001). Initial efforts at discriminating the green stage from healthy trees showed that wavelengths located at 500 nm, 521 nm, 685 nm, and 760 nm have the greatest potential (Ismail et al., 2008a). The laboratory based study by Ismail et al. (2008a) concentrated on the visible and near infrared regions but excluded the short wave infrared (SWIR), a domain which researchers have shown to be a good and consistent indicator of conifer mortality (Collins and Woodcock, 1996; Jin and Sader, 2005; Skakun et al., 2003). The present study intends to expand on the work by Ismail et al. (2008a) by determining if there are wavelengths in the SWIR region that will allow for the accurate discrimination between healthy trees and the green stage of infestation.

In anticipation of the future availability of hyperspectral data in South Africa (Scholes and Annamalai, 2006; van Aardt and Coppin, 2006), there is a keen interest amongst researchers in remote sensing to develop robust methods and techniques that will allow for the accurate discrimination of the green stage of *S. noctilio* infestation. Additionally, these methods need to be automated to some level with limited human interaction to allow for critical evaluation (Soh, 1999). However, hyperspectral data tends to be more difficult to process than the commonly used multispectral data due to the geometrical and statistical properties associated with high dimensional data (Langrebe, 2002). From a statistical perspective, the challenge is to identify the relevant wavelengths from a large set of candidate wavelengths (p) and a small number of

samples (*n*). The "small *n* large *p* problem" introduces multi-collinearity into the input data matrix which subsequently leads to instability in the classification process (Kavzoglu and Mather, 2002). Many variable selection approaches have been proposed to reduce the "curse of dimensionality" (Bajcsy and Groves, 2004; Bruzzone and Serpico, 2000; Kavzoglu and Mather, 2002; Vaiphasa et al., 2005; Vaiphasa et al., 2007). These approaches can be basically divided into two categories based on whether or not they use the classification algorithm as part of the evaluation process (Guyon and Elisseeff, 2003). If the variable selection is independent of the classification algorithm then the approach is defined as being a filter approach, and if the variable selection is dependent on the classification algorithm then the method is defined as a wrapper approach (Kohavi and John, 1997). The filter approach has been commonly used to reduce the number of wavelengths in hyperspectral applications (Ismail et al., 2008a; Schmidt and Skidmore, 2001; Vaiphasa et al., 2005). However, the wrapper approach is fast gaining popularity amongst some researchers in remote sensing (Chan and Paelinckx, 2008). A potential problem when using the faster filter approach is that the ranking of the wavelengths is carried out between pairs of wavelengths and without any direct relation to the classification algorithm. According to Granitto et al., (2006), in order to obtain unbiased estimates of error, especially in applications where $n < p$, the selection of variables should be included in the classification process and not treated as a separate pre-processing step. Therefore, an algorithm such as random forest (Breiman, 2001) that provides an additional direct measure of variable importance should be well suited for the classification of hyperspectral data.

While popular methods such as support vector machines and neural networks are useful for the classification of hyperspectral data (Mutanga and Skidmore, 2004a; Pal and Mather, 2004), these machine learning algorithms do not produce any insight regarding the wavelengths that would best contribute to the final classification (Archer and Kimes, 2008). Nevertheless,, the random forest algorithm has been successfully used for variable selection and for classification purposes in non-remote sensing domains (Diaz-Uriarte and Alvarez de Andres, 2006; Granitto et al., 2006; Svetnik et al., 2003; Svetnik et al., 2004). However, few applications in the remote sensing domain have evaluated the random forest algorithm for the combined purpose of classification and variable selection using hyperspectral data. Additionally, researchers

have used the random forest algorithm in applications that classify phenomena or objects that have distinct spectral characteristics. An exception was the study by Chan and Paelinckx (2008) which used the random forest algorithm and hyperspectral wavelengths to detect subtle changes in an ecosystem. However, in the study the number of samples far exceeded the total number of variables. In this study the random forest algorithm is evaluated for variable selection and classification in a hyperspectral application (i) where the number of samples are less than the number of variables ($n < p$) and (ii) where classes have similar spectral characteristics.

To summarize, this study evaluates whether the random forest algorithm can accurately discriminate between healthy trees and the early stages of *S. noctilio* infestation using SWIR wavelengths. More specifically, the potential role of three variable selection methods to produce a subset of wavelengths with the lowest misclassification error is examined. Furthermore, the study investigates if the random forest algorithm can recover the signal in the hyperspectral data when the class labels or the reflectance values of wavelengths are randomly altered. As no single machine learning algorithm has been demonstrated to be superior for all applications (Kohavi et al., 1996), it was necessary to test an additional competing ensemble based algorithm (that is, boosting trees) for the accurate discrimination of green stage and healthy spectra. Given that the number of samples were limited in this study, it was not practical to reduce the original observations for testing purposes. Therefore the .632+ bootstrap error (Efron and Thibshirani, 1997) was used to assess the classification accuracy of both algorithms.

5.2. Materials and methods

5.2.1. Spectral data acquisition and processing

During April 2006, needle samples from healthy and green stage *Pinus patula* trees were collected from a *S. noctilio* attacked compartment located at the Sappi Pinewoods plantation (centroid $30°4'13.83" E$ and $29°38'36.06" S$) in KwaZulu-Natal, South Africa (Ismail et al., 2008a). Before any samples or spectral measurements were acquired, the

trees were carefully examined with the assistance of experienced foresters and classified into mutually exclusive classes (that is, healthy or green stage trees). The green stage trees are characterized by the appearance of resin droplets along the trunk of the tree, the presence of ovipositors on the bark, and a dark fungal stain appearing along the cambium. There is minimal needle loss, and the canopy appears green, healthy, and visibly indistinguishable from a healthy tree (Neumann and Minko, 1981; Tribe and Cillie, 2004). Tree climbers obtained samples from five green stage trees and four healthy trees. The samples for each class were obtained from three branches (upper, middle and lower crowns) with two needle samples from the same branch. The pine trees were of the same age and no other damaging agents were observed (Ismail et al., 2008a).

Spectral measurements of the needle samples (n = 54) were acquired *in situ* on a clear sunny day between 10:00 am and 2:00 pm using the analytical spectral device (ASD) Field Spec Pro FR spectroradiometer. The ASD senses in the 350 nm to 2500 nm spectral range and has a bandwidth of between1.4 nm and 2.0 nm with a spectral resolution of 3 nm to 10 nm (Analytical Spectral Devices, 2002). The instrument (equipped with a field of view of 25°) was mounted on a tripod and positioned 0.5 m above each sample at the nadir position. Following ASD measurement protocols, reflectance spectra were obtained by calibrating the radiance of the target samples with the radiance of a standard (white reference panel, spectralon) of known spectral characteristics. Samples from the green stage (n = 30) and healthy trees (n = 24) were randomly stacked on a target platform. To minimize error, ten spectral reflectance measurements were averaged for each sample and individual samples were rotated 30° between scans (Pontius et al., 2005a). Since current hyperspectral sensors do not reach such a fine spectral resolution as the ASD does, the ASD spectra was subsequently resampled to HYMAP spectra using ENVI 4.2 (http://www.ittvis.com/envi). The method used a Gaussian model with a full width at half maximum (FWHM) equal to the band spacing provided (Mutanga and Skidmore, 2005). HYMAP provides 64 wavelengths covering the SWIR region. The resampled HYMAP spectra (Table 5.1) were used in all subsequent analyses.

Table 5.1: The spectral configuration of the HYMAP sensor. The sensor provides contiguous sampling of the spectra except for the water absorption wavelengths located at 1400 nm and 1900 nm

Module	Spectral range	Bandwidth across module	Average spectral sampling interval
SWIR1	1400 nm to 1800 nm	15 nm to 16 nm	13 nm
SWIR2	1950 nm to 2480 nm	18 nm to 20 nm	17 nm

5.2.2. Statistical analysis

5.2.2.1. The random forest algorithm

The Breiman-Cutler random forest algorithm (Breiman, 2001) is an ensemble method that grows multiple classification trees (*ntree*) and uses the entire forest as a complex composite classifier. As opposed to single classification trees, individual trees in the forest are maximally grown without any pruning, and the final classification of a given sample is decided by applying the majority rule over the votes of individual trees. The random forest (RF) algorithm introduces randomness in the classification process firstly by selecting only a random subset of candidate features (*mtry*) to determine the split at each node in a tree and secondly by using a bootstrap sample with replacement from approximately two thirds of the original training sample to create each tree in the forest. This implies that in some instances, training samples will be chosen more than once, while some training samples may not be used at all to grow individual trees in the forest. The excluded one third of the samples or the out of bag estimates are used to determine an internal measure of variable importance and an estimate of error (Breiman, 2001; Liaw and Wiener, 2002; Pal, 2005). The random forest library (Liaw and Wiener, 2002) developed for the R statistical software (R Development Core Team 2008) was used to implement the RF algorithm.

5.2.2.2. Boosting trees algorithm

The RF algorithm was compared to a competing ensemble method known as boosting trees (Freund and Shapiro, 1996). While the RF algorithm relies on bootstrapped aggregations of the original training data to generate individual trees in the ensemble, the boosting trees (BT) algorithm relies of the classification results from a previous iteration. Boosting methods grow a single classification tree, and weights are assigned to the training data. Hence, training data that are misclassified are increased in weight and training data that are correctly classified are decreased in weight. This forces subsequent classification trees to focus on the more difficult examples in the dataset. This entire process is repeated for a specific number of iterations and the resulting classification tree vote using a plurality rule (Chan and Paelinckx, 2008; Freund and Shapiro, 1996; Lawrence et al., 2004; Pal, 2007). For boosting 100 trees were created, however if the algorithm terminated earlier then a smaller iteration was subsequently used (Dietterich, 2000). The ada library (Culp et al., 2006) for the R statistical software (R Development Core Team 2008) was used to implement the AdaBoost version (Freund and Shapiro, 1996) of boosting by weighting.

5.2.3. Variable selection

The one-way analysis of variance (ANOVA) was used as a baseline filter approach. As mentioned earlier, when implementing the filter method, wavelengths with no statistical significance are discarded while the significant wavelengths are used as input variables into the relevant classification algorithm. In contrast, the wrapper approach searches for the best subset of wavelengths by using the classifier as part of the evaluation and the subset of wavelengths that produces the lowest misclassification error is then selected. The wrapper developed by Diaz-Uriarte and Alvarez de Andres (2006) was used for variable selection. Since the RF algorithm also provides an internal measure of variable importance, we also considered the top 10% and 20% of variables as ranked by the out of bag (OOB) samples. The sections below describe the OOB and wrapper methods in more detail.

5.2.3.1. Using the OOB method for variable selection

The RF algorithm returns three measures of variable importance (Breiman, 2001). The first measure is based on the number of times each candidate variable is selected, the second measure uses the Gini criterion, and the final measure utilizes the permutation of variables as an estimate of variable importance (Strobl et al., 2007). The most reliable measure is the permutation of variables which calculates variable importance as the mean decrease in classification accuracy using the out of bag samples (Breiman, 2001). As detailed in the study by Archer and Kimes (2008) the importance of each wavelength can be calculated as follows:

1. Predict the class membership for OOB sample using an unpruned classification tree, and then add the number of times the classification tree predicts the correct class for the OOB samples.
2. For wavelengths used in the study:
    - Permute the reflectance values of each wavelength in OOB sample.
    - Use the classification tree to predict the class membership for OOB sample using the permutated reflectance values, and then add the number of times the tree predicts the correct class for the OOB samples.
3. Finally, subtract the number of votes for the correct class in permutated dataset from the number of votes for the correct class in the original OOB sample.

Therefore, the relative loss in performance between the OOB dataset and the permutated dataset provides a ranking which can be used to select wavelengths for the final classification.

5.2.3.2. Using the wrapper method for variable selection.

The backward variable selection (BVS) approach was used for variable selection. The BVS method developed by Diaz-Uriarte and Alvarez de Andres (2006) built multiple random forests, and after building each random forest, iteratively discards 20% of the wavelengths with the smallest variable importance. After fitting all models, the BVS approach chooses the subset of wavelengths whose OOB error rate is within u standard errors of the minimum OOB error of all the forests created. Setting $u = 0$ selects the

subset of wavelengths with the smallest OOB error, and setting u = 1 selects the smallest subset of wavelengths but whose OOB error is within the sampling error from the best solution (Diaz-Uriarte and Alvarez de Andres, 2006). We used the varSelRF library (Diaz-Uriarte and Alvarez de Andres, 2006) for the R statistical software (R Development Core Team 2008) to implement the BVS method. According to Granitto et al. (2006) the selection of variables is an unstable process especially when $n < p$, and this could subsequently lead to the selection of very different subsets of explanatory variables for each replicate of the study. Therefore, we repeated the BVS approach (n = 1000) to determine the frequency with which the selected wavelengths appear in subsequent replicates of the study.

5.2.4. Classification accuracy

Classification accuracy in the absence of an independent test dataset can be determined by resampling of the original data (Molinaro et al., 2005). Several variants of the bootstrap resampling method have been introduced to estimate error (Efron and Tibshirani, 1993). For example, the leave-out-one bootstrap is based on a random sample that is drawn with replacement from n observations. For each draw, the observations (approximately 368 samples) that are excluded serve as a test dataset, and the training dataset has approximately 632 samples. However, this leads to an over estimation of error because a decrease of samples in the training dataset leads to an increase in bias (Molinaro et al., 2005). To correct the bias in error, the .632 and the .632+ estimators have been suggested (Efron and Thibshirani, 1997). Both estimators correct for bias by adding the underestimated resubstitution error. For the .632 estimator the weight (ω) is constant (ω=.632), whereas for the .632+ estimator , ω is determined by the no information rate (Efron and Thibshirani, 1997). The .632+ bootstrap error has been previously used to assess misclassification error in chemometric and genomic studies in applications where $n < p$ (Diaz-Uriarte and Alvarez de Andres, 2006; Granitto et al., 2006). In the present study the .632+ bootstrap method (n = 100) was applied as an “outer loop” to compare the three variable selection methods using the RF and BT algorithms, whereas the OOB error was used as an “inner loop” to guide the variable selection for the wrapper and the OOB variable selection methods (Diaz-Uriarte and Alvarez de Andres, 2006; Granitto et al., 2006). Therefore, the classification algorithms

are evaluated on a dataset which was not previously used for variable selection or for classification purposes. We used the errorest library (Peters et al., 2002) for the R statistical software (R Development Core Team 2008) to calculate the .632+ bootstrap error.

5.2.5. Class label and wavelength noise

In this study the robustness and stability of the classification algorithms are assessed against the introduction of noise. Remotely sensed data is likely to be noisy due to factors that include saturation of signal, missing scans, mislabelling, problems with the sensor, and geometry (DeFries and Chan, 2000). In order to determine if the RF algorithm would perform well under conditions where noise is introduced, we applied the algorithm to noisy data and then examined the resulting misclassification error as determined by 632+ bootstrap error (n = 100). Similar to the method implemented by Dietterich (2000), Breiman (2001), and Hamza and Larocque (2005), the values in the class labels and reflectance values were randomly altered in increments of 5% up to a maximum of 20%. As a result, the original values were replaced with alternate values chosen uniformly from all other possible values. Following suggestions by Zhu and Wu (2004), the impact of the two categories of noise (class labels and wavelength) were analysed independently because it would be difficult to consider the combined effects of both categories.

5.3. Results

5.3.1. Variable selection using the filter method

The ANOVA with a Tukey's HSD post hoc test was calculated at each measured wavelength for the healthy and green stage (H-G) class pair. The ANOVA results of the individual wavelengths (n = 64) are shown in Figure 5.1. The shaded areas in Figure 5.1 indicate specific wavelengths (n = 54) where the H-G class pair show a significant statistical difference in reflectance ($p < 0.001$). Wavelengths that have the potential to discriminate the H-G class pair are located in the following SWIR regions: between 1409 nm and 1613 nm (n = 16), 1739 nm and 1799 nm (n = 6), and 1952 nm and 2485 nm (n = 32). The majority of the significant wavelengths (59.25%) are predominately located between 1952 nm and 2485 nm (HYMAP SWIR 2). All significant wavelengths were then retained as input variables into the classification algorithms.

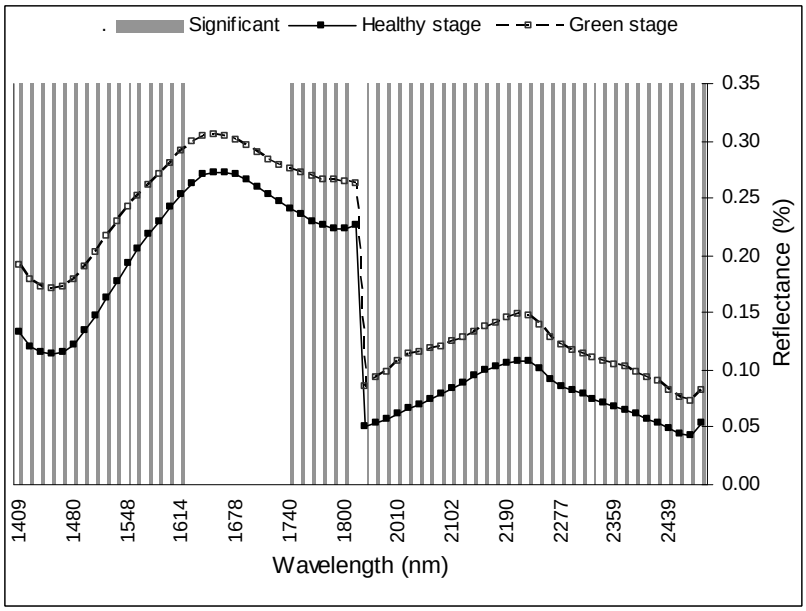


Figure 5.1: ANOVA results for the healthy-green stage using the resampled HYMAP spectra. The grey shades indicate regions of electromagnetic spectrum where there were significant differences ($p < 0.001$). For contextual purposes, the SWIR spectral reflectance for the green stage (n = 30) and for the healthy stage (n = 24) are shown.

5.3.2. Variable selection using the OOB method

Following suggestions by Diaz-Uriarte and Alvarez de Andres (2006), and Hamza and Larocque (2005), we optimized the *mtry* value by trying all possible values (n = 64). The default *mtry* value of eight produced the lowest OOB error (9.26%). Additionally, there was no significant increase in classification accuracy beyond 500 trees (*ntree*) in the ensemble. Figure 5.2 shows the importance of all wavelengths (n = 64) as determined by the OOB method. Noticeable in Figure 5.2 are the numbers of important wavelengths located between 1971 nm and 2101 nm (n = 8). Additional important wavelengths are located between 2326 nm and 2485 nm (n = 7). The wavelengths with the highest mean decrease in accuracy are located at 2028 nm and 2047 nm. Since the OOB method produces a ranking for all wavelengths, only the top 10% (n = 6) and 20% (n = 13) of the highest ranked wavelengths are considered for classification purposes.

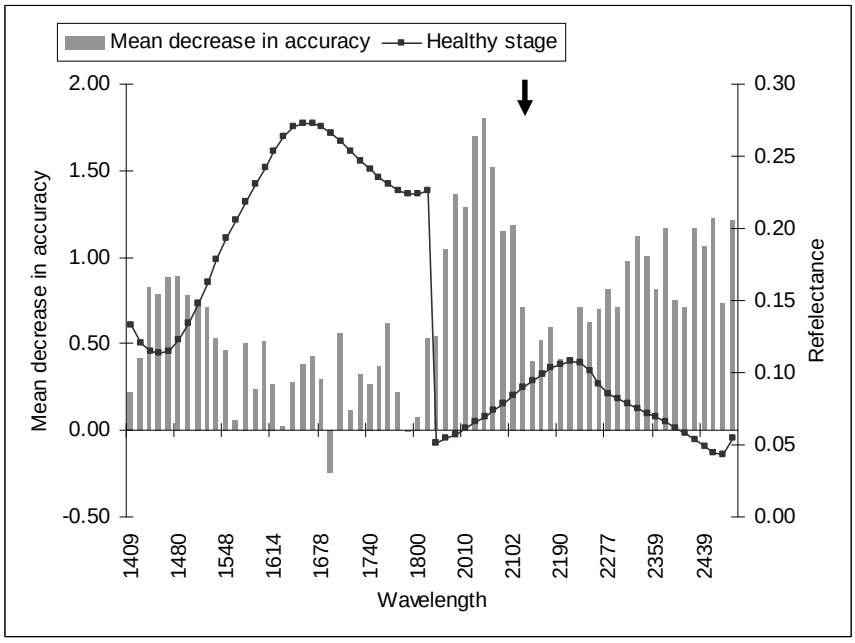


Figure 5.2: Variable selection using the random forest algorithm. The forest was created using all the resampled SWIR wavelengths. The wavelengths with the largest mean decrease in accuracy are shown by the black arrow .The average reflectance from the healthy stage is shown for contextual purposes.

5.3.3. Variable selection using the wrapper approach

The BVS approach iteratively built multiple random forests (n = 15) while discarding 20 % of the least important wavelengths as determined by the OOB error (Figure 5.3). We used an *ntree* value of 500 and the default *mtry* values for all iterations. Figure 5.3 shows that the lowest misclassification rate (7.41 %) as determined by the OOB error is obtained when using five variables located at the following wavelengths: 1990 nm, 2009 nm, 2028 nm, 2047 nm, and 2065 nm. Analogous to the findings of Diaz-Uriarte and Alvarez de Andres (2006), the *u* parameter had a minor effect on the results of the BVS. However, experiments showed that using $u = 1$ leads to a slightly more stable result with a smaller subset of wavelengths. Hence, we used $u = 1$ for all subsequent replicates of the study.

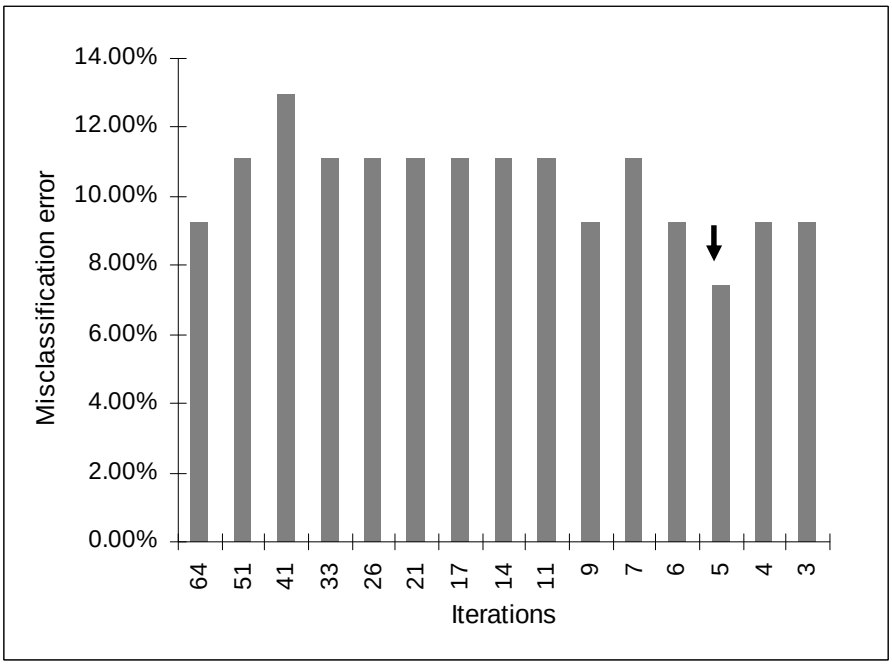


Figure 5.3: Misclassification error estimates for all random forest classifiers (n = 15) as determined by the OOB sample. The lowest misclassification error is shown by the black arrow.

