Field Spectroradiometer Data: Acquisition, Organisation, Processing and Analysis on the Example of New Zealand Native Plants

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

Master of Philosophy
in
Earth Science

at Massey University, Palmerston North,
New Zealand.

Andreas Hueni
2006

Abstract

The purpose of this research was to investigate the acquisition, storage, processing and analysis of hyperspectral data for vegetation applications on the example of New Zealand native plants. Data covering the spectral range 350nm-2500nm were collected with a portable spectroradiometer.

Hyperspectral data collection results in large datasets that need pre-processing before any analysis can be carried out. A review of the techniques used since the advent of hyperspectral field data showed the following general procedures were followed:

1. Removal of noisy or uncalibrated bands
2. Data smoothing
3. Reduction of dimensionality
4. Transformation into feature space
5. Analysis techniques

Steps 1 to 4 which are concerned with the pre-processing of data were found to be repetitive procedures and thus had a high potential for automation. The pre-processing had a major impact on the results gained in the analysis stage. Finding the ideal pre-processing parameters involved repeated processing of the data.

Hyperspectral field data should be stored in a structured way. The utilization of a relational database seemed a logical approach. A hierarchical data structure that reflected the real world and the setup of sampling campaigns was designed. This structure was transformed into a logical data model. Furthermore the database also held information needed for pre-processing and statistical analysis. This enabled the calculation of separability measurements such as the JM (Jeffries Matusita) distance or the application of discriminant analysis.

Software was written to provide a graphical user interface to the database and implement pre-processing and analysis functionality.

The acquisition, processing and analysis steps were applied to New Zealand native vegetation. A high degree of separability between species was achieved and using independent data a classification accuracy of 87.87% was reached. This outcome required smoothing, Hyperion synthesizing and principal components transformation to be applied to the data prior to the classification which used a generalized squared distance discriminant function.

The mixed signature problem was addressed in experiments under controlled laboratory conditions and revealed that certain combinations of plants could not be unmixed successfully while mixtures of vegetation and artificial materials resulted in very good abundance estimations.

The combination of a relational database with associated software for data processing was found to be highly efficient when dealing with hyperspectral field data.

Acknowledgements

I would firstly like to thank my supervisor Mike Tuohy for his time and thoughtful advice throughout the preparation of this thesis.

I would also like to thank Mike Hedley and Bambang H. Kusumo for their valuable input in terms of end user requirements for spectral processing software.