The entire BVS approach (n = 1000) was repeated to determine the frequency with which the 1990 nm, 2009 nm, 2028 nm, 2047 nm, and 2065 nm wavelengths occur in the best selected subset of wavelengths. Although no wavelengths are selected more than 50 % of the time, Figure 5.4 shows that the BVS selected wavelengths have a much higher frequency of being selected than any of the other wavelengths in the SWIR domain. More specifically, the wavelengths are selected with the following frequency: 1990 nm (n = 196), 2009 nm (n = 103), 2028 nm (n = 407), 2047 nm (n = 466), and 2065 nm (n = 271). The 2028 nm and 2047 nm wavelengths are the most frequently selected wavelengths.

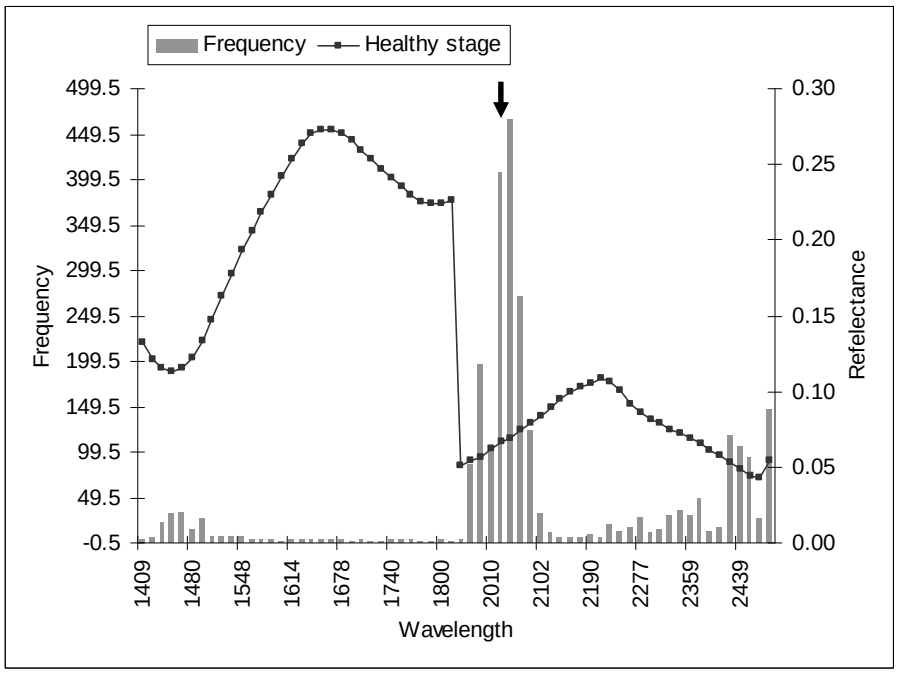


Figure 5.4: The frequency with which the backward variable selected wavelengths occur in the selected subset of wavelengths during each replicate (n =1000) of the study. The wavelengths that are most frequently selected by the approach are shown by the black arrow.

5.3.4. Classification results

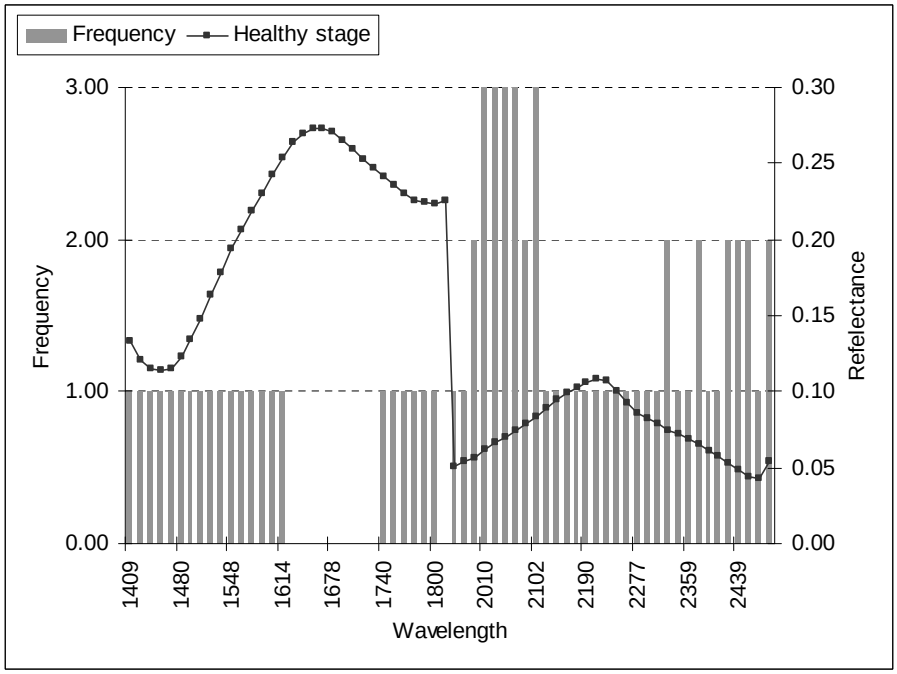


Figure 5.5: The selection of individual wavelengths by the analysis of variance (ANOVA), backward variable selection (BVS), and the out of bag (OOB) method.

In this section the results from the three variable selection methods are compared and the misclassification errors as determined by the .632+ bootstrap errors are reported on. For comparative purposes, Figure 5.5 shows the frequency and locations of wavelengths that were selected by all three variable selection methods (that is, ANOVA, OOB and BVS). The five wavelengths located at: 1990 nm, 2009 nm, 2028 nm, 2047 nm, and 2065 nm were selected by all three variable selection methods. Also noticeable is the absence of selected wavelengths between 1614 nm and 1740 nm. The various wavelengths were then used as input variables into the RF and BT algorithms. Table 5.2 reports on the misclassification errors for both algorithms as determined by their .632+ bootstrap errors.

Table 5.2: The misclassification error rate for the random forest (RF) and the boosting tree (BT) algorithms. Wavelengths selected by the analysis of variance (ANOVA), backward variable selection (BVS), and the OOB method (top 10 % and top 20 %) were used as input variables.

Algorithm	**All wavelengths**	**ANOVA**	**BVS**	**Top 10 %**	**Top 20 %**
Number of wavelengths	64	54	5	6	13
BT	7.43	7.51	7.54	7.57	6.82
RF	7.29	7.45	**6.14**	6.52	6.60

The .632+ bootstrap errors of both algorithms using the three variable selection methods are comparable and below eight percent. However, the RF algorithm produces slightly better classification results than the BT algorithm for all three variable selection methods including when all HYMAP SWIR wavelengths (n = 64) are used for classification. Using the wavelengths (1990 nm, 2009 nm, 2028 nm, 2047 nm, and 2065 nm) selected by the BVS method, as input variables into the RF algorithm (i) produces better accuracies than using all the HYMAP SWIR wavelengths (ii) produces the lowest overall misclassification error (6.14%) and (iii) produces the largest difference in error (1.4%) between the RF and BT algorithms. Wavelengths selected by the ANOVA (n = 54) produce the highest misclassification error (7.45 %) for the RF algorithm. Using the top 10% or 20% of wavelengths selected by the OOB method with the RF algorithm produces comparable results with the best solution. The BT algorithm achieves the lowest misclassification error (6.82%) when using the top 20% (n = 13) of wavelengths selected by the OOB method. Wavelengths selected by either the ANOVA or the BVS variable selection methods do not produce better classification results when compared to the use of all the wavelengths with the BT algorithm. Subsequently, the wavelengths selected by the BVS (n = 5) method were used to determine if the RF algorithm would perform well under conditions where noise is introduced.

5.3.5. Classification accuracy: Class label and wavelength noise

In this section the robustness of the RF algorithm is examined against permutations in the class labels and reflectance values. Additionally, the resulting misclassification error as determined by the .632+ bootstrap error is reported. For comparison purposes the misclassification error rate of the BT algorithm is also included. Figure 5.6 shows the .632+ bootstrap errors (n = 100) for all noise levels as a result of altering the class labels. Mislabelling of the class labels causes severe problems in the classification accuracy for both the machine learning algorithms. However, the BT algorithm shows a higher increase in misclassification error when compared to the RF algorithm at all noise levels. The RF algorithm has a minimum error of 12.56% and a maximum error of 18.15% compared to the original misclassification error of 6.14%. On the other hand, the BT algorithm has a minimum error of 13.27% and a maximum error of 21.62% compared to the original misclassification error of 7.54%.

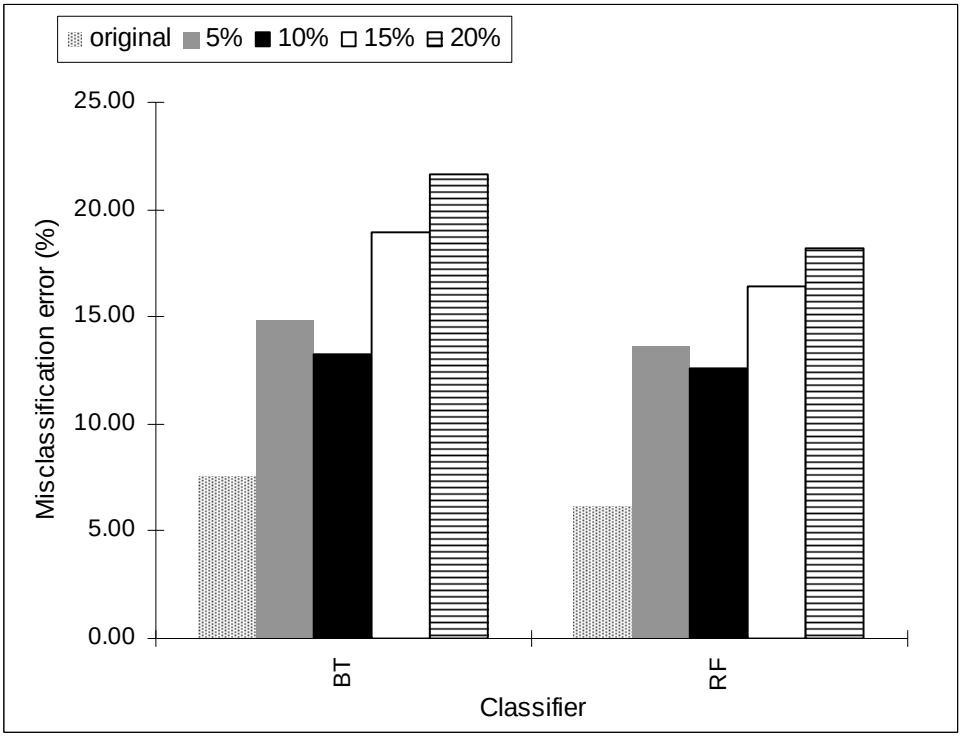


Figure 5.6: The .632+ bootstrap errors (n = 100) when random noise is introduced into the class labels.

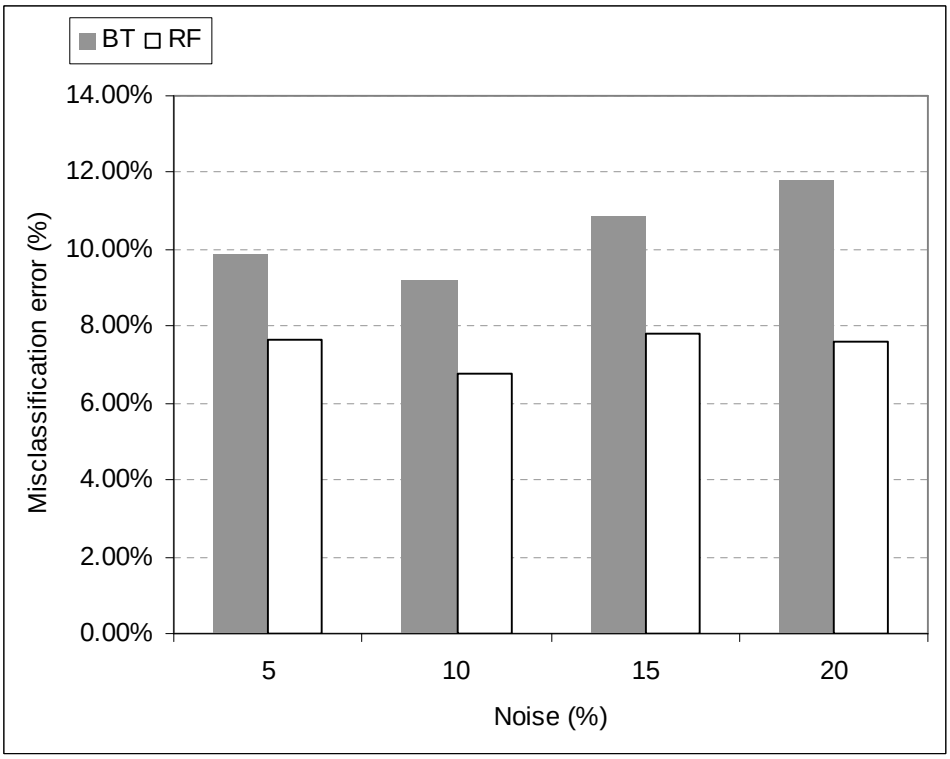


Figure 5.7: The .632+ bootstrap errors (n = 100) when random noise is introduced into the selected wavelengths.

Figure 5.7 shows the .632+ bootstrap errors (n = 100) when noise is introduced into the reflectance values of the BVS selected wavelengths. Altering the reflectance values has a more pronounced effect on the misclassification error of the BT algorithm rather than on the RF algorithm. The misclassification error rate for random forest remains below 8% for all noise levels, while the error rates for the BT algorithm are all above 8%. The misclassification error for the RF algorithm has a minimum error of 6.77% and a maximum error of 7.80% compared to the original misclassification error of 6.14%. In contrast the BT algorithm has a minimum error of 9.23% and a maximum error of 11.82% compared to the original misclassification error of 7.54%.

\

5.4. Discussion

5.4.1. Variable selection and classification accuracy

In recent years the RF algorithm has gained popularity as an effective classification method in the remote sensing domain (Chan and Paelinckx, 2008; Gislason et al., 2006; Ham et al., 2005; Lawrence et al., 2006; Pal, 2005). Results from the present study confirm that the RF algorithm is a robust and accurate method for the combined purposes of variable selection and for the classification of hyperspectral data in an application where (i) the number of samples is limited and (ii) where classes have similar spectral characteristics. Statistically, the RF algorithm deals with the "small *n* large *p*" problem, by employing a user defined random selection of variables *(mtry*) to grow each classification tree (Breiman, 2001). Hence, each classification tree is grown using only a small user defined subset of candidate variables, and the problems associated with "small *n* and large *p"* problem are thereby avoided. Results from this study additionally demonstrate that the suggested default *mtry* value ($\sqrt{p}$) described by Liaw and Wiener (2002) also achieves the best overall classification accuracies. Furthermore, by limiting the number of variables used for each split and using classification trees that are not pruned, the computational complexity of the RF algorithm is reduced, thus making it well suited for hyperspectral data (Gislason et al., 2006).

Researchers have shown that the RF algorithm achieves comparable if not better results than other competing classification methods (Chan and Paelinckx, 2008; Gislason et al., 2006; Pal, 2005). Results from the present study confirm that the RF algorithm has lower misclassification errors than the BT algorithm for all three variable selection methods including when all HYMAP SWIR wavelengths (n = 64) are used. However, differences in error between the RF and BT algorithms are minimal (< 2 %). Chan and Paelinckx (2008), and Gislason et al. (2006) reported similar findings on the small differences in the classification error between the RF and BT algorithms. Nevertheless, wavelengths selected by the wrapper approach produce the lowest misclassification error (6.14%) in conjunction with the RF algorithm. More importantly,

the wrapper method simplified the classification process and identified the smallest number of wavelengths that offer the best discriminatory power. Using the wrapper method, we only used approximately 8% (n = 5) of the total number of HYMAP wavelengths (n = 64) while still producing the best overall classification accuracies. Additionally, by using the wrapper method we did not have to specify the number of HYMAP wavelengths required for the classification process, rather the method adaptively selected the minimum number of wavelengths that provide the best classification accuracy.

While the RF algorithm has been shown to provide a sensible means for variable selection in a hyperspectral application, where $n < p$, however it should be noted that the data used in this study were continuous in nature. Since random forest is a nonparametric method and can handle continuous as well as discrete data the question remains whether the internal variable selection method will perform as well in situations where the data may vary in scale or in the number of categories present. Genomic studies by Strobl et al. (2007) show that the RF algorithm is especially biased towards variables that contain a large number of categories. This phenomenon needs to be investigated using remotely sensed data.

5.4.2. Model robustness and the introduction of noise

For all levels of noise, the RF algorithm produces lower misclassification errors than the BT algorithm when either the class labels or the reflectance values are altered. The BT algorithm is known to be particularly sensitive to noise in the training dataset because the algorithm places emphasis on the noisy data that is, after a few iterations most of the data with large weights are cases where the noisy data have been misclassified (Dietterich, 2000; Lawrence et al., 2004). Similarly, using 28 non-remote sensing datasets, Hamza and Larocque (2005) found that the RF algorithm is more robust with respect to noise than other tree based ensemble methods like boosting trees. The robustness of the RF algorithm can be explained by the ability of the classification algorithm to exploit the noise in the dataset to create a more diverse classifier (Breiman, 2001).

However, misclassification of the class labels results in a more severe decline in classification accuracy than alteration of the reflectance values of the wavelengths does. Similar results were reported by DeFries and Chan (2000) when they evaluated tree based classification algorithms for landcover classifications. Using non-remote sensing datasets, Zhu and Wu (2004) also reported that classification accuracies decline with an increase in class noise especially when there are limited samples. According to Zhu and Wu (2004) possible explanations for the severity of the class noise on the classification algorithm are that (i) there are multiple wavelengths in comparison to the uniqueness of class labels, and consequently the noise introduced from a limited number of wavelengths could have only a limited impact on the classification accuracy, and (ii) some of the wavelengths make only a limited contribution to the classification algorithm.

5.4.3. Understanding SWIR reflectance characteristics of green stage Sirex noctilio infestations

The importance of the RF algorithm is not only to improve classification accuracy or to reduce data dimensionality but also to deepen our understanding of which SWIR wavelengths are most suitable for discriminating the green stage of infestation. Results show that the HYMAP sensor provides the necessary spectral sensitivity to detect anomalies in SWIR reflectance that will allow researchers to effectively detect and monitor the green stage of infestation. More specifically, this study has shown that wavelengths located at 1990 nm, 2009 nm, 2028 nm, 2047 nm, and 2065 nm have the greatest potential for discriminating the green stage. According to Kumar et al. (2001), the main absorption features located in these regions are related to protein, starch, and nitrogen. Physiological evidence suggests that during the green stage, *S. noctilio* injects toxic mucus and a fungus that causes (i) an increase in enzyme activity associated with the conversion of foliar starch reserves to soluble sugars and (ii) a rise in respiratory activity which then results in the rapid depletion of soluble sugar levels. It is only during the later stages of infestation that the breakdown of chlorophyll occurs which is subsequently followed by the collapse of vascular tissue that causes chlorosis, wilting, and premature needle fall (Neumann and Minko, 1981). The results from this study

show that there is a link between the HYMAP spectral measurements and existing physiological research thereby improving the chances of detecting the green stage at an airborne level. However, for verification purposes, a more detailed study relating hyperspectral data to physiological measurements has to be carried out.

5.5. Conclusion

In this paper we demonstrated that the random forest algorithm can accurately discriminate between healthy trees and the early stages of *S. noctilio* infestation using SWIR wavelengths. More specifically we have shown that the wrapper method that uses random forest algorithm as part of the evaluation process produces the smallest subset of wavelengths with the lowest misclassification error. Wavelengths located at 1990 nm, 2009 nm, 2028 nm, 2047 nm and 2065 nm have the greatest potential for discriminating the green stage. Additionally, the random forest algorithm performs better than the boosting trees algorithm when noise is introduced into the class labels or the selected wavelengths. Overall, the results from this study additionally confirm that the random forest algorithm is a robust and accurate method for the combined purpose of variable selection and for the classification of hyperspectral data in an application where (i) the number of samples is limited and (ii) where classes have similar spectral characteristics.

Acknowledgements

We thank Sappi for allowing us access to the Pinewoods plantations. The contributions of Marcel Verleur in identifying *Sirex noctilio* infestations are gratefully acknowledged. We thank Wayne Jones for assisting with the sampling of pine needles. Additionally, we appreciate all the computer programming assistance provided by Chris Muncaster. Eric Economon from the Agricultural Research Centre (ARC) of South Africa provided assistance with the ASD spectroradiometer. Funding for this research was provided by the National Research Foundation (NRF) South Africa.

CHAPTER 6:

A comparison of regression tree based ensemble methods: Predicting *Sirex noctilio* induced water stress

* This chapter is based on:

Ismail, R. and Mutanga, O., *in review*. A comparison of regression tree based ensemble methods: Predicating *Sirex noctilio* induced water stress of pine forests in KwaZulu-Natal, South Africa. International Journal of Geoinformation and Earth Observation. Special edition on African remote sensing .

.

Abstract

In this study we evaluate the performance of various regression tree ensembles using hyperspectral data. More specifically, we compare the performance of bagging, boosting, and random forest ensembles to predict *Sirex noctilio* induced water stress in *Pinus patula* trees using several spectral parameters (n = 9) derived from hyperspectral data. Results from the study show that the random forest ensemble achieves the best overall performance (R^2 = 0.73) and that the predictive accuracy of the ensemble is statistically different ($p < 0.001$) from the bagging and boosting ensembles. Additionally, by using the random forest ensemble as a wrapper we simplified the modelling process and identified the minimum number (n = 2) of spectral parameters that offer the best overall predictive accuracy (R^2 = 0.76). The water index and the $Ratio_{975}$ indices have the best ability to assay the water status of *S. noctilio* infested trees thus making it possible to remotely predict and quantify the severity of damage caused by the wasp.

Keywords: Regression trees, ensembles, random forest, *Sirex noctilio*

6.1. Introduction

Sirex noctilio is currently the most destructive pest of conifers in South Africa, and the wasp is currently causing considerable tree mortality in *Pinus patula* forests located in the southern parts of the country. Recent estimates indicate that 35, 000 ha of *P .patula* forests are infested and dying (Hurley et al., 2007). In anticipation of the future availability of hyperspectral data in South Africa (van Aardt and Coppin, 2006), there is a keen interest amongst researchers to apply novel methods and techniques that will allow for the accurate prediction and quantification of *S. noctilio* infestations.

Regression trees (Breiman et al., 1984) have been widely used for prediction purposes in the remote sensing domain (DeFries et al., 1997; Hansen et al., 2002; Lobell et al., 2007; Michaelson et al., 1994). However, regression trees are very sensitive to small perturbations in the training dataset and have been identified as unstable learners that are prone to overfitting (Breiman, 1996). Simply stated, very small changes in the values of the training dataset can lead to significant changes in the selection of variables that are used to create the regression tree (Hastie et al., 2001). Therefore, the instability of regression trees introduces uncertainty in their interpretation and limits their predictive performance (Elith et al., 2008).

Bagging (Breiman, 1996), boosting (Freund and Shapiro, 1996; Friedman, 2002) and random forest (Breiman, 2001) are popular ensembles that have been used to improve the performance of unstable learners (Hamza and Larocque, 2005). As a result of their improved performance, these techniques have been implemented in a wide variety of remote sensing applications (Gislason et al., 2006; Ham et al., 2005; Lawrence et al., 2004; Lawrence et al., 2006; Pal, 2005). However, to the best of our knowledge (i) remote sensing applications thus far have focused on using classification trees rather than using regression trees as the base learner and (ii) the random forest ensemble has been more widely advocated as a classifier rather than the bagging or boosting ensembles.

The question then arises: How would bagging, boosting and random forest ensembles perform in regression type applications? Initial research carried out by Breiman (2001) on machine learning datasets revealed that the results were mixed. The random forest ensemble always produced better results than conventional bagging

ensembles while in some of the datasets, a modified version of bagging known as adaptive bagging outperformed the random forest ensemble. More recently, Prasad et al. (2006) compared the random forest and bagging ensembles for predicting species distribution under climate change scenarios. Results from the study concluded that the random forest and bagging ensembles have similar predictive abilities. To date, no research has compared the performance of regression tree ensembles using remotely sensed data. Consequently, the objective of this study is to compare the performance of the random forest, bagging and boosting ensembles for prediction purposes using hyperspectral data. More specifically, various regression tree ensembles are compared for predicting *S. noctilio* induced water stress in *P. patula* trees using several spectral parameters derived from hyperspectral data.

6.2. Materials and methods

6.2.1. Spectral reflectance and water content measurements

Using the analytical spectral devices (ASD) spectroradiometer (Analytical Spectral Devices, 2002), reflectance measurements (350 nm to 2500 nm) were obtained from *P. patula* trees located at the Sappi Pinewoods plantation (centroid $30°4'13.83" E$ and $29°38'36.06" S$) in KwaZulu-Natal, South Africa. Several spectral measurements were taken from healthy (no visible indication of *S. noctilio* infestation), green (appearance of resin droplets and presence of ovipositors), and from red (wilting of the infested tree and leaves appear reddish-brown) stage trees (Ismail et al., 2008a; Ismail et al., 2007). In total, 66 spectral measurements were acquired from healthy (n = 24), green (n = 30), and red (n = 12) stage trees. More details on the sampling scheme and the acquisition of the spectral reflectance measurements using the ASD are provided by Ismail et al. (2008a). Once the spectral measurements were completed, foliar samples from each stage (n = 66) were immediately sealed in a plastic bag and sent to the Institute of Commercial Forestry Research (ICFR) laboratory for water content analysis**.** Following the procedure described by Stimson et al. (2005), water content (WC) was calculated as follows:

$$WC(\%) = ((FW - DW) / FW) * 100 \quad (1)$$

Where FW is the fresh weight of the sample, and DW is the weight of the sample after being dried in an oven for approximately 24 hours at 60 ^{o}C.

6.2.2. Spectral parameters

Using the ASD reflectance measurements, several spectral indices (n = 7) were calculated. The spectral indices included: simple ratios, normalized ratios, and three band ratios (Table 6.1). Additionally, continuum removal was applied to water absorption features located at $R_{920\text{-}1120}$ and $R_{1070\text{-}1320}$ (Pu et al., 2003). The continuum is a convex hull fitted over the top of a spectrum utilizing straight-line segments that connect local spectra maxima (Clark and Roush, 1984; Kokaly and Clark, 1999). Although previous studies have calculated several parameters from the continuum-removed absorption features (Clark and Roush, 1984; Kokaly, 2001; Mutanga and Skidmore, 2003; Pu et al., 2003), we used only the band depth (BD) parameter, which is computationally efficient and therefore more suitable for the practical application of this study (Mutanga and Skidmore, 2004b).

Table 6.1: The various spectral indices (n = 7) that were used in the study.

Spectral Indices	Formula	Reference
Water index	$WI = \frac{\rho 900}{\rho 970}$	Peñuelas et al. (1997)
Normalized difference water index	$NDWI = \frac{\rho 860 - \rho 1240}{\rho 860 + \rho 1240}$	Gao (1996);
Normalized difference vegetation index	$NDVI = \frac{\rho 860 - \rho 690}{\rho 860 + \rho 690}$	Rouse et al. (1973)
$Ratio_{975}$	$Ratio_{975} = \frac{2\rho_{960-990}}{\rho_{920-940} + \rho_{1090-1110}}$	Pu et al. (2003)
$Ratio_{1200}$	$Ratio_{1200} = \frac{2\rho_{1180-1220}}{\rho_{1090-1110} + \rho_{1265-1285}}$	Pu et al. (2003)
Moisture stress index	$MSI = \frac{\rho 1600}{\rho 819}$	Hunt and Rock (1989)
Normalized difference infrared index	$NDII = \frac{\rho 819 - \rho 1600}{\rho 819 + \rho 1600}$	Hardinsky et al. (1983)

6.2.3. Statistical analysis

Regression tree ensembles were used to predict the water content as a function of multiple spectral parameters (n = 9) using a hold out sample. This was done by repeatedly and randomly (n = 1000) dividing the original dataset into training (70%) and test datasets (30%). For each run, regression tree ensembles developed on the training dataset (n = 46) were then used to predict the water content on the test dataset (n = 20). The final predictive accuracy used to compare the regression ensembles consisted of an averaged adjusted R^2 value for all the runs carried out. All statistical analysis was carried out using the R package (R Development Core Team 2008). The section below briefly describes the regression tree ensembles used in this study.

6.2.3.1. Bagging ensembles

Bagging or bootstrap aggregation (Breiman, 1996) is a relatively simple idea that uses many bootstrap samples (Efron and Tibshirani, 1993) with replacement from the original dataset and then applies a regression tree to each bootstrap sample. The results from each regression tree are then averaged to obtain the overall prediction. When a bootstrapped sample is drawn, approximately 37% of the dataset is excluded from the sample with the remaining data being replicated to bring the dataset to full size (Prasad et al., 2006). This implies that to grow regression trees in the ensemble, some training samples will be chosen more than once, while some training samples may not be used at all. The excluded one third of the samples is known as the out of bag samples (OOB), while the replicated dataset is known as the in bag samples (inBag).

6.2.3.2. Random forest ensembles

The random forest (Breiman, 2001) ensemble is similar to bagging ensembles, in that regression trees are grown on bootstrap samples and the final prediction of a given sample is decided by averaging the results of many regression trees. However, the random forest ensemble has the additional modification of selecting only a random subset of candidate features (*mtry*) to determine the split at each node of a tree. Regression trees in the forest are grown to maximum size until no further splits are possible and the trees are not pruned back. As each tree is grown, it makes predictions on the OOB sample for that particular tree. The prediction error can then provide an unbiased assessment of the predictive accuracy, since the OOB sample is not used in the training process. Additionally, the random forest ensemble provides an internal measure of variable importance using the OOB sample. The variables associated with the OOB sample are randomly permutated and regression trees are grown on the modified dataset. The importance measure of each variable is then calculated as the difference in the mean square error between the original OOB predicted dataset and the modified dataset (Breiman, 2001; Liaw and Wiener, 2002).

6.2.3.3. Boosting ensembles

While bagging and random forest ensembles rely on bootstrapped aggregations of the original training data to generate trees in the ensemble, boosting ensembles rely on the results from a previous iteration. Boosting ensembles use a forward stagewise procedure to iteratively fit trees to the training dataset and gradually increases emphasis on poorly modelled observations (Elith et al., 2008). For regression related problems, the boosting ensemble grows the first regression tree to maximally reduce the loss in predictive performance (such as deviance), and the next tree then focuses on the variation in the response (that is, residuals) that could not be explained by its predecessor. So, boosting regression trees assumes the form of a functional gradient descent (Friedman, 2002) that minimizes the loss function by adding at each step a new regression tree that best reduces the loss function. The final model is a linear combination of many trees with the contribution of each tree usually shrunk by a learning rate (*lr*) to achieve best performance (Elith et al., 2008).

6.2.3.4. Model optimization

Using the training dataset (n = 46), the optimal input parameters for models were selected based on the prediction error as calculated by the tenfold cross validation (CV). In tenfold CV, the dataset is divided into ten subsets of approximately equal size. The ensembles are then trained ten times, each time training on nine subsets and using the omitted subset to calculate the prediction error. In this study, the optimal input parameters for the ensembles are then selected based on the lowest prediction error as defined by the root mean square error (RMSE).

For random forest ensembles, the effect that the number of randomly selected variables (*mtry*) had on the prediction error was examined (Hamza and Larocque, 2005). The *mtry* value was optimized by creating random forest ensembles for all possible *mtry* values (n = 9) and then selecting the optimal *mtry* value based on the RMSE across all forests. For bagging ensembles, researchers have suggested that using between 25 trees to 50 trees in the ensemble is sufficient (Breiman, 1996; Sutton, 2005). However, Hastie et al. (2001) showed that there is a considerable amount of improvement in prediction

error if the number of trees is increased from 50 to 100. To optimize bagging ensembles, the number of trees (*nbag*) in the ensemble was varied by adding 25 trees at a time and then recording the resulting RMSE up to a maximum of 500 trees.

According to Elith et al. (2008), there are two important parameters that need to be optimized for boosting ensembles. The first parameter is the learning rate (*lr*) which determines the contribution of each tree to the final model and the second parameter is the tree complexity (*tc*) which controls whether interactions are fitted or not. Following Elith et al. (2008), boosting ensembles of 30,000 trees were fitted over a range of values for *tc* and *lr*. For each combination of *tc* and *lr*, we then identified the number of trees (*n.tree*) that achieved the lowest RMSE.

6.2.3.5. Variable selection

In order to simplify the modelling process it is important to identify the fewest number of spectral parameters that offer the best predictive power and help in the interpretation of the final model. To address this issue, a wrapper approach (Kohavi and John, 1997) was implemented. The wrapper approach searches for the best subset of spectral parameters by using the regression tree ensemble as part of the evaluation process. More specifically, a backward elimination greedy search function (Guyon and Elisseeff, 2003) was implemented. The search function starts with all the spectral parameters (n = 9) and then progressively eliminates the least promising spectral parameters. The nested subset of spectral parameters with the lowest RMSE is then selected. According to Kohavi and John (1997), the resulting subset of spectral parameters should be evaluated on an independent test set that was not used during the variable selection process. For comparative purposes the final subset of spectral parameters was evaluated using (i) the hold out test dataset (n = 20), (ii) ten fold CV, and (iii) the OOB samples.

6.3. Results

6.3.1. Model optimization

In an attempt to streamline the model building and evaluation process, we optimized the regression tree ensembles using the RMSE as our selection guide. Figure 6.1a shows that the lowest RMSE (8.19) for the random forest ensemble is obtained when using an *mtry* value of two. The default *mtry* value (1/3 of the total number of variables) as suggested by Breiman (2001) obtains a higher RMSE. Increasing the *mtry* value does not produce a lower RMSE. We therefore used an *mtry* value of two for all subsequent analyses. Figure 6.1b shows that the RMSE values for the bagging ensemble ranges from a minimum of 9.35 (*nbag* = 300) to a maximum of 10.79 (*nbag* = 475). Therefore an *nbag* value of 300 was used for all subsequent analyses. As suggested by Elith et al. (2008), for the boosting ensemble, the number of trees (*n.tree*) that achieved the lowest RMSE was identified for each combination of *tc* (1, 2, 3, 5, 7, and 10) and *lr* (0.1, 0.05, 0.01, 0.005, 0.001, and 0.005). Results indicated that the lowest RMSE (9.57) for the boosting ensemble is obtained with a *tc* value of one, a *lr* value of 0.01, and with the ensemble consisting of 1000 trees.

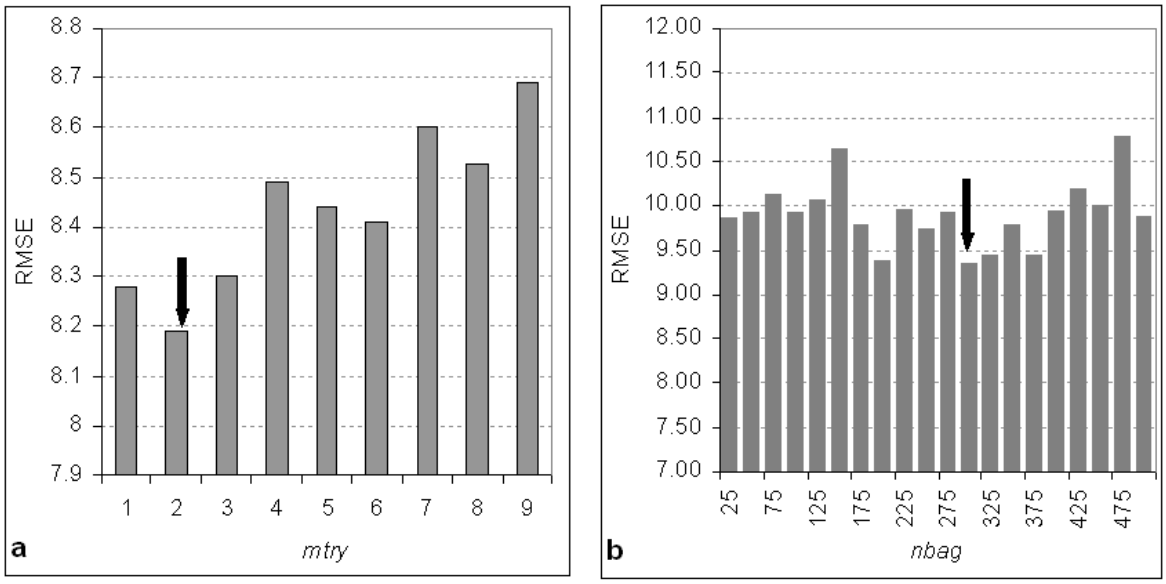


Figure 6.1: Insert (a) shows the root mean square error (RMSE) obtained for all possible *mtry* values (n = 9) for the random forest ensemble, and insert (b) shows resulting RMSE up to a maximum of 500 trees for the bagging ensemble. The lowest RMSE is indicated by the black arrow.

6.3.2. Comparison between bagging, boosting, and random forest

As mentioned earlier, the performance of the bagging, boosting, and random forest ensembles were compared using the adjusted R^2 values that were calculated from the repeated hold out samples. Figure 6.2 shows the histogram of adjusted R^2 values obtained over all runs used in this study (n = 1000). There is a narrow confidence interval for all the regression tree ensembles implying that the methods predicted with high precision (Mutanga et al., 2004). In order to assess whether the random forest ensemble is significantly better or worse than bagging and boosting ensembles, a Bonferroni corrected, one tailed paired *t* test was carried out. Results from paired *t* test indicated that there is a significant difference between random forest and boosting ensembles ($t = 6.24$, $p < 0.001$) and between the random forest and bagging ensembles ($t = 8.68$, $p < 0.001$). However, there was no significant difference ($t = 2.23$, $p > 0.05$) between the adjusted R^2 values of the boosting and bagging ensembles.

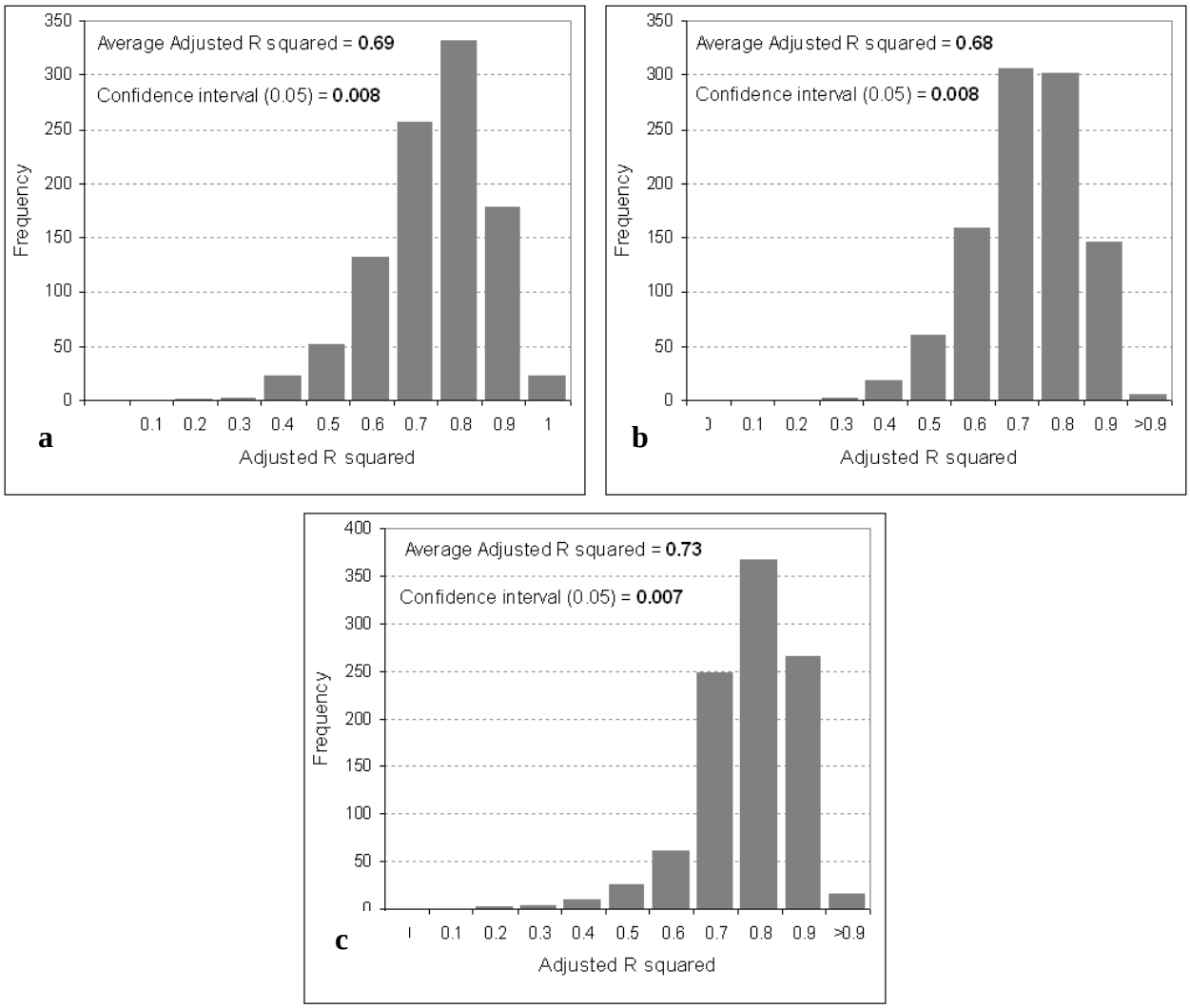


Figure 6.2: Histograms showing the frequency of the adjusted R^2 values for the regression tree ensembles used in this study. Insert (a) shows the distribution of the adjusted R^2 values for the bagging ensemble, insert (b) shows the distribution of the adjusted R^2 values for the boosting ensemble, and insert (c) shows the distribution of the adjusted R^2 values for the random forest ensemble.

The average performance of all three predictors is comparable, with the adjusted R^2 values ranging between 0.68 and 0.73 (Table 6.2). However, the random forest ensemble produces the best overall performance with an average adjusted R^2 value of 0.73. The adjusted R^2 value for a single regression tree was also calculated. As expected, the single regression trees obtain the lowest predictive performance with an adjusted R^2 value of 0.58. Using the random forest ensemble will produce a 15%

increase in predictive accuracy when compared to single regression trees, a 4% increase in accuracy when compared to bagging and a 5% increase in accuracy when compared to boosting. To check the validity of the comparisons among the regression tree ensembles, the RMSE was also calculated using the hold out samples. The results show that the random forest ensemble produces the lowest RMSE (Table 6.2).

Table 6.2: The average adjusted R^2 and RMSE values obtained by the bagging, boosting and random forest ensembles.

Model	Adjusted R^2	RMSE
Random forest	0.73	8.33
Boosting	0.68	10.27
Bagging	0.69	9.19

6.3.3. Variable selection

The random forest ensemble was used for variable selection since this ensemble produced the best predictive accuracy. However, before the variable selection process was carried out, the relative importance of individual spectral parameters were examined. Figure 6.3 shows the importance of individual spectral parameters as determined by the random forest OOB sample. The spectral parameters are ranked according to their importance during each run (n =1000) that was carried out during the ensemble comparison phase of the study (Section 6.3.2). For example, if the spectral parameter had the highest difference (RMSE) between the OOB predicted data and the permuted dataset it was ranked first for that particular run, and if the spectral parameter had the second highest difference in RMSE it was ranked second and so on.

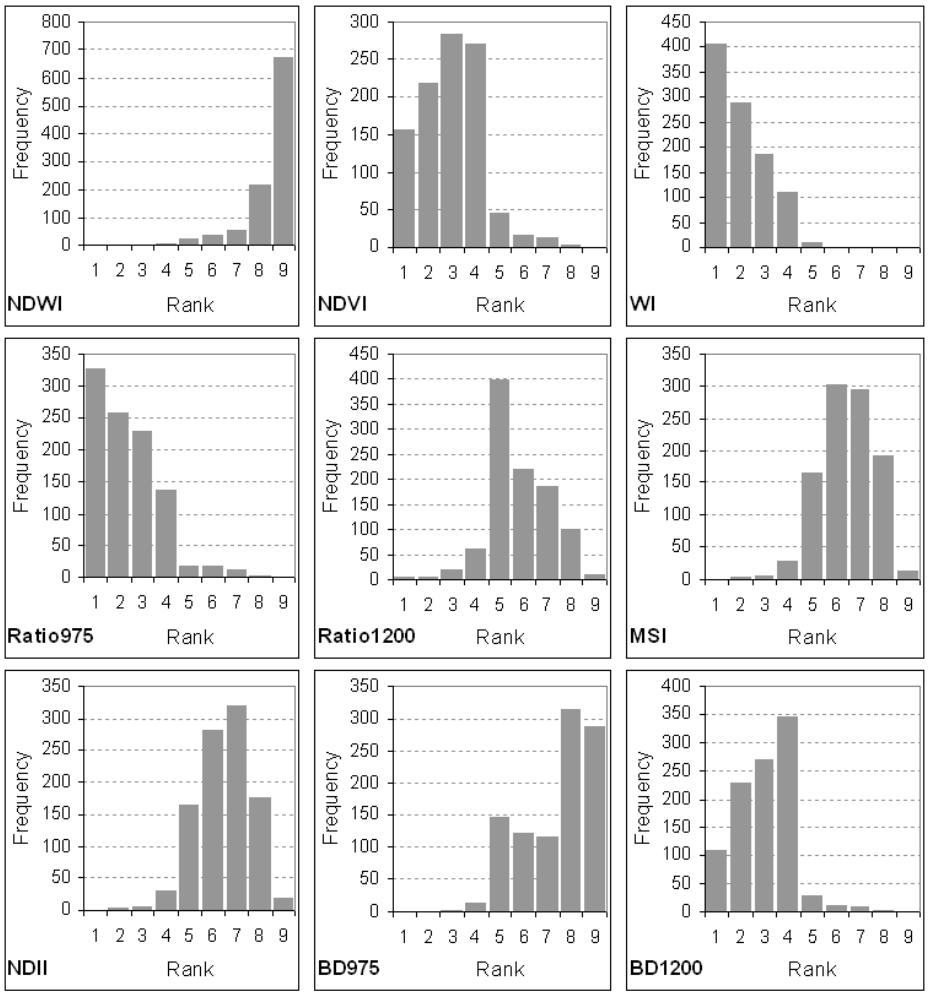


Figure 6.3: Histograms showing the ranked importance of the spectral parameters used in this study.