Table of Contents

Abstract iii

Acknowledgements v

1 Introduction 1

2 Literature Review 3

2.1 Hyperspectral Remote Sensing 3

2.2 Hyperspectral Sensors 4

2.2.1 Field Spectroradiometers 4

2.2.2 Airborne Hyperspectral Sensors 6

2.2.3 Spaceborne Hyperspectral Sensors 7

2.3 Hyperspectral Data 7

2.3.1 Overview and Principles 7

2.3.2 Data Processing 11

2.3.3 Analysis 16

2.4 Spectral Libraries and Spectral Databases 24

2.4.1 Spectral Libraries 24

2.4.2 Spectral Databases 25

2.5 Intermediate Conclusions 26

3 Methods 27

3.1 Acquisition and Storage of Field Data 27

3.1.1 Dataflow Overview 27

3.1.2 ASD FieldSpecPro 27

3.1.3 Study Sites 27

3.1.4 Structure of Field Data 28

3.1.5 Acquisition of Field Data 28

3.1.6 Species 29

3.2 Spectral Database 30

3.2.1 Spectral Database Model 30

3.2.2 Spectral Database Implementation 40

3.3 A Spectral Data Management and Processing Software 40

3.3.1 Programming Language, Libraries and Environment 41

3.3.2 Software Architecture 41

3.4 Data Processing 46

3.4.1 Waveband Filtering 46

3.4.2 Smoothing 46

3.4.3 Synthesizing of other Sensors 48

3.4.4 Derivative Calculation 52

3.4.5 Feature Space Transformation 53

3.5 Statistical Analysis 56

3.5.1 Classification 56

3.5.2 Discriminant Analysis 57

3.5.3 Separability Analysis 57

3.5.4 Most Discriminating Bands 57

3.6 Mixed Spectral Signatures 58

4 Results 63

4.1 Spectral Properties of New Zealand Native Plants 63

4.1.1 Smoothing 64

4.1.2 Sensor Synthesizing 68

4.1.3 Derivative Calculation 73

4.1.4 Feature Space Transformation 74

4.1.5 Statistical Analysis 81

4.2 Mixed Spectral Signatures 92

4.2.1 Paper/Plant Mixture 92

4.2.2 Paper/Plastic/Plant Mixture 93

4.2.3 Three plant mixture 95

4.2.4 Positional Dependence of Paper/Plastic Mixtures 97

4.2.5 Probe Rotation 98

5 Discussion 99

5.1 Collection of Spectral Data of New Zealand Native Plants 99

5.2 Spectral Databases 99

5.3 Spectral Processing Chain 99

5.4 Processing Speed of Smoothing Operations 100

5.5 Data Reduction 100

5.6 Discriminative Power of Feature Spaces 100

5.7 Discrimination and Classification 101

5.8 Principal Component Analysis 101

5.9 Linear Transformations 101

5.10 Most Discriminating Bands 102

5.11 Separability Analysis and Discriminant Analysis 102

5.12 Spectral Unmixing 102

5.13 Atmospheric Correction of Hyperion Imagery 103

6 Conclusion 105

7 Bibliography 107

8 Appendix 113

8.1 SpectraProc Graphical User Interface 113

List of Figures

Figure 1: Interaction between energy source, object and sensor 3

Figure 2: Examples of spectral signatures acquired in the preliminary stage of the project 4

Figure 3: Examples for spectral space and feature space (Data from a preliminary stage of this study) 8

Figure 4: Probability distributions in a 2d feature space (Richards, 1993) 9

Figure 5: An example of a data distribution in a 2d feature space, showing independent samples of a class and their mean 9

Figure 6: Two dimensional data with little correlation (a) and high correlation (b) (Richards, 1993) 10

Figure 7: Data reduced to mean values (left) and data including 2nd order statistics information and showing regression lines for each class (right) 10

Figure 8: An example of a spectrum showing water band noise in 3 wavelength ranges 11

Figure 9: An example of a mixed pixel (linear mixture model) 22

Figure 10: Dataflow and involved hardware 27

Figure 11: Hierarchical directory structure 28

Figure 12: Database model overview at entity level 31

Figure 13: ERD of the entities study, species, site and spectrum 31

Figure 14: ERD of the entities study and waveband_filter and waveband_filter_range 33

Figure 15: ERD of library, statistic, feature_space, sensor and associated entities 37

Figure 16: ERD of the entities species and mixture 40

Figure 17: File system interfaces 41

Figure 18: Non-MFC classes 43

Figure 19: Spectral data processing cascade 45

Figure 20: An example of pre and post filtering of noise bands 46

Figure 21: A smoothed signature of Pittosporum eugenoides before and after the removal of smoothing artefacts 48

Figure 22: Ratios for Landsat7 ETM+ band 1 49

Figure 23: Gaussian curve illustrating the FWHM measure 50

Figure 24: Sensor response functions for Hyperion sensor elements 8 and 9 and the FWHM of band 8 .. 52

Figure 25: Illustration of discrete reflectance values ρ and interpolated linear curves to form a continuous reflectance curve 54

Figure 26: General mixing setup 59

Figure 27: Mixture segments 59

Figure 28: Rotational positions of the bare fibre 62

Figure 29: Features of a vegetation curve 63

Figure 30: Mean Hyperion synthesized spectra of NZ native plants 64

Figure 31: Effects of variations of smoothing filter size and polynomial order on smoothed spectra 65

Figure 32: RMSE of raw minus smoothed spectra 66

Figure 33: Noise spectra of different Savitzky-Golay filter settings (raw minus filtered spectra) 67