To determine if the spectral parameters used in this study are statistically different in their importance in the modelling process, a rank analysis was performed using Friedman's ANOVA by ranks. The overall test was significant ($p < 0.001$) indicating that the spectral parameters are statistically different. The average rank as calculated by Friedman's ANOVA for the spectral parameters are as follows: WI (2.04), $Ratio_{975}$ (2.38), NDVI (2.94), BD_{1200} (3.03), $Ratio_{1200}$ (5.80), NDII (6.46), MSI (6.47), BD_{975} (7.43), and NDWI (8.45). These rankings were subsequently used to determine the sequence in which to eliminate variables using the backward elimination search function.

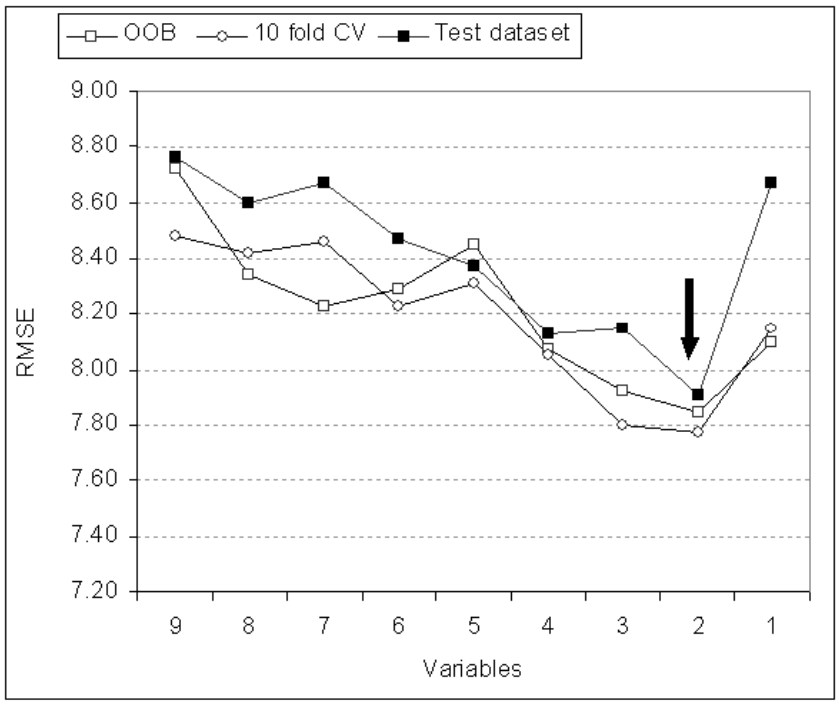


Figure 6.4: Variable selection using the backward elimination search function. The resulting RMSE for the OOB sample, tenfold CV, and the test dataset are shown.

Figure 6.4 shows the results of the variable selection process. As the spectral parameters were progressively eliminated by the backward elimination search function, the RMSE generally decreased, with the lowest RMSE obtained by using only two variables (WI and $Ratio_{975}$). Using WI and $Ratio_{975}$ produced the lowest RMSE using the hold out test dataset (7.91), tenfold CV (7.77), and the OOB sample (7.85). Subsequently, the adjusted R^2 value was recalculated for the random forest ensembles using WI and $Ratio_{975}$ as input variables. Results indicated that, by using WI and $Ratio_{975}$ an adjusted R^2 value of 0.76 is obtained by the random forest ensemble.

6.4. Discussion

In recent years, the random forest ensemble has gained popularity as an effective classification method in the remote sensing domain (Chan and Paelinckx, 2008; Gislason et al., 2006; Ham et al., 2005; Lawrence et al., 2006; Pal, 2005). Results from

the present study confirm that the ensemble is a robust and accurate method for regression applications as well. Regarding an adjusted R^2 value, the random forest ensemble produces the best overall performance (0.73), and the predictive accuracy of the ensemble is statistically different from the bagging and boosting ensembles. However, there was no statistically difference between the bagging and boosting ensembles. Similar results were obtained by Hamza and Larocque (2005), when they carried out an empirical comparison of ensemble methods using classification trees.

Besides obtaining the best overall predictive accuracy, using the random forest ensemble as a wrapper allowed the modelling process to be simplified, and identified the minimum number of spectral parameters that offer the best predictive accuracy. Using the backward elimination search function, only two spectral parameters were used while still producing an improved predictive accuracy (R^2 = 0.76). More specifically, results show that WI and the Ratio $_{975}$ indices have the best ability to assay the water content of *S. noctilio* infested trees thus making it possible to remotely quantify the severity of damage caused by the wasp.

The ability of WI and the Ratio $_{975}$ to quantify water content can be explained by the significant variation in water content in the healthy, green, and red stages. Physiological research has shown that tree mortality due to *S. noctilio* infestation is linked to the combined effects of a toxic mucus and the fungus *Amylostereum areolatum* that is injected into the tree by the female wasp during oviposition (Slippers et al., 2003). The mucus changes the water balance of the tree, thereby creating conditions that are ideal for the growth and spread of the fungus. In turn, the fungus rots and dries the wood, providing a suitable environment for the survival and development of the insect larvae (Slippers et al., 2003).

Additionally, in an experiment to assess the impact of mucus and the fungus on tree physiology, Couts (1970) found that, of the trees that died, their moisture content had decreased rapidly after only two to three weeks of infection. So, using indices like WI and the Ratio$_{975}$ which directly measure spectral variance caused by varying plant water status (Eitel et al., 2006), makes it possible to detect the initial stages of infestation or the green stage where there is minimal needle loss and the canopy appears green and indistinguishable from healthy trees.

6.5. Conclusion

The results from this study show that (i) there is a strong link between existing spectral indices (WI and the Ratio $_{975}$) and the water status of *P. patula* foliage of the tree thereby improving the chances of remotely detecting *S. noctilio* at a landscape level (ii) the random forest ensemble provides the best overall predictive accuracy when compared to the boosting and bagging ensembles, and (iii) using the random forest ensemble as part of a wrapper allowed the modelling process to be simplified, and identified the minimum number of spectral parameters that offer the best predictive accuracy. Ultimately, this study establishes the foundation for the potential upscaling of results to either an airborne or spaceborne platform. This is especially pertinent since it is envisaged that South Africa will soon launch the ZASat-003 satellite that will carry a hyperspectral sensor thus making high spectral resolution data more accessible and available to researchers in the country.

Acknowledgements

We thank Sappi for allowing us access to the Pinewoods plantations. The contributions of Marcel Verleur in identifying *Sirex noctilio* infestations are gratefully acknowledged. We thank Wayne Jones for assisting with the sampling of pine needles. Additionally, we appreciate all the computer programming assistance provided by Chris Muncaster. Eric Economon from the Agricultural Research Centre (ARC) of South Africa provided assistance with the ASD spectroradiometer. Funding for this research was provided by the National Research Foundation (NRF) South Africa.

CHAPTER 7:

Modelling the potential distribution of pine forests that are susceptible to *Sirex noctilio* infestations

* This chapter is based on:

Ismail, R., Mutanga, O., Kumar, L., *in review*. Modelling the potential distribution of pine forests that are susceptible to *Sirex noctilio* infestations in Mpumalanga, South Africa. *Transactions in GIS.*

Abstract

Reducing the impact of the siricid wasp, *Sirex noctilio* is crucial for the future productivity and sustainability of commercial pine resources in South Africa. The present study presents a machine learning model that serves as a spatial guide and allows forest managers to focus their existing detection and monitoring efforts on key areas and proactively adopt the most appropriate course of intervention. The random forest algorithm is implemented within a spatial framework to determine which pine forests in Mpumalanga are highly susceptible to *S. noctilio* infestations. Results indicate that the majority (63%) of pine forest plantations located in Mpumalanga have a high susceptibility (> 70%) to *S. noctilio* infestation. A KHAT value of 0.84 and F measures above 0.87 indicate that the random forest algorithm is a robust classifier that produces accurate results. Additionally, the use of the backward variable selection method enabled the modelling process to be simplified, and identified the minimum number of explanatory variables that would offer the best discriminatory power and help in the empirical interpretation of the final model. Overall, the results show that pine forests that experience stress caused by evapotranspiration and evaporation followed by rainfalls, especially during the summer months, are susceptible to *S. noctilio* infestations.

Keywords: Random forest, *Sirex noctilio*, susceptibility, variable selection

7.1. Introduction

Reducing the impact of the siricid wasp, *Sirex noctilio* (Hymenoptera: Siricidae) is crucial for the future productivity and sustainability of commercial pine resources in South Africa. *S. noctilio* has caused extensive damage to pine forests located in KwaZulu-Natal and the Eastern Cape (Hurley et al., 2007; Ismail et al., 2007; Slippers, 2006). It is now a major concern that the wasp will spread further north, to the province of Mpumalanga, where the majority of the country's pine forests are located (DWAF, 2005). Detection and monitoring methods have been identified as important tools that provide forest managers with valuable information on the current location and extent of *S. noctilio* infestations (Carnegie, 2005; Haugen et al., 1990; Hurley et al., 2007; Ismail et al., 2007).

Researchers have recommended the combined use of aerial and field surveys (Carnegie, 2005; Haugen, 1990) or the use of multispectral remote sensing (Ismail et al., 2007; Ismail et al., 2008b) to spatially quantify the location and extent of *S. noctilio* infestations. Due to operational limitations (namely, cost and labour) and the initial scattered pattern of infestations (Ciesla, 2003), it is not feasible to consistently implement any of the suggested detection and monitoring methods at national or provincial scales. Therefore, the strength of current detection and monitoring methods would be greatly enhanced if pine forests that are highly susceptible to *S. noctilio* infestations could be proactively identified before any concerted monitoring and detection methods are implemented. Maps showing the distribution of susceptible forests will then serve as a spatial guide and allow forest managers to focus their existing detection and monitoring efforts on these key areas (hotspots). Additionally, forest managers will have the ability to adopt the most appropriate remediation measures (Carnegie, 2005; Haugen, 1990; Haugen and Underdown, 1990; Neumann and Minko, 1981; Spradberry and Kirk, 1978; Taylor, 1981; Tribe and Cillie, 2004) before the wasp can colonize these uninfected pine forests.

Statistical modelling approaches have been increasingly recognized as being important tools that improve our understanding of forest pests and pathogens. When used within a spatial framework, these models have the ability to identify areas that are highly susceptible to infestations (Candau and Fleming, 2005; Carnegie et al., 2006;

Guo et al., 2005; Kelly and Meentemeyer, 2002; Negron, 1998; Rosso and Hansen, 2003; van Staden et al., 2004). For example, Carnegie et al., (2006) developed a model based on climate matching in order to understand the potential global distribution of *S. noctilio*. However, the explanatory variables used in the CLIMEX model (http://www.hearne.com.au/climex/) were based on the wasps' endemic habitat conditions in Eurasia and northern Africa. These areas experience dry warm summers and cool moist winters (Carnegie, 2005), whereas, in contrast, *S. noctilio* has successfully established itself in the summer rainfall areas of South Africa (Hurley et al., 2007; Hurley et al., 2008). With the exception of the CLIMEX model, spatially based studies that empirically relate the potential distribution of *S. noctilio* infestations to a set of explanatory variables (for example, environmental data) are non-existent. Therefore, it would be beneficial for pest management to model pine forests that are highly susceptible to infestations at a more regional scale in an effort to understand localized variations of environmental conditions in relation to the distribution of *S. noctilio* infestations.

Machine learning techniques such as classification and regression trees or C&RT (Breiman et al., 1984) have been used to model the damage associated with forest pests and pathogens (Candau and Fleming, 2005; Kelly et al., 2007; Kelly and Meentemeyer, 2002; Rosso and Hansen, 2003). C&RT are non-parametric models that construct a set of decision rules by recursively splitting the response variable (for example, species data) into smaller homogenous groups, where each split is based on a single explanatory variable. The final output is a tree diagram with the terminal nodes of the tree indicating the final response (De'ath and Fabricius, 2000; Prasad et al., 2006; Vayssieres et al., 2000). C&RT are popular amongst researchers because the model has the following benefits: no advanced variable selection is required, no assumptions are made regarding the Gaussian relationship between response and explanatory variables, the results are easy to interpret due to the graphical nature of the tree,, the model can use a combination of categorical and continuous explanatory variables , the model has the ability to capture hierarchical and non-linear relationships, and finally the model provides insight into the spatial influence of the explanatory variables (De'ath and Fabricius, 2000; Kelly et al., 2007; Kelly and Meentemeyer, 2002; Prasad et al., 2006; Vayssieres et al., 2000). However, C&RT are very sensitive to small changes in the

training dataset and have been identified as being unstable classifiers that are prone to overfitting (Breiman, 1996). Researchers have suggested that by bootstrapping (Efron and Tibshirani, 1993) the original training dataset and then averaging the class predictions, C&RT can be stabilized (Archer and Kimes, 2008).

The Breiman-Cutler, random forest (RF) algorithm is an improvement of C&RT that includes bootstrap aggregation (bagging) and randomly selects a subset of explanatory variables to create an ensemble classifier that avoids overfitting and is successful in combining unstable learners like C&RT (Breiman, 2001). Notably, the RF algorithm has been exploited for the analysis of microarray data (Archer and Kimes, 2008; Diaz-Uriarte and Alvarez de Andres, 2006; Jiang et al., 2004; Strobl et al., 2007). In recent years, researchers have successfully applied the RF algorithm to a variety of spatial datasets. Within a spatial framework the RF algorithm has been used to map invasive plants (Lawrence et al., 2006), land cover (Gislason et al., 2006; Pal, 2005), tick-borne disease (Furlanello et al., 2003), climate change (Leng et al., 2007; Prasad et al., 2006), and habitat suitability (Garzon et al., 2006) The present study intends to expand the RF algorithm to susceptibility mapping.

The aim of this study was to model pine forests that are susceptible to *S. noctilio* infestations in an effort to enhance current initiatives on monitoring and detection. The RF algorithm is implemented within a spatial framework to determine which pine forests in an unaffected area (that is, Mpumalanga) are highly susceptible to *S. noctilio* infestations. It is assumed that if pine forests in Mpumalanga share similar environmental conditions with those areas with confirmed *S. noctilio* infestations in KwaZulu-Natal, they are relatively more likely to be susceptible to infestation. More specifically, the robustness of the RF algorithm is examined, firstly, in terms of its classification accuracy and secondly, for the empirical selection of explanatory variables. This study ultimately focuses on developing a geographic information systems (GIS) susceptibility model that could eventually be implemented for all pine forests located in South Africa. For the first time, this study introduces the RF algorithm for mapping pine forests that are susceptible to *S. noctilio* infestations. Although the RF algorithm is capable of carrying out regression as well as classification (Liaw and Wiener, 2002), this study will focus on classification trees (CT), since the response

variable in this study is binary and denotes the absence or presence of *S. noctilio* infestations.

7.2. Materials and methods

7.2.1. Response and explanatory variables

The robustness and accuracy of the RF algorithm was assessed by applying the algorithm to 1301, Sappi and Mondi, *Pinus patula* compartments located in the southern region of KwaZulu-Natal (Figure 7.1). These compartments were visually checked by experienced forestry personnel for the presence or absence over a period of three years of *S. noctilio* infestations (that is, from 2004 to 2006). Additionally, in compartments that were classified as being infested, a subset of infested trees was destructively sampled by foresters to verify the presence or absence of *S. noctilio* larvae. Of the 1301 response variables, 458 (35.20%) had *S. noctilio* infestations, and 843 (64.80%) had no *S. noctilio* infestations. The response variables were further divided into a training dataset for model development and a test dataset for independent accuracy assessments (Table 7.1). Additionally, class frequencies of absence or presence were approximately balanced in both the test and training datasets.

Table 7.1: Training and test datasets used in the study.

	Training dataset	**Test dataset**
Presence (Y)	321 (35.24%)	137 (35.13%)
Absence (N)	590 (64.76%)	253 (64.87%)
Total	911 (70.02%)	390 (29.98%)

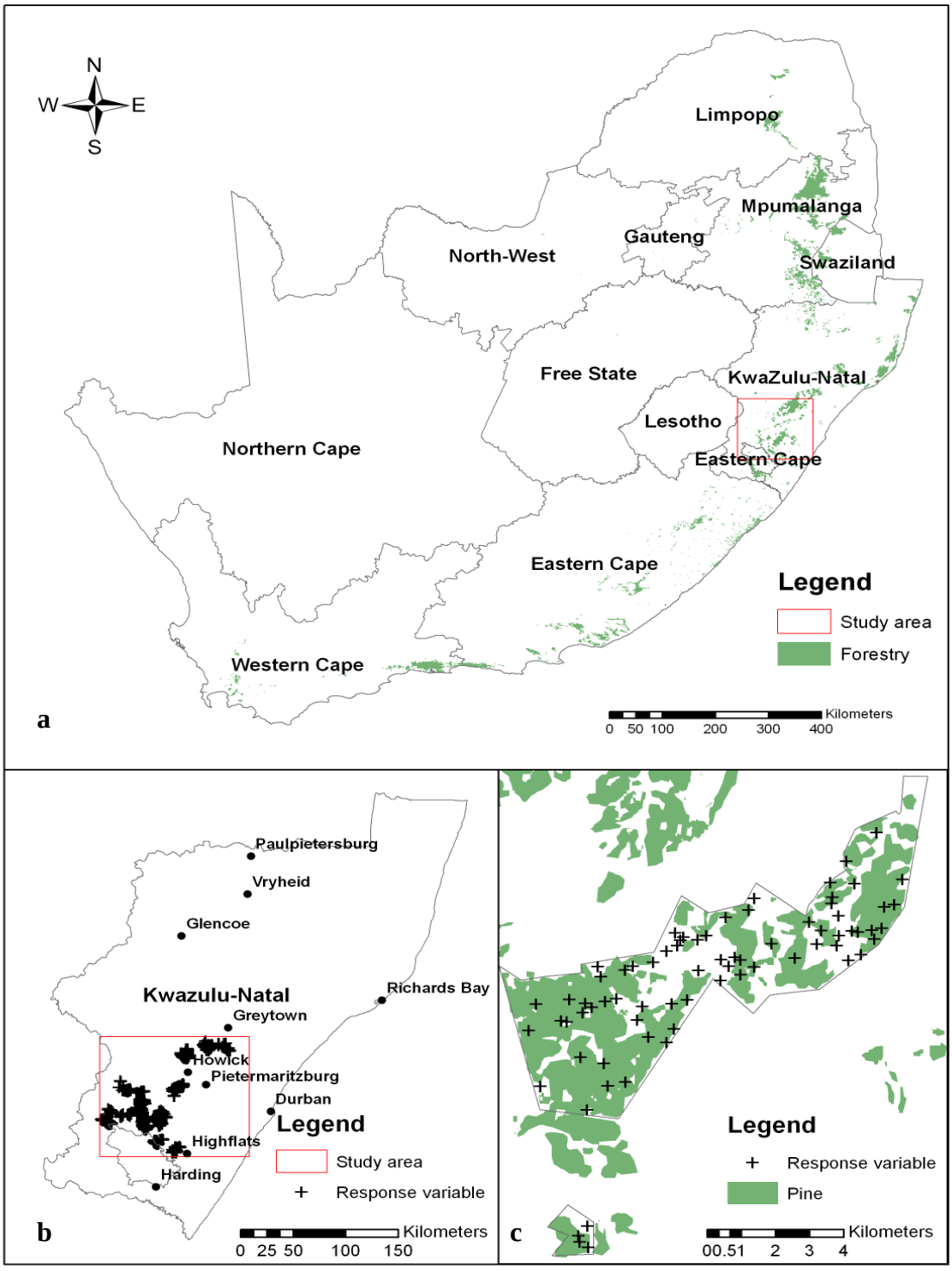


Figure 7.1: Inserts (a) and (b) show the study area in relation to the spatial distribution of commercial forestry plantations in South Africa. Insert (c) provides a detailed view of the samples that were collected at the Sappi Pinewoods plantation.

The explanatory variables used in this study consisted of one minute by one minute, historical climatic as well as topographic layers projected to Transverse Mercator (Hartebeesthoek datum: central meridian 31). The climatic variables used for developing the model were obtained from the South African agrohydrology atlas (Schulze et al., 1997) and included: mean annual precipitation, mean annual temperature, monthly median rainfall, monthly minimum temperature, monthly maximum temperature, monthly solar radiation, monthly evapotranspiration, and monthly potential evaporation. These historical climatic datasets (1990 to 1997) were derived from one thousand meteorological stations located across South Africa (van Staden et al., 2004) and a detailed methodological description of the datasets is provided by Schulze et al. (1997).

The topographic variables used in the study consisted of a digital elevation model (DEM) slope and aspect. The DEM (90 m spatial resolution) was derived from shuttle radar topographic mission (STRM) data and was obtained from the global landcover facility (GLCF) at the University of Maryland (http://glcf.umaics.umd.edu). Slope (percentage) and aspect (degrees) were then calculated from the DEM using Spatial Analyst (ESRI, 2006). Data from the climatic and topographic datasets (n = 77) were then extracted for the test and training datasets using the zonal statistics functionality in ArcGIS 9.1 (ESRI, 2006). The complete list of explanatory variables used in this study is shown in Table 7.2.

Table 7.2: Climatic and topographic datasets used in the study.

Variable	Abbreviation	Description	Coverage
Solar radiation	SR	Monthly solar radiation	January to December
Precipitation	MAP	Mean annual precipitation	
	MR	Median rainfall	January to December
Temperature	MAXT	Daily maximum temperature	January to December
	MINT	Daily minimum temperature	January to December
	MAT	Mean annual temperature	
Evaporation	APAN	Potential evaporation	January to December
	PEMO	Potential evapotranspiration	January to December
Digital elevation model	DEM	Elevation	
	SLOPE	Slope	
	ASPECT	Aspect	

7.2.2. Model description

7.2.2.1. The random forest algorithm

Firstly, the RF algorithm generates an ensemble of classification trees with each tree in the ensemble grown to maximum size without any pruning. The classification trees in the ensemble then vote by plurality on the correct classification. Secondly, the RF algorithm searches only across randomly selected explanatory variables (*mtry*) to determine the split at each node. Each tree in the ensemble is then constructed using a different bootstrapped sample (that is, with replacement) and contains randomly drawn samples from approximately two thirds of the samples from the original training dataset. The excluded 1/3 of the random samples which are left out from each bootstrapped sample, are known as the out of bag (OOB) samples. Finally, the OOB samples are then used to determine misclassification error and variable importance. The misclassification error or the OOB error is calculated by putting each OOB sample down the corresponding classification tree from which it was excluded. The error estimate is then calculated as the misclassified proportion of that OOB sample (Breiman, 2001; Garzon

et al., 2006; Liaw and Wiener, 2002; Pal, 2005; Peters et al., 2007; Prasad et al., 2006). A more detailed statistical description of the algorithm is provided by Breiman (2001). The *randomForest* library (Liaw and Wiener, 2002) developed for the R statistical software (R Development Core Team, 2008) was used to implement the RF algorithm. The calculation of the variable importance is described in the section below.

7.2.2.2. Using random forest for variable selection

The RF algorithm calculates the importance of each explanatory variable by random permutation of all values of the explanatory variables in the OOB sample. The number of votes for the correct class in the permutated data is subtracted from the number of correct votes in the original data which is then averaged over all trees in the forest. This represents the importance value for each variable and is the percentage increase in the misclassification rate as compared to the OOB rate of the non-permutated data (Prinzie and Van den Poel, 2008). As opposed to other methods of calculating variable importance (for example, the Gini index), the permutation method is regarded as being the most reliable measure for determining variable importance (Breiman, 2001). However, it is often difficult to set a cut-off value when there are many explanatory variables and when most of them have very similar importance measures (Jiang et al., 2004). Also, in order to simplify the modelling process, this study identified the smallest number of explanatory variables that offer the best discriminatory power and help in the empirical interpretation of the final *S. noctilio* susceptibility model. To address these issues, two variable selection methods were examined. These methods iteratively measured the importance of each explanatory variable (as determined by the RF algorithm) and then removed the less relevant explanatory variables. The backward variable selection method built multiple RF and after building each RF iteratively discarded those explanatory variables with the smallest variable importance as determined by the OOB error rate (Diaz-Uriarte and Alvarez de Andres, 2006). The recursive variable selection method is very similar to the backwards approach except that variable importance is recalculated for each RF that is built thus producing a new ranking of variables before the variables with the smallest importance are discarded (Jiang et al., 2004; Svetnik et al., 2003). We used the varSelRF library (Diaz-Uriarte

and Alvarez de Andres, 2006) for the R statistical software (R Development Core Team, 2008) to implement the recursive and the backward variable selection methods.

7.2.2.3. Accuracy assessments

It has been suggested that when using the RF algorithm there may be no need for cross validation or a separate test dataset to determine the misclassification error because the OOB error provides an unbiased estimate of error (Lawrence et al., 2006; Prasad et al., 2006; Prinzie and Van den Poel, 2008). However, according to Diaz-Uriarte and Alvarez de Andres (2006), and Granitto et al. (2006), using the OOB error to determine the misclassification error could result in a biased estimation of the error because the samples used to calculate the error are not independent of the model being evaluated. In the present study, the OOB error was used to fine-tune the user defined *mtry* parameter and for the empirical selection of explanatory variables (Diaz-Uriarte and Alvarez de Andres, 2006).

To avoid bias in the accuracy assessments, we used an independent test dataset (n = 390) to calculate the misclassification error (Reunanen, 2003), and the results were tabulated using a confusion matrix (Table 7.3). Several measures can be calculated from the confusion matrix (Fielding and Bell, 1997). However, the precision and recall measures were calculated because the main interest in this study was to correctly model presence rather than absence of the wasp (Peters *et al.*, 2007). From the confusion matrix (Table 7.3), precision (p) is calculated as the proportion of predicted presences that are observed to be present rather than absent and is defined as: $p = TP/(TP + FP)$. Recall (r) is calculated as the proportion of observed presences that were predicted correctly and is defined as: $r = TP/(TP + FN)$ (Peters et al., 2007).

Table 7.3: The confusion matrix used in the study.

	presence	**absence**	**row total**
presence	True Positive (TP)	False Positive (FP)	$TP + FP$
absence	False Negative (FN)	True Negative (TN)	$FN + TN$
column total	TP + FN	FP + TN	Total

The weighted F measure (van Rijisbergen, 1979) that combines precision and recall is stated as:

$$F_{\beta}(p,r) = (\beta^2 + 1)pr \Big/ \beta^2 p + r \quad (1)$$

β is the weighting factor that controls the relative importance of precision versus recall. If $\beta = 1$, precision and recall have equal importance; if $\beta = 0.5$, precision is twice as important as recall; and if $\beta = 2$ then recall is twice as important as precision. According to Peters et al. (2007), the magnitude of F varies from no predictive power (0) to perfect prediction (1). Furthermore, we calculated the *k* (KHAT) statistic to determine if the overall classification as determined by RF was better than if it was classified by a random classifier. KHAT values range from -1 to +1, and if the values are one or close to one then there is perfect agreement between the test and training datasets (Congalton and Green, 1999; Lillesand et al., 2004; Skidmore, 1999).

7.3. Results

7.3.1. Fine tuning the random forest algorithm

Before using the RF algorithm to model the potential distribution of pine forests that are susceptible to *S. noctilio* infestations, the effect that the number of randomly selected variables (*mtry*) had on the classification error was examined. According to Peters et al. (2007), reducing the *mtry* value decreases (i) the strength of individual trees which results in an increase in classification error, and (ii) the correlation between any two trees in the forest which results in a decrease in classification error. Therefore, the user defined *mtry* value has to be optimized in order to achieve a minimal classification error. Four different sized RF algorithms (*ntree*) were constructed for all possible unique values (n = 77) of *mtry*. The lowest OOB error was used to determine the optimal *mtry* value (Diaz-Uriarte and Alvarez de Andres, 2006; Granitto et al., 2006; Peters et al., 2007; Svetnik et al., 2003).

Table 7.4: Maximum and minimum OOB errors obtained using four different *ntree* values and all possible *mtry* values.

***ntree* value**	**100**	**200**	**500**	**1000**
Minimum OOB error	9.22%	8.89%	9.00%	9.22%
Optimal *mtry* value	2	3	6	11
OOB error (*mtry* = 3)	9.55%	8.89%	9.33%	9.44%

Table 7.4 shows that the lowest OOB error (8.89%) is obtained when 200 trees are built using an *mtry* value of three. Furthermore, using an *mtry* value of three for the other models (100, 500, and 1000) produced a negligible increase in OOB error (< 1%). Similarly, when classifying microarray data, Diaz-Uriarte and Alvarez de Andres (2006) showed that the OOB error rate is largely independent of *ntree* sizes even for *ntree* values ranging from 1,000 to 40,000 trees. Due to the low OOB error produced, an *mtry* value of 3 and an *ntree* value of 200 was used for all subsequent analyses.

7.3.2. Variable selection using backward and recursive approaches

As mentioned earlier, the RF algorithm estimates the importance of an explanatory variable by looking at how much the OOB error increases when the OOB data for that particular explanatory variable is permutated and others are not permutated. Figure 7.2 shows the mean decrease in accuracy of explanatory variables (n = 35) as determined by the OOB error. For visualization purposes, variables that have greater than 30% decrease in accuracy are shown in the Figure 7.2.

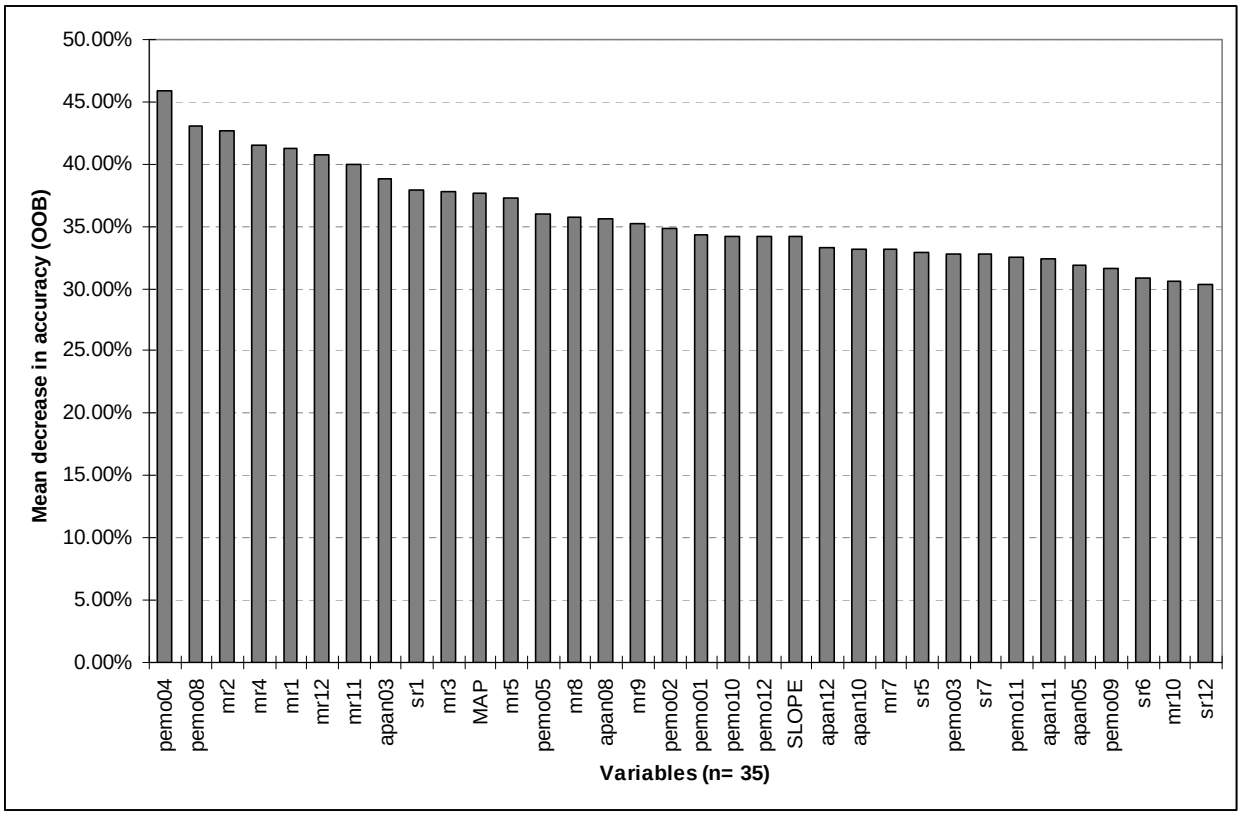


Figure 7.2: Variable importance as determined by random forest (*mtry* = 3 and *ntree* = 200). The full names for the variables are shown in Table 7.2.

Results show that the highest ranked variables in respect to their importance include: evapotranspiration (April and August), followed by the median rainfall during the summer months (February, April, January, November, and December). Additional high ranked variables include: solar radiation, evaporation and slope. Temperature (minimum and maximum), aspect and the digital elevation model have very low importance scores (not shown in Figure 7.2). To determine the minimum number of explanatory variables required to accurately model the potential distribution of *S. noctilio* infestations the backward and recursive variable selection methods were implemented. For both variable selection methods, 20% of the least important explanatory variables were discarded from the previous iteration. This allowed for faster computations and is based on an aggressive variable selection approach (Diaz-Uriarte and Alvarez de Andres, 2006). Figure 7.3 shows the results for both variable selection methods.

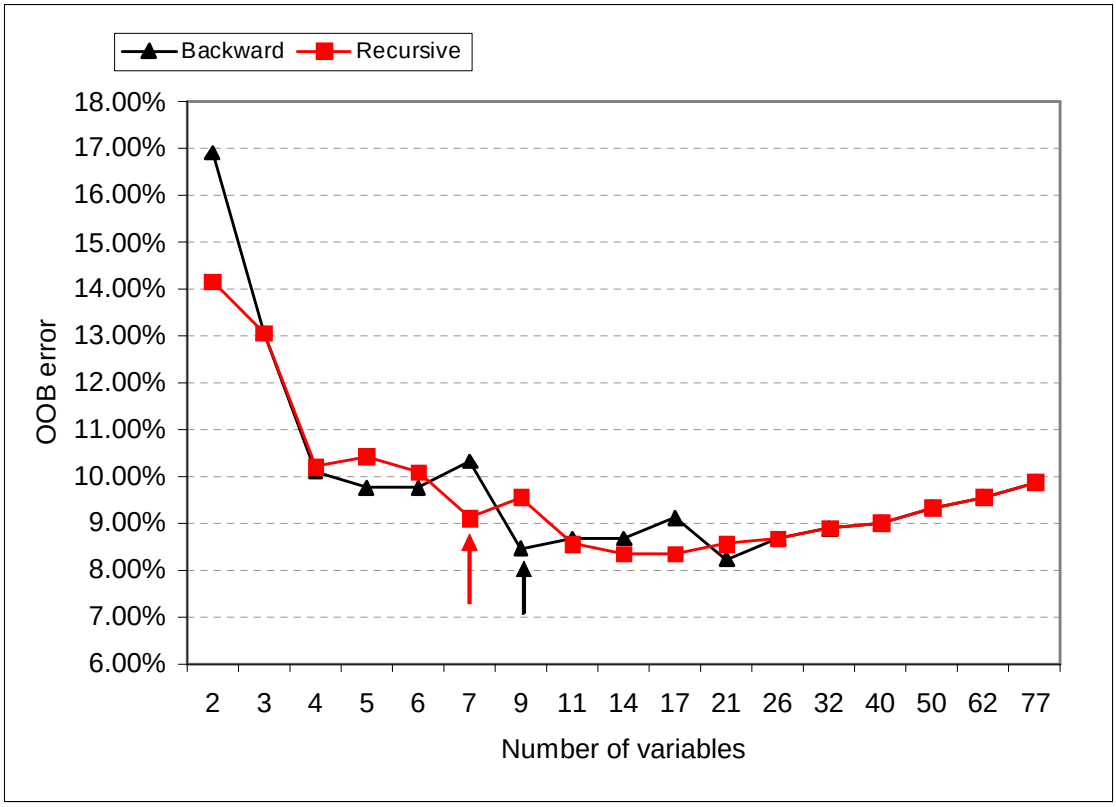


Figure 7.3: The OOB error obtained during the backward and recursive variable selection process. The arrows indicate the number of explanatory variable that produce an OOB error within one standard deviation of the lowest OOB error.

Results show that by using 14 variables the minimum OOB error obtained for the recursive variable selection method was 8.34% and by using 21 variables the minimum OOB error was 8.23% for the backward variable selection method. However, the best solution is based on selecting the least number of variables with the proviso that the final solution has an OOB error rate that is within one standard error of the minimum error rate of all forests created (Diaz-Uriarte and Alvarez de Andres, 2006). Under these conditions the recursive method then selected the best solution based on seven variables with an OOB error of 9.11%, while the backward method selected nine variables with an OOB error of 8.45%. The variables selected by the recursive method were as follows: median rainfall (January, February, April, and November), evapotranspiration (April and August), and potential evaporation (August). The variables selected by the backward method included: the mean annual precipitation, monthly median rainfall

(January, February, April, and December), monthly evapotranspiration (April, August, and October), and monthly potential evaporation (August). The backward variable selection method provides the better solution with a lower OOB error than the recursive approach.

7.3.3. Stability of the backward variable selection method

According to Granitto et al. (2006) the selection of explanatory variables is an unstable process and could lead to the selection of very different subsets of explanatory variables for each replicate of the study. To examine the stability of the model, it was determined the number of times an explanatory variable (MAP, MR1, MR2, MR4, MR12, PEMO4, PEMO8, PEMO10 and APAN08) is selected when the backward variable selection approach is bootstrapped (n = 100) (Efron and Thibshirani, 1997). Results indicate that all the variables selected using the backward method have a very high selection probability (90% and greater). As displayed in Figure 7.4, the explanatory variables with the highest probability of selection are mean annual precipitation (99%), April median rainfall (97%), August potential evaporation (97%), August evaporation (96%), followed by April potential evaporation (96%). The variables selected using the backward method (n = 9) were then used as the input explanatory variables for the final model.

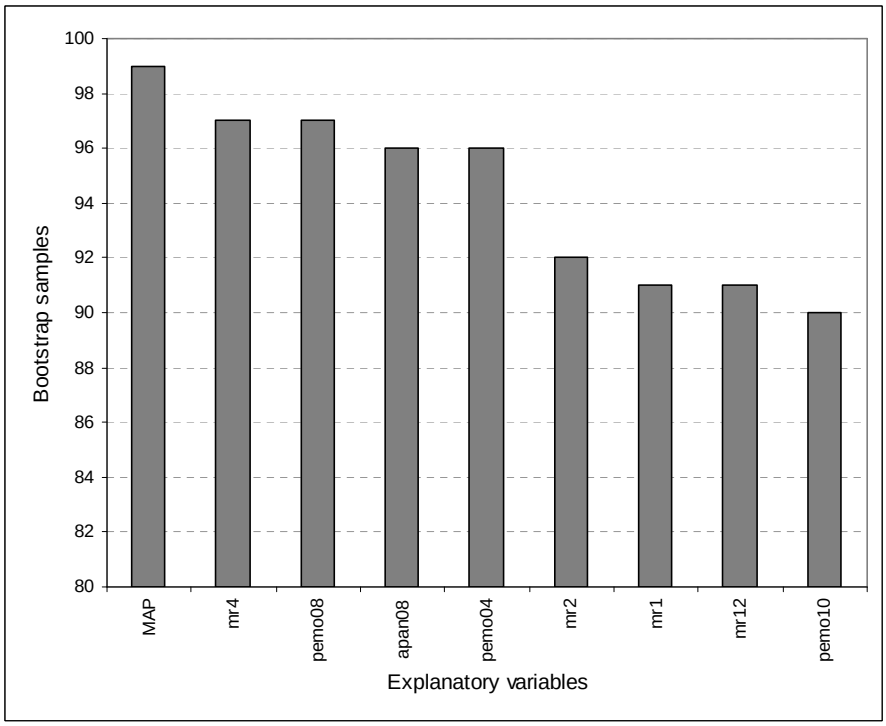


Figure 7.4: The number of times each explanatory variable is selected during the bootstrap process (n = 100). The full names for the variables are shown in Table 7.2.

7.3.4. Classification accuracy

To evaluate the accuracy and robustness of the RF algorithm for mapping the potential spatial distribution of *S. noctilio* infestations we compared the accuracy assessments of the RF algorithm against the widely used classification trees (CT) algorithm. Table 7.5 shows the accuracy assessments for both machine learning models. A KHAT value of 0.84 was obtained when using the RF algorithm, indicating that there is a strong agreement between the observations (n= 911) and the model predictions (n = 390). The KHAT value obtained by the RF algorithm is much higher than the KHAT value obtained by the CT (0.74) algorithm. For both algorithms, precision and recall are high, implying that there was more correctly predicted presence rather than absence of *S. noctilio* infestations. However, the weighted F measures ranged from 0.78 to 0.87 for CT algorithm, while the weighted F measures for the RF algorithm were all above 0.87. Overall, the RF algorithm produces better results than CT algorithm as determined by the weighted F measure as well as by the kappa analysis.

Table 7.5: Accuracy assessments using the test dataset (n = 390)

	Classification trees (CT)	**Random forest (RF)**
Precision	0.90	0.91
Recall	0.76	0.88
F_2	0.78	0.89
F_1	0.82	0.90
$F_{0.5}$	0.87	0.90
KHAT	0.74	0.84

7.3.5. Modeling Sirex noctilio susceptibility

Finally, we extrapolated the model developed for KwaZulu-Natal to all pine forest plantations located in Mpumalanga (Figure 7.5a). Each pixel (one minute by one minute) that contained pine forests was classified 200 times and the proportion of votes over all 200 trees indicated the susceptibility to *S. noctilio* infestations. Figure 7.5b shows the potential distribution of pine forest plantations that are susceptible to *S. noctilio* infestations in Mpumalanga as determined by the final model. Of the 1,909 pixels that were classified, 1,204 pixels showed that there was a high susceptibility (>70%) to *S. noctilio* infestation and the remaining pixels (n = 705) showed that there was a moderate (50% to 70%) to low (< 50%) susceptibility to *S. noctilio* infestation. Overall, the majority (63%) of pine forest plantations located in Mpumalanga showed a high susceptibility to *S. noctilio* infestation with the exception of pine plantations located in the vicinity of Lothair (Figure 7.5c), which showed a moderate to low susceptibility to *S. noctilio* infestation.