Figure 34: Red-NIR region of the noise spectra after filtering with order 3 smoothing filters (raw minus filtered spectra) 67

Figure 35: Raw and Hyperion synthesized spectra of Pittosporum eugenioides 69

Figure 36: Raw and Hyperion synthesized in NIR and SWIR2 parts of the spectrum 69

Figure 37: Noise spectrum of Pittosporum eugenioides (raw minus Hyperion synthesized) 69

Figure 38: Raw and decimated by factor 10 and 5 spectra of Pittosporum eugenioides (offset for clarity) 70

Figure 39: Raw and decimated by factor 10 and 5 (NIR part of the spectrum) 70

Figure 40: Raw and decimated by factor 10 and 5 (SWIR2 part of the spectrum) 71

Figure 41: Noise spectrum of Pittosporum eugenioides (Raw minus decimated by factor 5) 71

Figure 42: Noise spectrum of Pittosporum eugenioides (Raw minus decimated by factor 10) 71

Figure 43: RMSE of Hyperion synthesizing and Decimation 10 and 5 72

Figure 44: Landsat7 ETM+ and Hyperion signatures 72

Figure 45: Derivatives based on different pre-processing and derivative calculations 73

Figure 46: DGVI regions overlaid with a typical plant spectrum (Pittosporum eugenioides) 74

Figure 47: Frequency of statistically significant differences of DGVIs and their dependence on pre-processing 76

Figure 48: Example of the discrimination of species by DGVIs (calculation based on Hyperion synthesized data) 77

Figure 49: Frequency of statistically significant differences of NDVIs and their dependence on pre-processing 78

Figure 50: Scree plots of eigenvalues (Hyperion synthesized and Decimation 5 data) 79

Figure 51: Histogram of statistically significant differences between species pairs for PC transformed Hyperion synthesized data 80

Figure 52: Average PC Factor Loadings for the first five components 80

Figure 53: PC factor loadings for PC1 and PC2. The mean reflectance of Pittosporum eugenioides is displayed to relate the factors to typical vegetation reflectance features 81

Figure 54: Graphical comparison of the number of bands with frequencies higher than the threshold (mean + standard deviation) per spectrum segment 89

Figure 55: Histogram of the statistically significant differences in reflectance calculated using raw data of all library relevant species. The mean reflectance of Pittosporum eugenioides is displayed to relate the frequency to typical vegetation reflectance features. 90

Figure 56: Histogram of the statistically significant differences in reflectance calculated using Hyperion synthesized data of all library relevant species. The mean reflectance of Pittosporum eugenioides is displayed to relate the frequency to typical vegetation reflectance features....90

Figure 57: Histogram of the statistically significant differences in the first derivative of reflectance calculated using Hyperion synthesized data of all library relevant species. The mean reflectance of Pittosporum eugenioides is displayed to relate the frequency to typical vegetation reflectance features 91

Figure 58: Spectral curves of mixtures of vegetation and paper 92

Figure 59: Scatterplot of reflectance values for simulated Landsat bands 1 vs 7 for mixtures of paper and vegetation 93

Figure 60: Spectral curves of mixtures of vegetation, paper and plastic. Endmembers are plotted in thick lines 94

Figure 61: Scatterplot of reflectances for simulated Landsat bands 1 vs 7 for mixtures of vegetation, paper and plastic. Endmembers are plotted in bigger symbols 94

Figure 62: Spectral curves for mixtures of three plants. Endmembers are plotted as dashed lines. 95

Figure 63: Scatterplot of reflectances for simulated Landsat bands 1 vs 7 for mixtures of 3 plants 96

Figure 64: Positional dependence for paper/plastic mixtures 97

Figure 65: First derivative of 50% paper/plastic mixtures 98

Figure 66: Probe Rotation on paper/plastic mixture 98

Figure 67: Spectral curves for the endmembers of the 3 plant mixing experiment 103