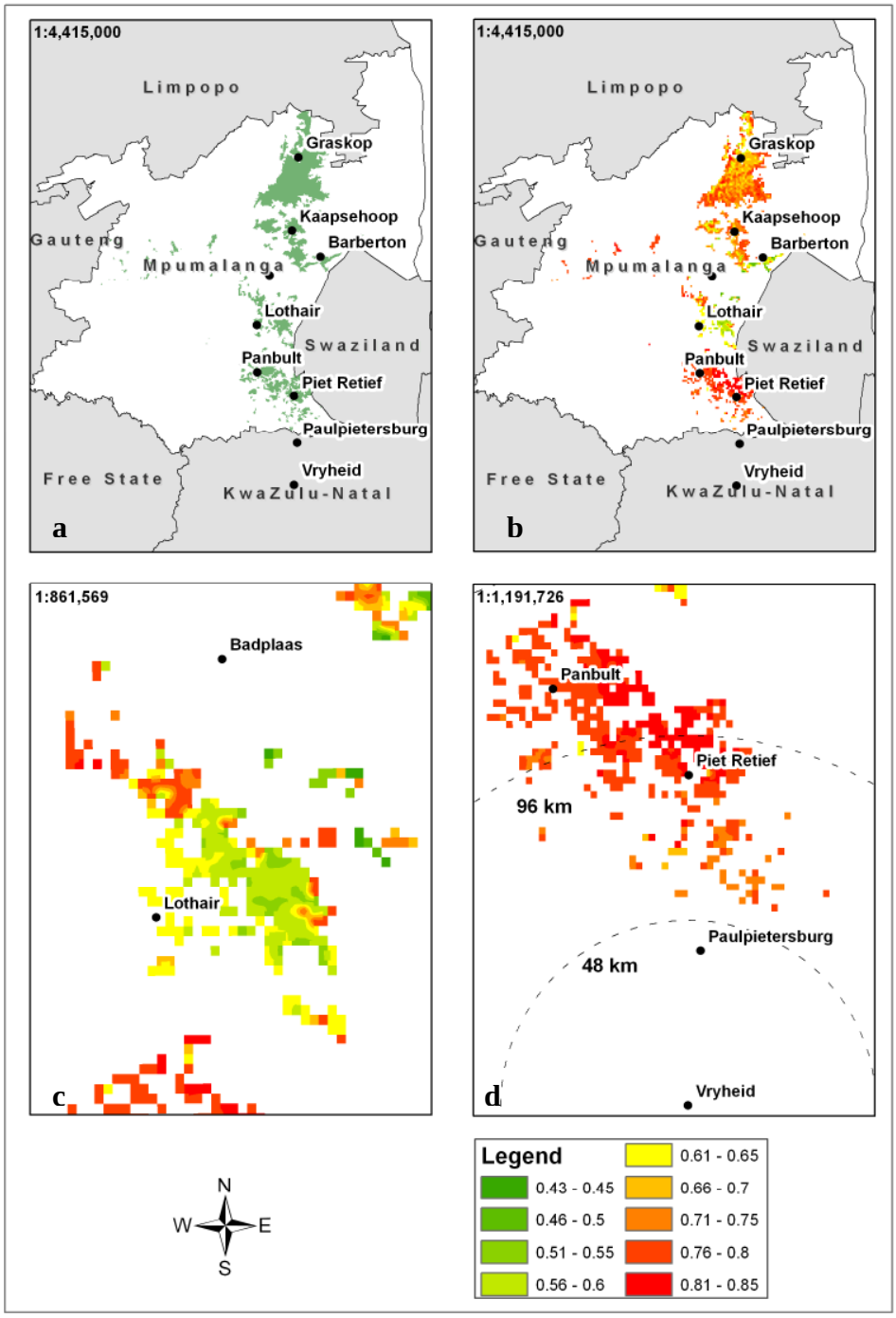


Figure 7.5: Insert (a) shows the current distribution of pine forests in Mpumalanga. Insert (b) shows the potential distribution of pine forests that are susceptible to *Sirex noctilio* infestations in Mpumalanga. Inserts 7.5 (c) and (d) provide a detailed view of pine forests that are susceptible to *Sirex noctilio* infestations. Insert 7.5 (d) shows the annual flight distance of the wasp by using 48 km radial buffers.

7.4 Discussion

7.4.1 Modeling susceptibility

Developing a model that spatially defines the potential distribution of those pine forests that are susceptible to *S. noctilio* infestations is an important step in understanding the nature of the epidemic in South Africa. *S. noctilio* is currently the most important pest of pines in South Africa (Hurley et al., 2008). Knowledge of the potential distribution of those pine forests that are susceptible to *S. noctilio* infestations is important because it serves as a spatial guide and allows forest managers to focus their existing detection and monitoring efforts on key areas and to proactively adopt the most appropriate course of intervention before the wasps colonize these unaffected pine forests in Mpumalanga. For example, results show that a potential 'hotspot' exists around the town of Piet Retief (Figure 7.5d). Pine forests located in the vicinity are highly susceptible to *S. noctilio* infestations and are within proximity to the current *S. noctilio* infestation in KwaZulu-Natal (Vryheid). With an annual flight radius of 48 km (Tribe and Cillie, 2004), the wasps will most probably colonize pine forests in the area within the next two years. It is recommended that pine forests located in the area should be continuously monitored for the early symptoms of *S. noctilio* infestation and prioritized for remediation efforts.

Remediation of established *S. noctilio* populations is achieved by biological means using the nematode *Beddingia siricidicola* and by using various parasitic wasps (Carnegie, 2005; Ciesla, 2003; Hurley et al., 2007; Tribe and Cillie, 2004). However, in unaffected pine forests in Mpumalanga, silvicultural practices, especially thinnings have been recommended to improve tree vigour and to increase resistance to future *S. noctilio* infestations (Hurley et al., 2007). However, it is important that thinnings are not carried out during the flight season as the practice could increase stress and favour a build-up of *S. noctilio* infestation (Carnegie, 2005).

7.4.2. Classification accuracy

The results obtained from this study are very encouraging and show that the random forest algorithm is a robust classifier that produces accurate results and better accuracies than a single classification tree. Using 20 machine learning datasets (Breiman, 2001) showed that the RF algorithm gave better predictive accuracies than C&RT. Similarly, Hamza and Larocque (2005) showed that the RF algorithm obtains the best overall classification results even when compared to other ensemble methods that use tree classifiers as the base model. The results obtained in this study are consistent with the above studies and additionally demonstrate the performance and accuracy (KHAT value of 0.84, and F measures above 0.87) of the RF algorithm when applied to spatial data. In addition to providing accurate classification results, the RF algorithm also provided insight into which covariates are most important in the modelling process.

7.4.3. Variable importance

The backward variable selection method successfully simplified the modelling process, and identified the smallest number of explanatory variables that offer the best discriminatory power and help in the empirical interpretation of the model. Variables that were selected included: MAP, MR1, MR2, MR4, MR12, PEMO4, PEMO8, PEMO10, and APAN08.

Although the RF algorithm and the backward variable selection process is not expected to describe the causal relationship between the explanatory variables and presence or absence of *S. noctilio* infestations, empirically derived results show that pine forests that experience stress caused by evapotranspiration and evaporation (PEMO4, PEMO8, PEMO10 and APAN08) followed by rainfalls especially during the summer months (MR1, MR2, MR4, MR12) are relatively more susceptible to *S. noctilio* infestations. The results are consistent with the view of Madden (1974) who hypothesized that intermittent stress (for example drought) contributes significantly to outbreaks by increasing tree attractiveness and susceptibility through rapid physiological changes following rains of short duration. It is well documented that trees that are experiencing stress are more likely to be attacked by *S. noctilio*. For example,

researchers have suggested that pine forests that experience drought, or fire, or have a high density of tree plantings are more likely to be attacked by *S. noctilio* (Ciesla, 2003; Haugen and Underdown, 1990; Tribe and Cillie, 2004).

7.5. Conclusion

The random forest algorithm when used in conjunction with GIS provides a useful and robust tool that can assist with current initiatives in forest pest management. The added benefit of using the random forest algorithm is that it requires the fine tuning of only one user defined parameter in order to achieve good classification. Overall, there is a high probability of *S. noctilio* infestation for the majority (63%) of pine forest plantations located in Mpumalanga. Compared to previous studies the final model identified highly susceptible pine forests at a more regional scale and provided an understanding of localized variations of environmental conditions in relation to the distribution of the wasps. Knowledge of the potential distribution of pine forests that are susceptible to *S. noctilio* infestations is important because it serves as a guide and allows forest managers to focus their existing detection and monitoring efforts on key areas and to proactively adopt the most appropriate course of intervention before the wasps colonize these as yet unaffected pine forests.

Acknowledgements

We thank Mondi and Sappi for allowing us access to field data. This project was partially funded by the National Research Foundation (South Africa). The early contributions of Marcel Verleur, Philip Croft, and Mark Norris-Rogers are gratefully acknowledged.

CHAPTER 8:

Remote sensing of forest health: A Synthesis

8.1. Introduction

Why do we need to utilize remote sensing technologies to detect and map *Sirex noctilio* infestations? *S. noctilio* is currently the most devastating pest of pines (Hurley et al., 2008) and poses a serious threat to the future sustainability of the South African forestry industry. Approximately 720,000 ha of pine forests in the country are potentially susceptible to *S. noctilio* infestations. Presently, 35,000 ha of pine forests are infested and dying in KwaZulu-Natal and the Eastern Cape (Hurley et al., 2007). Remote sensing has the ability to spatially quantify the severity and extent of these infestations, so that forest managers can adopt the most appropriate course of intervention before the forest reaches a point of non-recovery.

In retrospect, researchers have commented that the severe outbreak of *S. noctilio* in Australia and the failure of the pest management initiatives were primarily due to the following: (i) trees were experiencing high levels of stress due to overstocked stands and (ii) the detection and monitoring of the wasps were neglected (Haugen et al., 1990; Haugen and Underdown, 1990). As such, the surveillance of the wasps is a key component of management initiatives and remote sensing technology can significantly improve the current ability to detect and map *S. noctilio* infested pine forests relatively more accurately and effectively.

However, for remote sensing to be effective and accurate in detecting and mapping *S. noctilio* infestations, a sound understanding of the progression and pattern of symptoms of *S. noctilio* infestation across leaf, canopy, and landscape levels is required. So, the challenge in this thesis was to assess the observed symptoms of *S. noctilio* infestation and associate each level of observation with different remote sensing technologies in order to provide the appropriate level of detail and accuracy.

As a result, in this thesis the objectives were: (1) determining if vegetation indices derived from high spatial resolution digital multispectral imagery DMSI could characterize *S. noctilio* induced stress at a canopy level, (2) defining the appropriate spatial resolution that would allow for the accurate detection and mapping of *S. noctilio* infestations at a canopy level, (3) determining if high spectral resolution data would successfully discriminate between the healthy, green, and red stages of *S. noctilio* infestation, (4) determining if machine learning algorithms would accurately

discriminate between healthy trees and the green stage of *S. noctilio* infestations using resampled HYMAP data, (5) quantifying *S. noctilio* induced water stress in *Pinus patula* trees using regression tree ensembles and hyperspectral indices, and (6) modelling the potential distribution of pine forest that are susceptible to *S. noctilio* infestations. The findings pertaining to the objectives of this thesis are described in subsequent sections in this chapter.

8.2. The ability of high spatial resolution imagery to detect and map *Sirex noctilio* infestations

The ability of high spatial resolution imagery to detect and map *S. noctilio* infestations was addressed in Chapter 2. Results show that ratio (NDVI, RVI, DVI, and GNDVI) and linear (TCB and TCG) based indices derived from high spatial resolution DMSI (0.5 m x 0.5 m) have the ability to detect and map the healthy, red, and grey stages of *S. noctilio* infestations. However, discriminating the green stage of infestation from healthy trees remains elusive when using high spatial resolution DMSI (remote sensing of the green stage is adressed in Chapters 4, 5, and 6). Canonical variate analysis further revealed that when compared to the other ratio and linear based indices used in this study, NDVI has the best ability to detect and map *S. noctilio* infestations (Table 8.1).

Table 8.1: Factor structure matrix representing the correlation between variables and canonical functions (the healthy, red, and grey classes).

VI	Function 1	Function 2
NDVI	.633	.369
GNDVI	.629	.605
DVI	.559	.550
TCG	.500	.669
RVI	.484	.463
Eigenvalue	0.961	0.025
% Variance	97.5	2.5

Accuracy assessments indicated that a KHAT value of 0.79 is obtained when high spatial resolution DMSI is used to map the healthy, red, and grey stages of infestation. More noticeable was the user (70%) and producer (84%) accuracy values for the red stage of infestation. The high classification accuracy obtained for the red stage, emphasizes the importance of the visible and the near infrared (NIR) bands when

detecting and mapping declining forest health, especially when the damaging agent causes foliar discolouration. Overall, the study demonstrated that high spatial resolution imagery (i) is capable of replacing visual assessments of forest health, and (ii) provides a quantitative spatial framework for accurately detecting and mapping the visual stages of *S. noctilio* infestation. To date, this was the first study that investigated the ability of remote sensing technology to improve the ability to detect and map *S. noctilio* infested pine forests.

8.3. The appropriate spatial resolution to map *Sirex noctilio* infestations

In Chapter 3 the effects of spatial resolution on detecting and mapping red stage infestation levels were examined. By using the minimal variance of resampled normalized difference vegetation index (NDVI) images, the study defined the appropriate pixel sizes that captured the spatial variability (%) of red stage infestations. Results show that the appropriate pixel size should be chosen between the upper and lower limits proposed in Table 8.2.

Table 8.2: Appropriate spatial resolution for varying infestation levels (%).

Infestation Rate (%)	Resolution (Minimum Variance)	Infestation Level	Resolution (Sampling Theorem) (m)
2	3.50	Low	1.75
3	3.50	Low	1.75
4	3.50	Low	1.75
5	3.64	Low	1.82
6	3.50	Medium	1.75
7	3.64	Medium	1.82
8	3.64	Medium	1.82
9	3.92	Medium	1.96
10	3.86	Medium	1.93
11	3.98	High	1.99
12	3.95	High	1.98
13	4.18	High	2.09
14	4.37	High	2.19
15	4.61	High	2.31

A summary of, the major findings from the study are:

1. The method of calculating minimum variance for localized sub-samples is a useful and simple tool for identifying the appropriate spatial resolution needed for a particular investigation.
2. When using a spectral classifier (for example, NDVI) to detect infestation levels, pixel sizes larger than 2.3 m will not provide adequate information for high infestation levels (11% to 15%), and using pixel sizes smaller than the 1.75 m for detecting low to medium infestation levels (1% to 10%) could mean an unnecessarily large volume and cost of data.
3. Although the identified range of appropriate spatial resolutions is narrow (< 0.5 m), using the appropriate spatial resolutions as determined by this study would result in reduced costs of future image data acquisitions.
4. By establishing the spatial limitations of image data under variable infestation levels, the study provided the necessary guidelines for the operational detection and mapping of *S. noctilio* at compartment or plantation scales.

8.4. Testing the potential of hyperspectral data to detect *Sirex noctilio* infestations

Hyperspectral data (400 nm to 1300 nm) measured in a controlled laboratory environment can successfully discriminate between all stages of *S. noctilio* infestation. Results indicated that there is a significant difference ($p < 0.001$) between the mean spectral reflectance for the healthy, green and red stages of infestation, with large number of significant wavelengths located in the visible region of the electromagnetic spectrum (Figure 8.1). The majority of the significant bands (82.10%) are located in the visible part (300 nm to 700 nm) of the electromagnetic spectrum, with the remainder of the significant wavelengths located in the shoulder of the NIR.

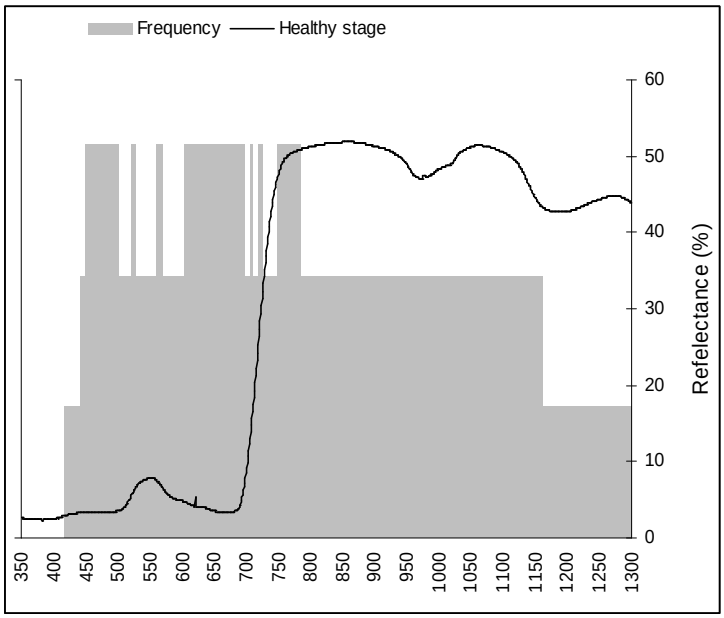


Figure 8.1: Frequency of statistically significant wavelengths ($p < 0.001$) for healthy, green, and red stages. The maximum grey shadings indicate the wavelengths that could discriminate between all class combinations of damaged trees. The spectral signature for the healthy needle is included for comparative purposes.

Due to the singularity problem of matrix inversion (Vaiphasa et al., 2005), the total number of significant bands (n = 190) had to be reduced prior to calculating the Jeffries–Matusita (J-M) distances (Richards and Jia, 1999). Using sensitivity and coefficient of variation analysis we selected bands located at 500 nm, 521 nm, 565 nm, 685 nm, 690 nm, 695 nm, 707 nm, 720 nm and 760 nm for further investigation. J-M distances were then calculated to determine the best combinations of these bands for separating the healthy, green, and red stages of infestation from each other.

Table 8.3: Results of the average Jeffries–Matusita distance analysis for all *S. noctilio* infestation classes (healthy, green, and red). The symbol (*) indicates the optimal bands that were selected in each band combination.

Best combination	**Wavelength(nm)**									**J-M value**	
	500	521	565	685	690	695	707	720	760		%
Single band				*						1.186	83.88
Two band							*	*		1.354	95.76
Three band	*	*		*						1.392	98.44
Four band	*	*		*					*	1.404	99.29
Five band	*	*		*	*				*	1.410	99.72
Six band	*	*	*	*	*				*	1.412	99.86
Seven band	*	*	*	*	*	*			*	1.413	99.93
Eight band	*	*	*	*	*		*	*	*	1.413	99.93
Nine band	*	*	*	*	*	*	*	*	*	1.413	99.93

Results showed that, although no single band or band combination is capable of total separability, the best average J-M values (1.413) is reached when using a seven band combination, with individual bands located at 500 nm, 521 nm, 565 nm, 685 nm, 690 nm, 695 nm and 760 nm (Table 8.3). However, using a four band combination, consisting of bands located at 500 nm, 521 nm, 685 nm and 760 nm produces an acceptable J-M separability (99.29 %) for all the stages of *S. noctilio* infestation. The results from this study provide the basis for future algorithms or spectral indices that can be used to detect and map *S. noctilio* attacked trees at airborne or spaceborne platforms.

8.5. Examining the ability of machine learning algorithms to detect *Sirex noctilio* infestations

The ability of machine learning algorithms to accurately discriminate between healthy trees and the green stage of *S. noctilio* infestations was investigated using resampled HYMAP data. The approach of using a wrapper (backward variable selection) in tandem with the random forest algorithm produces the smallest subset of wavelengths in the short wave infrared (SWIR) domain with the lowest misclassification error (Table 8.4). More specifically, HYMAP wavelengths located at 1990 nm, 2009 nm, 2028 nm,

2047 nm, and 2065 nm have the greatest potential for discriminating between healthy trees and the green stage of infestation. By using resampled HYMAP data instead of *in situ* hyperspectral measurements, the present study showed that it is possible to upscale the findings of this study to an airborne hyperspectral platform.

Table 8.4: The misclassification rate for random forest and boosting tree algorithms as determined by the .632+ bootstrap errors using the wavelengths selected by the analysis of variance (ANOVA), backward variable selection (BVS) and the out of bag (OOB) method (top 10% and top 20%).

Algorithm	All wavelengths	ANOVA	BVS	Top 10 %	Top 20 %
Number of wavelengths	64	54	5	6	13
Boosting trees	7.43	7.51	7.54	7.57	6.82
Random forest	7.29	7.45	**6.14**	6.52	6.60

Although researchers (Chan and Paelinckx, 2008) have previously used random forest as a wrapper for classification and variable selection purposes, this study additionally proved that the method works well in hyperspectral applications (i) where the number of samples are limited and (ii) where classes have similar spectral characteristics (that is, healthy and green stage of *S. noctilio* infestation). Additionally, comparisons between a competing machine algorithm, known as boosting trees revealed that (i) random forest produces a lower misclassification error and (ii) is relatively more robust to the introduction of noise. To summarize, the results of this study show that the random forest algorithm is a novel and robust method for the classification of hyperspectral data.

8.6. Predicting *Sirex noctilio* induced water stress

The objective of this chapter was to compare the performance of various regression tree ensembles for prediction purposes using hyperspectral remotely sensed data. More specifically, random forest, bagging and boosting ensembles were evaluated for predicting *S. noctilio* induced water stress in *P. patula* trees using several spectral parameters (n = 9) derived from hyperspectral data. Results from paired *t* test

indicated that there is a significant difference between random forest and boosting (t = 6.24, p < 0.001) and between the random forest and bagging ensembles (t = 8.68, p < 0.001). However, there was no significant difference (t = 2.23, p > 0.05) between the adjusted R^2 values of the boosting and bagging ensembles. In terms of an average adjusted R^2 value (Figure 8.2), the random forest ensemble produces the best overall performance (R^2 = 0.73).

Additionally, using the random forest ensemble as part of a wrapper, allowed the modelling process to be simplified, and identified the minimum number of spectral parameters that offer the best predictive accuracy. The backward elimination search function used only two spectral parameters while still producing an improved predictive accuracy (R^2 = 0.76). Results show that WI and the Ratio $_{975}$ indices have the best ability to assay the water content of *S. noctilio* infested trees thus making it possible to remotely quantify the severity of damage caused by the wasp.

While other remote sensing studies (Chan and Paelinckx, 2008; Gislason et al., 2006; Ham et al., 2005; Lawrence et al., 2006; Pal, 2005) have shown the value of using classification tree based ensembles, to the best of our knowledge, this was the first study to use regression trees as the base learner for bagging, boosting, and random forest ensembles.

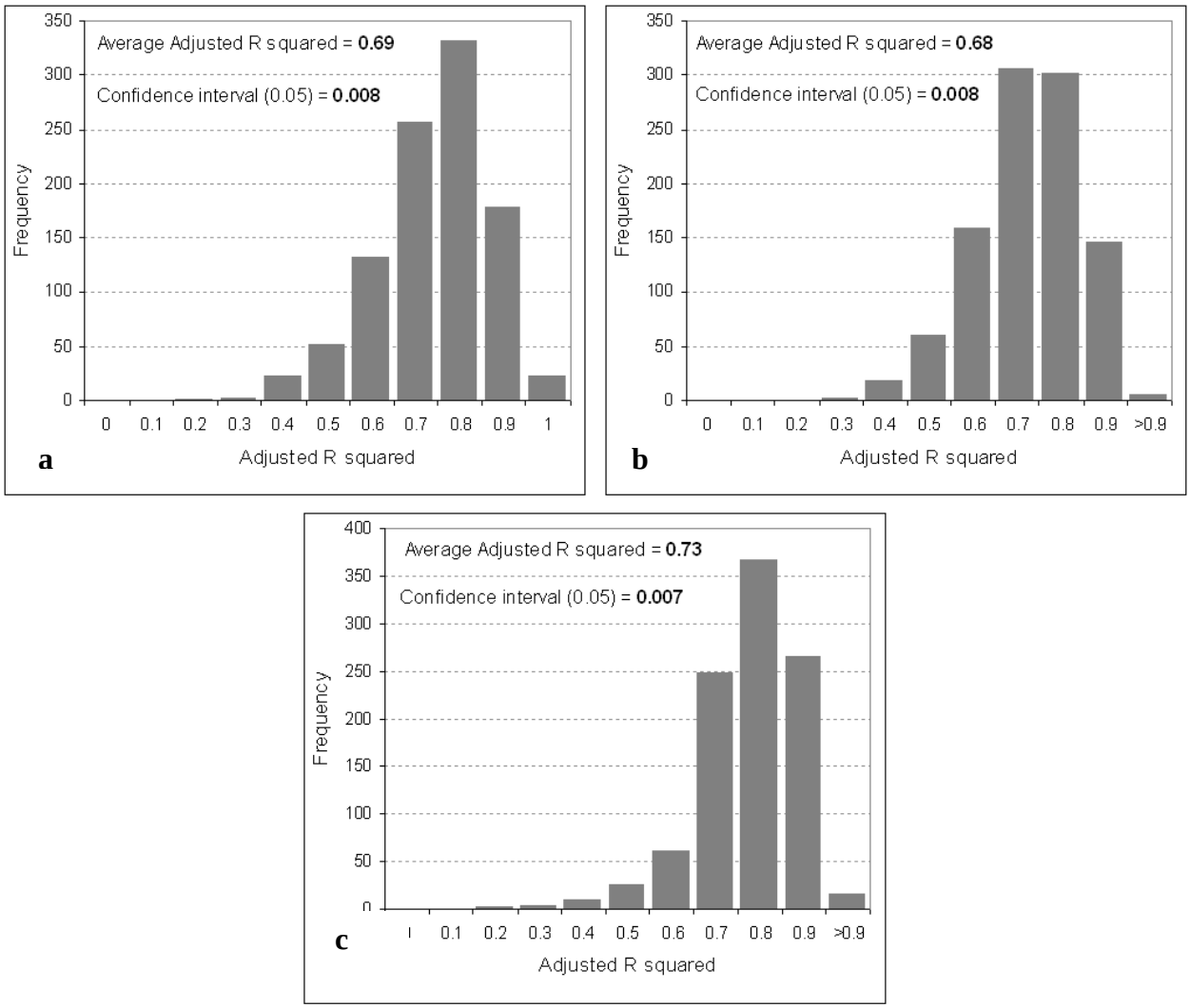


Figure 8.2: Histograms showing the frequency of the adjusted R^2 values for the regression tree ensembles used in this study. Insert (a) shows the distribution of the adjusted R^2 values for bagging ensembles, insert (b) shows the distribution of the adjusted R^2 values for boosting ensembles and insert (c) shows the distribution of the adjusted R^2 values for the random forest ensemble.

`

8.7. Modelling the potential distribution of pine forests that are susceptible to *Sirex noctilio* infestations

While previous chapters contained remote sensing solutions for the accurate detection and mapping of existing green and red stage *S. noctilio* infestations, the last aspect of this thesis deals with developing a GIS framework that would proactively identify pine forests that are highly susceptible to *S. noctilio* infestations. Results show that there is a high probability of *S. noctilio* infestation for the majority (63%) of pine forest plantations located in the province of Mpumalanga, South Africa (Figure 8.3). It is strongly recommended that these pine forests should be continuously monitored for the early symptoms of *S. noctilio* infestation and prioritized for remediation efforts.

Compared with previous models (Carnegie et al., 2006) the model developed in this thesis identified highly susceptible pine forests at a more regional scale and provided an understanding of localized variations of environmental conditions in relation to the potential distribution of the wasps. Environmental variables that offer the best explanatory power in the final model included: the mean annual precipitation, monthly median rainfall (January, February, April, and December) monthly evapotranspiration (April, August, and October) and monthly potential evaporation (August). Results are consistent with view of Madden (1974) and show that susceptibility of pine forest to *S. noctilio* infestation is dependent on stress followed by rainfalls especially in the summer months.

To summarize, the random forest model developed in this study provides a useful and robust tool that can assist with current initiatives for the management of forest pest.

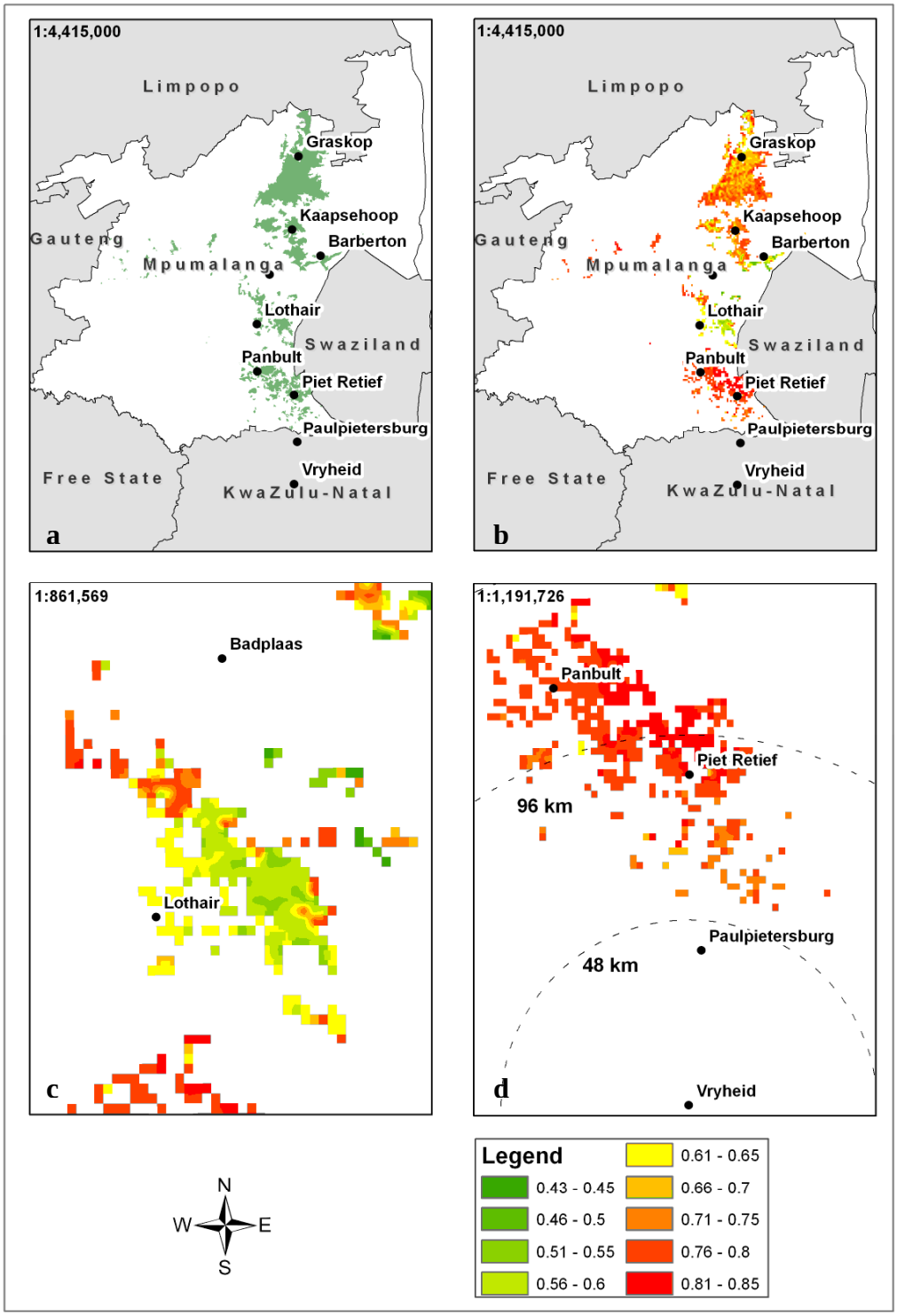


Figure 8.2: Insert (a) shows the current distribution of pine forests in Mpumalanga. Insert (b) shows the potential distribution of pine forests that are susceptible to *S. noctilio* infestations in Mpumalanga. Inserts (c) and (d) provide a detailed view of pine forests that are susceptible to *S. noctilio* infestations.

8.8. Conclusion

In this thesis the aim was to investigate the potential of remote sensing technologies to detect and map *S. noctilio* infestations. The research carried out in this thesis showed that it is possible to detect and map *S. noctilio* infestations using a combination of high spatial and spectral remote sensing technologies. The final conclusion was based on the following observations preserved in this thesis:

1. NDVI derived from high spatial resolution DMSI (0.5 m x 0.5 m) have the ability to detect and map the healthy, red, and grey stages of *S. noctilio* infestations. Accuracy assessments indicated that a KHAT value of 0.79 is obtained when high spatial resolution DMSI is used to map the healthy, red and grey stages of infestation.
2. A pixel size of 2.3 m should be used to detect high (11% to 15%) infestation levels, and a pixel size of 1.75 m should be used for detecting low to medium infestation levels (1% to 10%).
3. Hyperspectral data (400 nm to 1300 nm) measured in a controlled laboratory environment can successfully discriminate among the healthy, green, and red stages of *S. noctilio* infestation, with bands located at 500 nm, 521 nm, 685 nm and 760 nm producing an acceptable separability (99.29 %) for healthy, green, and red stages of infestation.
4. As determined by the random forest classifier, resampled HYMAP wavelengths located at 1990 nm, 2009 nm, 2028 nm, 2047 nm, and 2065 nm have the greatest potential for discriminating between healthy trees and the green stage of infestation.
5. WI and the Ratio $_{975}$ indices have the best ability to assay the water content of *S. noctilio* infested trees thus making it possible to remotely quantify the severity of damage caused by the wasps using the random forest ensemble.
6. The random forest model developed within a GIS framework showed that there is a high probability of *S. noctilio* infestation for the majority (63%) of pine forest plantations located in the province of Mpumalanga and the model provided an understanding of localized variations of environmental conditions in relation to the potential distribution of the wasps.

8.9. The future

The results from this study provide an alternative method for detecting and mapping forest health. In the future, the operational use of high spectral and spatial remote sensing technologies will improve the ability to detect and map *S. noctilio* infested pine forests. In line with the findings of this thesis, the detection and mapping of *S. noctilio* infestations is imperative for an effective pest management programme.

While findings of this thesis indicated that the red stage of infestation can be observed using remotely sensed data, forest managers should implement the detection and mapping of red stage *S. noctilio* infestations on a yearly basis. Temporal data on the severity and extent of infestations will provide useful information on the effectiveness of the current mitigation efforts as well as an indication of the spread of the wasps at a plantation level. The recent launch of the RapidEye satellite constellation provides an excellent source of affordable remotely sensed data that will allow for the continuous monitoring of the pest.

It is hoped that South Africa will hopefully launch the ZASat-003 satellite that will carry a hyperspectral sensor thus making high spectral resolution data more accessible and available to researchers in the country. The prospect of readily available airborne hyperspectral data is exciting and will provide researchers with an opportunity to upscale results that were developed in this thesis to an airborne platform. Thus the detection as well as the mapping of green stage *S. noctilio* infestation will become an operational reality.

Finally, the machine learning methods tested, applied, and discussed in this thesis present an alternative way of analysing remotely sensed data. For example, the random forest algorithm was shown to be a robust and accurate technique for classification as well as for regression applications. Opportunities exist to explore the usefulness of the method as a means to calculate an independent estimate of error without the need for a test dataset and to provide a variable selection method that will perform well in situations where the data may vary in scale or in the number of categories present.

References

Ahern, F., 1988. The effects of bark beetle stress on the foliar spectral reflectance of lodgepole pine. International Journal of Remote Sensing, 63: 61-72.

Ahmed, F. and Mthembu, I., 2006. Assessing the utility of HYPERION in extracting vegetation indices and leaf area index, AARSE 2006: Proceeding of the 6th AARSE international conference on earth observation and geoinformation sciences in support of Africa's development. The National Authority for Remote Sensing and Space Science (NARSS), Cairo, Egypt, pp. 8.

Analytical Spectral Devices, 2002. FieldSpec Pro users guide, Boulder, Colorado.

Anon, 2004. Sirex Woodwasp, FABI: Tree Protection Co-Operative Programme, Pretoria.

Archer, K.J. and Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Computational Statistics and Data Analysis, 52(4): 2249-2260.

Atkinson, P.M., 1993. The effect of spatial resolution on the experimental variogram of airborne MSS imagery. International Journal of Remote Sensing, 14: 1005-1011.

Atkinson, P.M., 1997. Selecting the spatial resolution of airborne MSS imagery for small scale agricultural mapping. International Journal of Remote Sensing, 18: 1903-1917.

Atkinson, P.M. and Aplin, P., 2004. Spatial variation in landcover and choice of spatial resolution for remote sensing. International Journal of Remote Sensing, 25(2): 3687-3702.

Bajcsy, P. and Groves, P., 2004. Methodology for hyperspectral band selection Photogrammetric Engineering and Remote Sensing, 70(7): 793-802.

Bian, L. and Butler, R., 1999. Comparing effects of aggregation methods on statistical and spatial properties of simulated data. Photogrammetric Engineering and Remote Sensing, 65: 73-84.

Breiman, L., 1996. Bagging predictors. Machine Learning, 26(2): 123-140.

Breiman, L., 2001. Random forests. Machine Learning, 45: 5-32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C., 1984. Classification and regression trees. Wadsworth and Brooks, Monterey, California.

Bruzzone, L. and Serpico, S.B., 2000. A technique for feature selection in a multiclass problem. International Journal of Remote Sensing, 21(3): 549-563.

Candau, J. and Fleming, R.A., 2005. Landscape-scale spatial distribution of spruce defoliation in relation to bioclimatic conditions. Canadian Journal of Forest Resources, 35: 2218-2232.

Carnegie, A., Matsuki, M., Haugen, D.A., Hurley, B., Ahumada, R., Klasmer, P., Sun, J. and Iede, E., 2006. Predicting the potential distribution of *Sirex noctilio* Fabricius (Hymenoptera: Siricidae), a significant exotic pest of *Pinus* plantations. Annals of Forest Science, 63: 119-128.

Carnegie, A.J., 2005. History and management of sirex wood wasp in pine plantations in New South Wales, Australia New Zealand Journal of Forestry, 35(1): 3-24.

Carter, G.A., 1994. Ratios of leaf reflectance in narrow wavebands as indicators of plant stress. International Journal of Remote Sensing, 15: 697-703.

Carter, G.A., Cibula, W.G. and Miller, R.L., 1996. Narrow-band reflectance imagery compared with thermal imagery for early detection of plant stress. Journal of Plant Physiology, 148: 515-522.

Carter, G.A. and Knapp, A.K., 2001. Leaf optical properties in higher plants: Linking spectral characteristics to stress and chlorophyll concentration. American Journal of Botany, 88: 677-684.

Chan, J.C. and Paelinckx, 2008. Evaluation of random forest and adaboost tree based ensembles classification and spectral band selection for ecotope mapping airborne hyperspectral imagery. Remote Sensing of the Environment, 112: 2999-3011.

Chen, D., Stow, D.A. and Gong, P., 2004. Examining the effect of spatial resolution and texture window size on classification accuracy: an urban environment case. International Journal of Remote Sensing, 25(11): 2177-2192.

Cibula, W.G. and Carter, G.A., 1992. Identification of a far-red reflectance response to ectomycorrhizae in slash pine. International Journal of Remote Sensing, 13(5): 925-932.

Ciesla, W.M., 2003. European woodwasp: a potential threat to North America's conifer forests. Journal of Forestry, 101(2): 18-23.

Clark, R.N. and Roush, T.L., 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. Journal of Geophysical Research, 89: 6329-6340.

Collins, J.B. and Woodcock, C., 1996. An assessment of several linear change detection techniques for mapping forest mortality using multitemporal Landsat TM data. Remote Sensing of Environment, 56: 66-77.

Colombo, S., Chica-Olmo, M., Abarca, F. and Eva, H., 2004. Variographic analysis of tropical forest cover from multi-scale remotely sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 58: 330-341.

Congalton, R.G. and Green, K., 1999. Assessing the accuracy of remotely sensed data : principles and practices. Mapping Sciences Series. Lewis Publishers, Boca Raton etc., 137 pp.

Coops, N., Dury, S., Smith, M.L., Martin, M. and Ollinger, S., 2002. Comparison of green leaf eucalypt spectra using spectral decomposition. Australian Journal of Botany, 50: 567-576.

Coops, N., Johnson, M., Wulder, M and J. White, 2006. Assessment of QuickBird high spatial resolution imagery to detect red-attack damage due to mountain pine beetle infestation. Remote Sensing of Environment, 103: 67-80.

Coops, N., Stone, C., Culvenor, D.S. and Chisholm, L., 2004. Assessment of crown condition in eucalypt vegetation by remotely sensed optical indices. Journal of Environmental Quality, 33: 956-964.

Coops, N.C., Stanford, M., Old, K., Dudzinski, M., Culvenor, D.S. and Stone, C., 2003. Assessment of *Dothistroma* needle blight of *Pinus radiata* using airborne hyperspectral imagery. Phytopathology, 33(12): 1524-1532.

Cosmopoulous, P. and King, D.J., 2004. Temporal analysis of forest structural condition at an acid mine site using multispectral digital camera imagery. International Journal of Remote Sensing, 25(12): 2259-2275.

Couts, M.P., 1970. The Physiological effects of the mucus secretion of *Sirex noctilio* on *Pinus Radiata*. Australian Forest Research, 4(4): 23-26.

Croft, P., 2006.Personal communication. Resource manager, Mondi Shanduka.

Culp, M., Johnson, K. and Michailides, G., 2006. ada: An R Package for Stochastic Boosting. Journal of Statistical Software, 17(2):1-27.

De'ath, G. and Fabricius, K.E., 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology, 81(11): 3178-3192.

DeFries, R.S. and Chan, J.C., 2000. Multiple criteria for evaluating machine learning algorithms for landcover classification from satellite data. Remote Sensing of the Environment, 74: 503-515.

DeFries, R.S., Hansen, M., Steininger, M., Dubayah, R., Sohlberg, R. and Townshend, J., 1997. Subpixel forest cover in Central Africa from multisensor, multitemporal data. Remote Sensing of Environment, 60: 228-246.

Diaz-Uriarte, R. and Alvarez de Andres, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(3): 1-13.

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision tree: Bagging, boosting and randomization. Machine Learning, 40(2): 1-19.

DWAF, 2005. Commercial timber resources and primary roundwood processing in South Africa 2003/2004, Department of Water Affairs and Forestry, Pretoria.

Efron, B. and Thibshirani, R.J., 1997. Improvements on cross-validation: the .632+ bootstrap method. Journal American Statistical Association, 92: 548-560.

Efron, B. and Tibshirani, R., 1993. An introduction to bootstrapping. Monographs on statistics and applied probability. Boca Raton, Florida. Chapman and Hall/CRC, New York, 436 pp.

Eitel, J.U.H., Gessler, P.E., Smith, A.M.S. and Robberecht, R., 2006. Suitability of existing and novel spectral indices to remotely detect water stress in *Populus* spp. Forest Ecology and Management, 229: 170-182.

Ekstrand, S., 1994. Assessment of forest damage with Landsat TM: correction for varying forest stand characteristics. Remote Sensing of the Environment, 47: 291-302.

Elith, J., Leathwick, J.R. and Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology, 77: 802-813.

Elvidge, C.D. and Chen, Z., 1995. Comparison of broad-band and narrow band red and near infrared vegetation indices. Remote Sensing of Environment, 54: 38 - 48.

Entcheva, P.K., Rock, B.N., Martin, M.E., Neefus, C.D., Irons, J.R., Middleton, E.M. and Albrechtova, J., 2004. Detection of initial damage in Norway spruce canopies using hyperspectral airborne data. International Journal of Remote Sensing, 25(24): 5557-5583.

Erdas, 2004. Erdas Imagine 8.7, Leica Geosystems GIS & Mapping., Atlanta , GA, USA.

ESRI, 2006. ArcGIS 9.1, Redlands, California.

Ferwerda, J.G., 2005. Charting to quality of forage: mapping and measuring the chemical composition of foliage using hyperspectral remote sensing, ITC, Wageningen, Enschede, 183 pp.

Fielding, A.H. and Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24: 38-49.

Freund, Y. and Shapiro, R.E., 1996. Experiments with a new boosting algorithm Machine learning proceedings of the thirteenth international conference. Morgan-Kaufman, San Francisco, California, pp. 148-156.

Friedman, J., 2002. Stochastic gradient boosting. Computational Statistics and Data Analysis, 38(4): 367-378.

Furlanello, C., Neteler, M., Merler, S., Menegon, S., Fontanari, S., Donini, A., Rizzoli, A. and Chemini, C., 2003. GIS and random forest predictor: integration in R for tick-borne disease risk assessment. In: K. Hornik, F. Leisch and A. Zeileis (Editors), Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003), Vienna, Austria.

Gao, B.C., 1996. NDWI- a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment, 58: 257-266.

Garrigues, S., Allard, D., Baret, F. and Weiss, M., 2006. Quantifying spatial heterogeneity at the landscape scale using variogram models. Remote Sensing of Environment, 103: 81-96.

Garzon, M.B., Blazek, R., Neteler, M., de Dios, R.S., Ollero, H.S. and Furlanello, C., 2006. Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula. Ecological Modelling, 197: 383-393.

Gilabert, M.A., Gonzalez-Piqueras, J., Garcia-Haro, F.J. and Melia, J., 2002. A generalized soil-adjusted vegetation index. Remote Sensing of the Environment, 82: 303-310.

Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R., 2006. Random Forests for land cover classification. Pattern Recognition Letters, 27: 294-300.

Gitelson, A. and Merzlyak, M., 1998. Remote sensing of chlorophyll concentration in higher plant leaves. Advances in Space Research, 22(5): 689-692.

Gong, P., Mahler, S., Biging, G. and Newburn, D., 2003. Vineyard identification in an oak woodland landscape with airborne digital camera. International Journal of Remote Sensing, 24(6): 1303-1315.

Gong, P., Pu, R. and Heald, R., 2002. Analysis of insitu hyperspectral data for nutrient estimation of giant sequoia. International Journal of Remote Sensing, 23(9): 1827-1850.

Granitto, P.M., Furlanello, C., Biasioli, F. and Gasperi, F., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems, 83: 83-90.

Guo, Q., Kelly, M. and Graham, C., 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling, 182: 75-90.

Guyon, I. and Elisseeff, A., 2003. An introduction to varaible and feature selection Journal of Machine Learning Research, 3: 1157-1182.

Haara, A. and Nevalainen, S., 2002. Detection of dead or defoliated spruces using digital aerial data. Forest Ecology and Management, 160: 97-107.

Ham, J., Chen, Y., Crawford, M. and Ghosh, J., 2005. Investigation of the Random Forest Framework for Classification of Hyperspectral Data. IEEE Transactions on Geoscience and Remote Sensing, 43(3).

Hamza, M. and Larocque, D., 2005. An empirical comparison of ensemble methods based on classification trees. Journal of Computation and Simulation, 75(8): 629-643.

Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Sohlberg, R., Dimiceli, C. and Carroll, M., 2002. Towards an operational MODIS continuous field of percent tree cover algorithm: examples using AVHRR and MODIS data. Remote Sensing of Environment, 83: 303-319.

Hardinsky, M.A., Klemas, V. and Smart, M., 1983. The influence of soil salinity, growth form, and leaf moisture on the spectral radiance of *Spartina alterniflora* canopies. Photogrammetric Engineering and Remote Sensing, 49: 77-83.

Hastie, T., Tibshirani, R. and Friedman, J., 2001. The elements of statistical learning : Data mining, inference and prediction Springer-Verlag, New York, 560 pp.

Haugen, D.A., 1990. Control procedures for *Sirex noctilio* in the Green Triangle : review from detection to severe outbreak (1977-1987). Australian Forestry, 53(1): 24-32.

Haugen, D.A., 2000. *Sirex noctilio*, USDA Forest Service.