Figure 68: Screen capture of SpectraProc 113

List of Tables

Table 1: Widely used airborne hyperspectral sensor systems 7

Table 2: Collected species 29

Table 3: Short description of MFC derived classes 42

Table 4: Short description of non-MFC derived classes 43

Table 5: Mixtures of paper and kawakawa 59

Table 6: Mixtures of paper, plastic and kawakawa 60

Table 7: Mixtures of kawakawa, lemonwood and karaka 61

Table 8: Paper/plastic mixtures and positions 61

Table 9: Mean frequencies of statistically significant differences in species pairs for DGVIs calculated for differing pre-processing parameters 75

Table 10: NTBI and mean frequencies of statistically significant differences in species pairs 77

Table 11: First 18 components of the eigenanalysis of Hyperion-synthesized and Decimation by 5 data ... 78

Table 12: The 10 principal components with the highest significances (according to the Wilcoxon test) ordered by significance 79

Table 13: Classification results for calibration and independent datasets (accuracy in percentage) 82

Table 14: Error matrix for DGVIs of smoothed Hyperion synthesized data classified using the quadratic distance discriminant function 83

Table 15: Producer and user accuracy statistics for DGVIs of smoothed Hyperion synthesized data classified by the quadratic distance discriminant function (accuracy in percentage) 84

Table 16: Statistics of separability analysis 85

Table 17: JM distances (upper triangle) and B distances (lower triangle) between species in DGVI feature space 86

Table 18: Significance statistics 88

Table 19: Number of bands with frequencies higher than mean plus one standard deviation 88

Table 20: Unmixing results for vegetation/paper mixtures 93

Table 21: Unmixing results for vegetation/paper/plastic mixtures 95

Table 22: Unmixing results for 3 plant mixtures 96

1 Introduction

Spectroradiometry has become increasingly popular in the last few years. The technology has advantages over conventional techniques, allowing the non destructive sampling of objects and enabling users to gain critical information more quickly and cheaply. The operation of the equipment tends to be relatively easy and data are collected quickly. However, the interpretation of these data is not dealt with quite as easily. The main issue when dealing with hyperspectral data is their dimensionality. Hyperspectral data are more complex than previous multispectral data and different approaches for data handling and information extraction are needed (Vane and Goetz, 1988; Landgrebe, 1997).

The Institute of Natural Resources, Massey University, had acquired a spectroradiometer built by ASD (Analytical Spectral Devices) and a study utilizing this instrument was considered to be of interest.

The goals of this study were: Enhance the knowledge of the Institute in the field of hyperspectral remote sensing utilizing the recently acquired FieldSpecPro spectroradiometer; study the processes of field data acquisition, data processing and analysis; create a spectral database of New Zealand native vegetation; analyze the spectral separability of New Zealand native vegetation; investigate the problem of mixed signatures; suggest a basis for the classification of land cover using Hyperion data

While the main focus of this research was on hyperspectral data, the simulation of Landsat7 ETM+ was also undertaken, mainly to provide a basis for further investigation of the problem of atmospheric correction. Landsat7 imagery of New Zealand has been successfully corrected for atmospheric influences by Landcare Research, Palmerston North.

During the project, support was given to a Soil Science PhD study at Massey University and to a study on soils and pastures at Landcare Research, Palmerston North, in terms of sharing expertise, collecting data and subsequent processing. These collaborations led to further development of the database and processing requirements and widened the focus of this study to include data from soil and pasture studies. As a result of this, a section on correlation of spectral data with other physical properties was added to the literature review. It serves to complete the picture of the analysis that can be applied to hyperspectral data. The above mentioned collaborations also supported the hypothesis that tools for efficient data handling, organisation and processing were of high interest to scientists.

2 Literature Review

2.1 Hyperspectral Remote Sensing

Hyperspectral remote sensing is a relatively recent development based on the principles of spectroscopy. Spectroscopy which originated from the area of analytical chemistry is the study of the interaction between electromagnetic radiation and matter (Milton, 2001).

In order to gain spectral data from an object, its chemical bonds must be stimulated by external energy. In laboratory conditions, artificial energy sources are usually employed while field measurements mostly rely on the sun, although some technologies use artificial light sources.