Haugen, D.A., Bedding, R.A., Underdown, M.G. and Neumann, F.G., 1990. National strategy for control of *Sirex noctilio* in Australia. Australian Forest Grower, 13(2): 8.

Haugen, D.A. and Underdown, M.G., 1990. *Sirex noctilio* control program in response to the 1987 Green Triangle outbreak. Australian Forestry, 53: 33-40.

Healey, S., Cohen, W., Zhiqiang, Y. and Krankina, O., 2005. Comparison of tasseled cap based Landsat data structures for use in forest disturbance detection. Remote Sensing of Environment, 97: 301-310.

Hunt, E.R. and Rock, B.N., 1989. Detection of changes in leaf water content using near and middle infrared reflectances. Remote Sensing of Environment, 30: 43-54.

Hurley, B., Slippers, B. and Wingfield, J., 2007. A comparison of the control results for the alien invasive woodwasp, *Sirex noctilio*, in the southern hemisphere. Agricultural and Forest Entomology, DOI:10.1111/j.1461-9563.2007.00340.x.

Hurley, B.P., Slippers, B., Croft, P.K., Hatting, H.J., van der Linde, M., Morris, A.R., Dyer, C. and Wingfield, M.J., 2008. Factors influencing parasitism of *Sirex noctilio* (Hymenoptera: Siricidae) by the nematode *Deladenus siricidicola* (Nematoda:Neotylenchidae) in summer rainfall areas of South Africa. Biological Control, 45(3): 450-459.

Hyppanen, H., 1996. Spatial autocorrelation and optimal spatial-resolution of optical remote sensing data in Boreal forest environment. International Journal of Remote Sensing, 17: 3441-3452.

Ismail, R., Brink, A., Verleur, M. and Boreham, G., 2005. An integrated approach to managing *Sirex noctilio* infestations. The Leaflet (Sappi forest quarterly newsletter), 41: 7.

Ismail, R., Mutanga, O. and Ahmed, F., 2008a. Discriminating *Sirex noctilio* attack in pine forest plantations in South Africa using high spectral resolution data. In: M. Kalacska and A. Sanchez-Azofeifa (Editors), Hyperspectral Remote Sensing of Tropical and Sub-Tropical Forests. Taylor and Francis: CRC Press, pp. 350.

Ismail, R., Mutanga, O. and Bob , U., 2006. The use of high resolution airborne imagery for the detection of forest canopy damage by *Sirex noctilio*. In: P.A. Langin and

M.C. Antonides (Editors), Precision forestry in plantations, semi-natural areas and natural forest: proceedings of the international precision forestry symposium. Stellenbosch University, Stellenbosch University, South Africa, pp. 119-134.

Ismail, R., Mutanga, O. and Bob , U., 2007. Forest health and vitality: The detection and monitoring of *Pinus patula* trees infected by *Sirex noctilio* using digital multispectral imagery (DMSI). Southern Hemisphere Forestry Journal, 69(1): 39-47.

Ismail, R., Mutanga, O., Kumar, L. and Bob, U., 2008b. Determining the optimal resolution of remotely sensed data for the detection of *Sirex noctilio* infestations in *Pinus patula* plantations in KwaZulu-Natal,South Africa. The South African Geographical Journal, 90(1): 196-204.

Jackson, R., 1983. Spectral indices in n-space. Remote Sensing of the Environment, 13: 409-421.

Jackson, R.D. and Huete, A.R., 1991. Interpreting vegetation indices. Preventative Veterinary Medicine, 11: 185-200.

Jiang, H., Deng, Y., Chen, H., Tao, L., Sha, Q., Chen, J., Tsai, C. and Zhang, S., 2004. Joint analysis of two microarray gene-expression datasets to select lung adenocarcinoma marker genes. BMC Bioinformatics, 5(81).

Jin, S. and Sader, S., 2005. Comparison of time series tasseled cap wetness and the normalised moisture index in detecting forest disturbances. Remote Sensing of Environment, 94: 364-372.

Jordan, C.F., 1969. Derivation of leaf area index from quality of light on the forest floor. Ecology, 50: 663-666.

Kavzoglu, T. and Mather, P.M., 2002. The role of feature selection in artificial neural network applications. International Journal of Remote Sensing, 23(15): 2919-2937.

Kelly, M., Guo, Q., Liu, D. and Shaari, D., 2007. Modeling the risk for a new invasive forest disease in the United States: An evaluation of five environmental niche models. Computers, Environment and Urban Systems, 31(6): 689-710.

Kelly, M. and Meentemeyer, R.K., 2002. Landscape dynamics of the spread of Sudden Oak Death. Photogrammetric Engineering and Remote Sensing, 68(10): 1001-1009.

Kohavi, R. and John, G.H., 1997. Wrappers for feature subset selection. Artifical Intelligence Journal, 97(1-2): 273-324.

Kohavi, R., Sommerfield, D. and Dougherty, J., 1996. Data Mining using MLC++: A Machine Learning Library in C++, Tools with Artifical Intelligence. IEEE Computer Society Press, pp. 234-245.

Kokaly, R.F., 2001. Investigating a physical basis for spectroscopic estimates of leaf nitrogen concentration. Remote Sensing of Environment, 75(2): 153-161.

Kokaly, R.F. and Clark, R.N., 1999. Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. Remote Sensing of Environment, 67(3): 267-287.

Kumar, L., Schmidt, K.S., Dury, S. and Skidmore, A., 2001. Review of hyperspectral remote sensing and vegetation science. In: F. van der Meer and S.M. de Jong (Editors), Imaging spectrometry: basic principles and prospective applications. Kluwer Academic Press, Dordrecht, The Netherlands.

Langrebe, D., 2002. Hyperspectral imagae data analysis IEEE Signal Processing Magazine, 19(1): 17-28.

Lawrence, R.L., Bunn, A., Powell, S. and Zambon, M., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote Sensing of Environment Remote Sensing of Environment, 90: 331-336.

Lawrence, R.L. and Labus, M., 2003. Early detection of Douglas-fir beetle infestation with subcanopy resolution hyperspectral imagery. Western Journal of Applied Forestry, 18(3): 1-5.

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

Leckie, D.G., Cloney, E. and Joyce, S., 2005. Automated detection and mapping of crown discolouration caused by jack pine budworm with 2.5m resolution multispectral imagery. International Journal of Earth Observation and Geoinformation, 7: 61-77.

Leckie, D.G., Jay, C., Gougeon, F., Sturrock, R. and Paradinee, D., 2004. Detection and assessment of trees with *Phellinus weirii* (laminated root rot) using high resolution multi-spectral imagery. International Journal of Remote Sensing, 25(4): 793-818.

Leckie, D.G., Teillet, P.M., Ostaff, D.P. and Fedosjevs, G., 1988. Sensor band selection for detecting current defoliation caused the spruce budworm. Remote Sensing of Environment, 26(31-50).

Leckie, D.G., Yuan, X., Ostaff, D.P., Piene, H. and Maclean, D.A., 1992. Analysis of high resolution multispectral MEIS imagery for spruce budworm damage assessment on a single tree basis. Remote Sensing of Environment, 40: 125-136.

Leng, W., He, S.H., Bu, R., Dai, L., Hu, Y. and Wang, X., 2007. Predicting the distribution of suitable habitat for three larch species under climate warming in Northern China. Forest Ecology and Management, doi;10.1016/j.foreco.2007.08.031.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R News, 2/3: 18-22.

Lillesand, T., Kiefer, R. and Chipman, J., 2004. Remote sensing and image interpretation. John Wiley and Sons, New York, 763 pp.

Lobell, D.B., Oritiz-Monasterio, J.I., Asner, G.P., Naylor, R. and Falcon, W., 2007. Combining field surveys, remote sensing and regression trees to understand yield variations in an irrigated wheat landscape. Agronomy Journal, 97: 241-249.

Lucas, R., Rowlands, A., Niemann, O. and Merton, R.N., 2004. Hyperspectral sensors and applications. In: P.K. Varshney and M.J. Arora (Editors), Advanced image processing techniques for remotely sensed hyperspectral data. Springer-Verlag, Berlin, pp. 11-49.

Macfarlane, D., 2004. State of the environment report: Comrie-HCV areas & important rivers and streams, SAPPI, Pietermaritzburg.

Madden, J.L., 1974. Oviposition behaviour of the woodwasp. Australian Journal of Zoology, 22: 341-351.

Marceau, D.J., Gratton, D.J., Fournier, R.A. and Fortin, J., 1994. Remote sensing and the measurement of geographical entities in a forested environment.2. the optimal spatial resolution. Remote Sensing of Environment, 49: 105-117.

McConnell, T.J., Johnson, E.W. and Burns, B., 2000. A guide to conducting aerial sketchmapping surveys. FHTET 00-001., USDA Forest Service, Forest Health Technology Enterprise Team, Fort Collins,Colorado.

McGarigal, K., Cushman, S. and Stafford, S., 2000. Multivariate statistics for wildlife and ecology research. Springer, New York, 283 pp.

McGrew, J.C. and Monroe, C.B., 2000. An introduction to statistical problem solving in geography. McGraw-Hill, Boston, 110 pp.

Menges, C.H., Hill, G.J.E. and Ahmad, W., 2001. Use of airborne video data for the characterization of tropical savannas in north Australia: Optimal spatial resolution for remote sensing applications. International Journal of Remote Sensing, 22(5): 727-740.

Michaelson, J., Schimel, D.S., Friedl, M.A., Davis, F.W. and Dubayah, R.O., 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. Journal of Vegetation Science, 5: 673-696.

Molinaro, A.M., Simon, R. and Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21(15): 3301-3307.

Mutanga, O. and Skidmore, A., 2004a. Integrating imaging spectrometry and neural networks to map grass quality in the Kruger National Park, South Africa. Remote Sensing of Environment, 90(1): 104-115.

Mutanga, O. and Skidmore, A., 2005. Discriminating tropical grass canopies grown under different nitrogen treatments using spectra resampled to HYMAP. International Journal of Geoinformatics, 1(2): 21-32.

Mutanga, O. and Skidmore, A.K., 2003. Continuum - removed absorption features estimate tropical savanna grass quality in situ, 3rd EARSeL Workshop on Imaging spectroscopy, 13 - 16 May 2003. EARSeL, Hersching, Germany.

Mutanga, O. and Skidmore, A.K., 2004b. Integrating imaging spectroscopy and neural networks to map tropical grass quality in the Kruger National Park, South Africa. Remote Sensing of Environment, 90(1): 104-115.

Mutanga, O., Skidmore, A.K. and Prins, H.H.T., 2004. Predicting in situ grass quality in the Kruger National Park, south Africa, using continuum-removed absorption features. Remote Sensing of Environment, 89(3): 393-408.

Murwira, A., 2003. Scale matters, PhD Thesis, ITC, Wageningen, Enschede, 150 pp.

Negron, J.F., 1998. Probability of infestation and extent of mortality associated with Douglas-fir beetle in the Colorado Front Range. Forest Ecology and Management, 107: 71-85.

Nelson, R.F., 1983. Detecting forest canopy change due to insect activity using Landsat MSS. Photogrammetric Engineering and Remote Sensing, 49: 1303-1314.

Neumann, F.G. and Minko, G., 1981. The Sirex woodwasp in Australian radiata pine plantations. Australian Forestry, 44: 46-63.

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

Pal, M., 2007. Ensemble learning with decision tree for remote sensing classification. Proceedings of world academy of science, engineering and technology, 26: 735-737.

Pal, M. and Mather, M., 2004. Assessment of the effectiveness of support vector machines for hyperspectral data. Future Generation Computer Systems 20(7): 1215-1225.

Penuelas, J., Pinol, J., Ogaya, R. and Filella, I., 1997. Estimation of plant water concentration by the refelctance water index (R900/R970). International Journal of Remote Sensing, 18: 2869-2875.

Peters, A., Hothorn, T. and Lausen, B., 2002. ipred: Improved Predictors. R News, 2/2: 33-36.

Peters, J., De Baets, B., Verhoest, N., Samson, R., Degroeve, S., De Becker, P. and Huybrechts, W., 2007. Random forest as a tool for ecohydrological distribution modelling Ecological Modelling, 207: 304-318.

Pontius, J., Hallet, R. and Martin, M., 2005a. Assessing Hemlock decline using visible and near-infrared spectroscopy: Indices comparison and algorithm development. Applied Spectroscopy, 59(6): 836-843.

Pontius, J., Hallet, R. and Martin, M., 2005b. Using AVIRIS to assess hemlock abundance and early decline in Catskills, New York. Remote Sensing of Environment, 97: 163-173.

Pontius, J., Martin, M., Plourde, L. and Hallett, R., 2008. Ash decline assessment in emerald ash borer-infested regions: A test of tree-level, hyperspectral technologies. Remote Sensing of the Environment, 112(5): 2665-2676

Prasad, A., Iverson, L. and Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, 9: 181-199.

Price, K.P. and Jakubauskas, M.E., 1998. Spectral retrogression and insect damage in lodgepole pine successional forest. International Journal of Remote Sensing, 19(8): 1627-1632.

Prinzie, A. and Van den Poel, D., 2008. Random Forests for multiclass classification: Random MultiNomial Logit. Expert systems with Applications, 34(3): 1721-1732.

Pu, R., Ge, S., Kelly, N.M. and Gong, P., 2003. Spectral absorption features as indicators of water status in coast live oak (*Quercus agrifolia*) leaves. International Journal of Remote Sensing, 24: 1799-1810.

Quattrochi, D. and Goodchild, M., 1997. Scale in Remote Sensing and GIS. FL:CRC/Lewis Publishers, Boca Raton, 432 pp.

R Development Core Team 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Radeloff, V.C., Mladenoff, D.J. and Boyce, M.S., 1999. Detecting Jack pine budworm defoliation using spectral mixture analysis. Remote Sensing of the Environment, 69(2): 156-169.

Rahman, A.F., Gamon, J.A., Sims, D.A. and Schmidt, M., 2003. Optimal pixel size for hyperspectral studies of ecosystem function in southern California chaparral and grassland. Remote Sensing of Environment, 84: 192-207.

Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. Journal of Machine learning Research, 3: 1371-1382.

Richards, J.A. and Jia, X., 1999. Remote Sensing Digital Image Analysis. Springer, Berlin, 363 pp.

Rosso, P.H. and Hansen, M.E., 2003. Predicting Swiss needle cast disease distribution and severity in young Douglas-Fir plantations in coastal Oregon. Epidemiology, 93(7): 790-798.

Rouse, J.W., Haas, R.H., Schell, J.A. and Deering, D.W., 1973. Monitoring vegetation systems in the Great Plains with ERTS, Third ERTS Symposium. NASA SP-351, Goddard Space Flight Center, Washington D.C, pp. 309-317.

Runesson, U.T., 1991. Considerations for early remote detection of mountain pine beetle in green-foliaged lodgepole pine, University of British Columbia, British Columbia, 240 pp.

Ruth, B., Hoque, E., Weisel, B. and Hutzler, P., 1991. Reflectance and fluorescence parameters of needles of Norway spruce affected by forest decline. Remote Sensing of Environment, 38: 35-44.

Schmidt, K.S. and Skidmore, A.K., 2001. Exploring spectral discrimination of grass species in African rangelands. International Journal of Remote Sensing, 22(17): 3421 - 3434.

Scholes, B. and Annamalai, L., 2006. CSIR imaging expertise propels SA to a science high, Aerospace Science Scope, pp. 19-21.

Schulze, R.E., Maharaj, M., Lynch, S.D., Howe, B.J. and Melvil-Thomson, B., 1997. South African atlas of agrohydrology and climatology, Water Research Commission Report,TT82/96.

Sharma, R. and Murtha, P., 2001. Application of Landsat TM tasseled cap transformation in detection of mountain pine beetle infestations., Final Proceedings: 23rd Canadian Symposium of Remote Sensing. Canadian Aeronautics and Space Institute, Quebec, 9 pp. CD Proceedings.

Skakun, R.S., Wulder, M.A. and Franklin, S.E., 2003. Sensitivity of the thematic mapper enhanced wetness difference index to detect mountain pine beetle red-attack damage. Remote Sensing of the Environment, 86: 433-443.

Skidmore, A., 1999. Accuracy assessment of spatial information. In: A. Stein, F.v.d. Meer and B. Gorte (Editors), Spatial statistics for remote sensing. Kluwer Academic Publishers, Netherlands, pp. 197-209.

Slippers, B., 2006. The sirex epidemic in KwaZulu-Natal and its control. Wood and Timber Times Southern Africa, 31(7): 24-25.

Slippers, B., Coutinho, T.A., Wingfield, B.D. and Wingfield, M.J., 2003. A review of the genus Amylostereum and its association with woodwasps. South African Journal of Science (99): 70-74.

Soh, L.K., 1999. Segmentation of satellite imagery of natural scenes using data mining IEEE Transactions on Geoscience and Remote Sensing, 37(2): 1086-1099.

Spradberry, J.P. and Kirk, A., 1978. Aspects of the ecology of siricid woodwasps (Hymenoptera:Siricidae) in Europe, North Africa and Turkey with special reference to the biological control of *Sirex noctilio* F. in Australia. Bulletin of Entomological Resources, 68: 341-359.

Stimson, H., Breshears, D., Ustin, S. and Kefauver, S., 2005. Spectral sensing of foliar water conditions in two co-occurring conifer species: *Pinus edulis* and *Juniperus monosperma*. Remote Sensing of Environment, 96: 108-118.

Stone, C., Chisholm, L. and Coops, N., 2001. Spectral reflectance characteristics of eucalypt foliage damaged by insects. Australian Journal of Botany, 49: 687-698.

Stone, C., Chisholm, L. and McDonald, S., 2003. Spectral reflectance characteristics of *Pinus radiata* needles affected by *Dothistroma* needle blight. Canadian Journal of Botany, 81: 560-569.

Stone, C. and Coops, N.C., 2004. Assessment and monitoring of damage from insects in Australian eucalypt forests and commercial plantations. Australian Journal of Entomology, 43: 283-292.

Strobl, C., Boulesteix, A., Zeileis, A. and Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8: 25.

Sutton, C.D., 2005. Classification and regression trees, bagging and boosting. In: C.R. Rao, E.J. Wegman and J.L. Solka (Editors), Handbook of Statistics, Volume 24: Data Mining and Data Visualization. North-Holland Publishing Co, Amsterdam, The Netherlands, pp. 303-329.

Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R. and Feuston, B., 2003. Random Forest: A classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Science, 43(6): 1947-1958.

Svetnik , V., Liaw, A., Tong, C. and Wang, T., 2004. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules, Lecture Notes in Computer Science, pp. 334-343.

Taylor, K.L., 1981. The sirex woodwasp: Ecology and control of an introduced forest insect. In: R.L. Kitching and R.E. Jones (Editors), The Ecology of Pests - Some Australian Case Histories. CSIRO, Melbourne, Australia, pp. 231-248.

Thenkabail, P., Smith, R. and de Pauw, E., 2002. Evaluation of narrowband and broadband vegetation indices for determining optimal hyperspectral wavebands for agricultural crop characterization. Photogrammetric Engineering and Remote Sensing, 68(6): 607-621.

Treitz, P.M. and Howarth, P.J., 2000. High spatial resolution remote sensing data for forest ecosystem classification: an examination of spatial scale. Remote Sensing of Environment, 72: 268-289.

Tribe, G.D., 1995. The woodwasp *Sirex noctilio* Fabricius (Hymenoptera: Siricidae), a pest of *Pinus* species, now established in South Africa. African Entomology, 3: 215-217.

Tribe, G.D. and Cillie, J.J., 2004. The spread of Sirex *noctilio* Fabricius (Hymenoptera: Siricidae) in South African pine plantations and the introduction and establishment of its biological control agents. African Entomology, 12(1): 9-17.

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment, 8: 127 - 150.

Vaiphasa, C., Ongsomwang, S., Vaiphasa, T. and Skidmore, A.K., 2005. Tropical mangrove species discrimination using hyperspectral data: a laboratory study. Estuarine, Coastal, and Shelf Science, 65: 371-379.

Vaiphasa, C., Skidmore, A., de Boer, W. and Vaiphasa, T., 2007. A hyperspectral band selector for plant species discrimination. ISPRS Journal of Photogrammetry and Remote Sensing, 62: 225-235.

van Aardt, J. and Coppin, P., 2006. Current state and potential of the IS-HS project: Integration of insitu data and hyperspectral remote sensing for plant production modelling. In: Ackerman PA, Längin DW and A. MC (Editors), Precision Forestry in plantations, semi-natural and natural forests. Proceedings of the International Precision Forestry Symposium. Stellenbosch University, Stellenbosch, Stellenbosch University.

van Rijisbergen, C.J., 1979. Information retrieval. Butterworths, London.

van Staden, V., Erasmus, B.F.N., Roux, J., Wingfield, M.J. and van Jaarsveld, A.S., 2004. Modelling the spatial distribution of two important South African plantation forestry pathogens. Forest Ecology and Management, 187: 61-73.

Vayssieres, M.P., Plant, R.E. and Allen-Diaz, B.H., 2000. Classification trees: An alternative non parametric approach for predicting species distributions. Journal of Vegetation Science, 11: 679-694.

Vogelmann, J.E., 1990. Comparison between two vegetation indices for measuring different types of forest damage in the north-eastern United States. International Journal of Remote Sensing, 11: 2281-2297.

Wingfield, M.J., Roux, J., Coutinho, T., Govender, P. and Wingfield, B.D., 2001. Plantation disease and pest management in the next century. Southern African Forestry Journal, 190: 67-71.

Woodcock, C.E. and Strahler , A.H., 1987. The factor of scale in remote sensing. Remote Sensing of Environment, 21: 311-332.

Wulder, M., White, J. and Bentz., B., 2004a. Detection and mapping of mountain pine beetle red attack: matching information needs with appropriate remotely sensed

data, Proceedings from the Joint Annual Meeting of the Canadian Institute of Forestry and Society of American Foresters. Society of American Foresters, Edmonton, CA, pp. 1-17.

Wulder, M.A., 1998. Optical remote sensing techniques for the assessment of forest inventory and biophysical parameters. Progress in Physical Geography, 22: 449-476.

Wulder, M.A., Hall, R.J., Coops, N. and Franklin, S.E., 2004b. High spatial resolution remotely sensed data for ecosystems characterization. BioScience, 54(6): 511-521.

Yarbrough, L., Easson, G. and Kuszmaul, J., 2005. Tasseled cap coefficients for the QuickBird 2 sensor: multiple derivation techniques and comparisons, Pecora 16: Global Priorities in Land Remote Sensing, Sioux Falls, South Dakota.

Yuan, X., King, D.J. and Vlcek, 1991. Sugar maple decline assessment based on spectral and textural analysis of multispectral aerial videography. Remote Sensing of Environment, 37: 47-54.

Zhu, X. and Wu, X., 2004. Class noise vs attribute noise: A quantitative study. Artifical Intelligence Review, 22: 177-210.

Zwolinski, J.B., South, D.B. and Droomer, E.A.P., 1998. Pine mortality after planting on post-agricultural lands in South Africa. Silva fennica, 32(3): 271-280.