Figure 1 illustrates the interaction between the energy source, object and sensor.

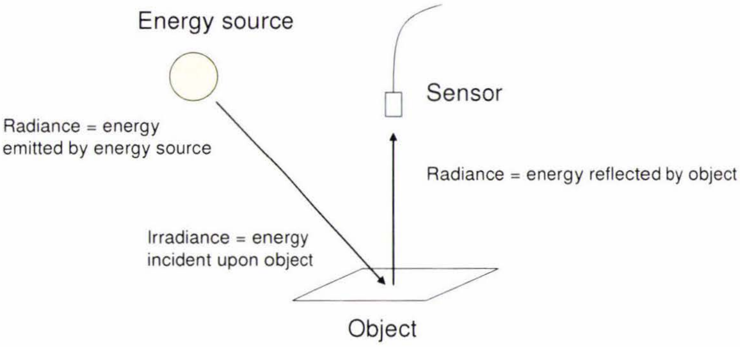

Energy source
Sensor
Radiance = energy emitted by energy source
Radiance = energy reflected by object
Irradiance = energy incident upon object
Object

Figure 1: Interaction between energy source, object and sensor

A range of instruments are used to capture spectral data. Photometers and radiometers are multiband instruments; the former are restricted to visible wavelengths only, whereas the latter make use of a wider range of wavelengths. The prefix 'spectro' designates instruments that are used to measure electromagnetic radiation in many narrow, contiguous wavebands, resulting in detailed, continuous spectra of the sampled objects (Milton, 2001).

The spectral range covered by spectroradiometers usually starts at blue, visible wavelengths (~400nm) and goes up to near infrared (~1000nm) or mid-infrared (~2500 nm). Thus, most of the reflectance data captured consist of responses at wavelengths that are not visible to the human eye.

It is expected that such detailed spectral data permit the identification of most surface types (Price, 1994). Figure 2 shows examples of rock, snow and vegetation spectra. Note that in the visible part of the vegetation spectrum (400-700 nm), green wavelengths (500-600nm) show a higher reflectance than blue (400-500 nm) and red (600-700nm). It is this local reflectance peak that lets humans perceive vegetation in shades of green.

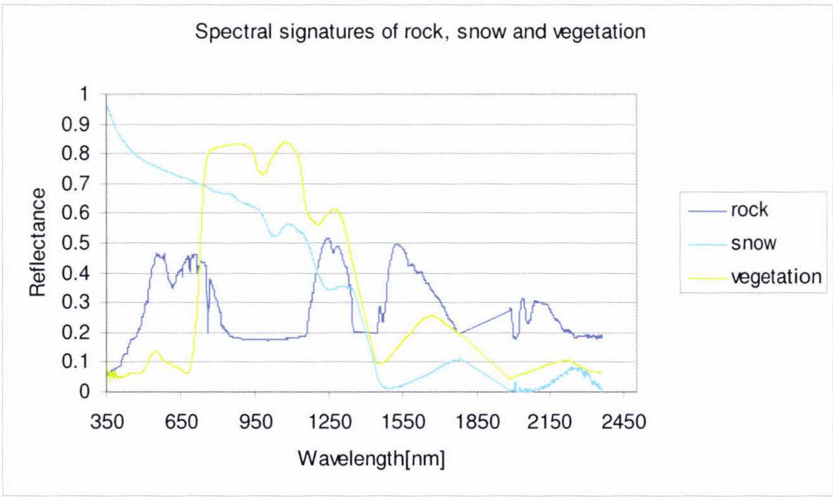


Figure 2: Examples of spectral signatures acquired in the preliminary stage of the project

In contrast to chemistry, remote sensing tends to concentrate on reflectance rather than absorbance. Reflectance is defined by:

$$\rho(\lambda) = \frac{\mathrm{E_r}(\lambda)}{\mathrm{E_i}(\lambda)} = \frac{\text{energy reflected from object at wavelength } \lambda}{\text{energy incident on object at wavelength } \lambda}$$

In order to convert measured radiance to reflectance, spectrometers must either be calibrated against a reflectance panel or directly measure the incident energy.

2.2 Hyperspectral Sensors

Four major groups of hyperspectral sensors are discernible:

1. Laboratory spectroradiometers
2. Field spectroradiometers
3. Airborne imaging spectroradiometers
4. Spaceborne imaging spectroradiometers

Laboratory spectroradiometers are not much used in remote sensing studies as field spectroradiometers can also be used indoors and usually offer all the needed data. Therefore laboratory spectroradiometers are not further discussed in this section.

2.2.1 Field Spectroradiometers

2.2.1.1 Overview

First field sensors emerged in the 1960's. They were usually modified laboratory instruments and had limited spectral coverage in the range of 400-1100nm.

Specifically developed, portable field spectroradiometers appeared in the late 1980's.

The PIDAS (Portable Instant Display and Analysis Spectrometer) instrument was completed in 1987. It sampled 833 bands in less than 2 seconds, covering a wavelength region from 450nm to 2500nm. Field and library spectra could be displayed simultaneously (Vane and Goetz, 1988).

The IRIS instrument covered wavelengths from 360 – 3000nm (Hutsinpiller, 1988).

For both of these instruments, spectral resolutions were not uniform over the whole spectral range.

Modern day field spectroradiometers still cover the same wavelength region as their predecessors. However, the spectral resolution tends to be uniform; around 1nm over the whole bandwidth and the integration times have increased by about factor 10. As well, the use of portable field computers with some instruments facilitates their operation and the subsequent data transfer to other systems.

Field spectroradiometers can be divided into two classes: single beam and dual beam instruments. Dual beam instruments are capable of measuring two energy sources simultaneously, i.e. incident energy and reflected energy can be recorded at the same time. This gives the dual beam instruments an advantage over single beam instruments, as the latter have to acquire these data consecutively, i.e. there is a time delay during which the incident energy level can change.

2.2.1.2 Acquisition of Field Data

As mentioned above, if no artificial light source is employed, field measurements rely on the illumination of the object by the sun. One of the problems posed is the rapidly changing light condition, even on clear days.

Milton (2001) lists three stages of illumination changes:

Cause	**Time period**	**Expected changes in irradiation**
Streams of atmospheric particulates	Few milliseconds	1%
Probably high altitude cirrus clouds (invisible to the human eye)	Seconds to minutes	5%
Visible clouds passing in front of the sun	NA	Major changes

The regular calibration of the instrument against a white reference is therefore of high importance if consistent readings are to be achieved. These references such as the Spectralon® (Labsphere Inc.) panels are assumed to have a Lambertian surface with a reflectance of 1, thus acting as ideal diffuse reflectors.

The above also implies that sampling should only be done on clear days to exclude the possibility of visible clouds changing the incident energy. High sun elevations are preferable due to the shorter path of the sun rays through the atmosphere, resulting in less atmospheric interference.

Consequently, field data collection usually happens in the summer months on cloud free days between 0900/1000 h and 1600 h (Hutsinpiller, 1988; Fyfe, 2003; Schmidt and Skidmore, 2003).

The field of view of spectroradiometers is around 20 degrees and less. Some instruments have inbuilt lenses while others use fibre-optics as an input device. There is however a constraint to the length of fibre optics which is currently 2-3 metres. Longer fibres result in loss of signal strength and are not employed. These technical issues have implications on the size of the sampled area.

Most studies report a nadir view of the optic and a distance of about 1–3 metres to the object, in some cases, step ladders, cherry pickers and helicopters have been used to raise the instrument into a suitable position (Thenkabail et al., 2000; Schmidt and Skidmore, 2003; Thenkabail et al., 2004a; Ramsey et al., 2005).

The field of view (FOV) is dependent on the type of foreoptics used and the distance to the target. The diameter of the FOV for a given FOV angle α and a height h above target is then calculated by:

$$d_{FOV} = 2 \cdot h \cdot \tan(\alpha)$$

2.2.1.3 Applications of Field Spectroradiometers

Milton (2001) differentiates between the following applications of field spectroradiometers:

1. As a remote sensing technique in its own right.

   Basic research and applied technology in areas like soil science, agriculture and horticulture.
2. In education and training.

   To teach the interaction of energy with matter and give an understanding of the principles of remote sensing.
3. Calibration of airborne and spaceborne sensors.

   The collection of ground truth data is important for the analysis of airborne and spaceborne hyperspectral data.
4. As a source of data for quantitative models and spectral libraries.

   The assembly of spectral data in libraries forms the base for physical and numerical models concerned with the interactions between electromagnetic radiation and matter.

2.2.2 Airborne Hyperspectral Sensors

The first airborne hyperspectral sensor AIS (Airborne Imaging Spectrometer) was first flown in 1982 (Vane and Goetz, 1988). The system collected data in 128 bands of 9.3nm width, covering a range from 400-1200nm in 'tree-mode' and 1200-2400nm in 'rock-mode'. It had a swath width of 32 pixels, every pixel covering approximately 8x8 metres of ground when flown at an altitude of 4200 metres (Lillesand et al., 2004). Being a prototype system, a series of problems were found such as: excessive electronic noise, non-uniformity of detector response, optical contamination due to vibrations, vertical and horizontal striping. The problematic issues found during the tests of AIS were addressed by AIS2 (Vane and Goetz, 1988). The AIS sensors led to the highly successful AVIRIS sensor generation, which is still being improved and used today (for technical specifications see Table 1).

Another widely used airborne sensor is the Australian developed HyMap (see specifications in Table 1). The sensor can be customized to suit demands of clients in terms of spectral coverage and number of bands. A new version of the system is being engineered offering an additional 32 bands in the thermal infrared (8-12 um) (Integrated Spectronics Pty Ltd).

Some of the widely used airborne hyperspectral sensor systems are listed in Table 1 (Olsen et al., 1997; Cocks et al., 1998; GER, 2000; Riedmann, 2003; Lillesand et al., 2004)

Table 1: Widely used airborne hyperspectral sensor systems

Name	Number of bands	Wavelength range	Bandwidth	Swath width	Comment
CASI 2	288	400-1000nm	1.8 nm	512 pixels	Fully programmable
AVIRIS	224	400-2450nm	9.6nm	614 pixels	
HYDICE	210	400-2500nm	10nm	208 pixels	
GER EPS-H	136 + 12	300-2500nm 8-12um	8-67nm	512-2048 pixels	Customisable system
HyMap	100-200	450-2500nm	10-20nm	60-70 degrees	Customisable system

2.2.3 Spaceborne Hyperspectral Sensors

There are currently two spaceborne hyperspectral sensors in orbit: Hyperion and CHRIS.

Hyperion is flown aboard the EO-1 satellite which was launched in late 2000. Hyperion collects 242 bands from 360-2600nm with bandwidths around 11nm. Some of these bands do not yield valuable data due to poor signal to noise ratios. The level 1 product subsequently contains only 198 calibrated bands. The spatial resolution is 30 metres at a swath width of 7.5 km.

This system is experimental and the data shows striping and other irregularities. Nonetheless Hyperion data has been used successfully in numerous hyperspectral studies.

CHRIS (Compact High Resolution Imaging Spectrometer) is carried on board the PROBA platform that was launched by ESA (European Space Agency) in October 2001. It samples a spectral range from 410-1050nm in 19 bands at a spatial resolution of 18 metres or in 63 bands at 36 metre resolution. The image area is 14km by 14km.

2.3 Hyperspectral Data

2.3.1 Overview and Principles

The main issues when dealing with hyperspectral data are their dimensionality and storage profile. The physical data size is an especially important issue with imaging spectrometers.

The dimensionality of the data is the result of sampling a wide spectral range in very narrow bands. This is in itself a problem because the influence of noise on narrow channels is much higher than on traditional broadband channels.

Hyperspectral data are more complex than previous multispectral data and different approaches for data handling and information extraction are needed (Vane and Goetz, 1988; Landgrebe, 1997).

2.3.1.1 Spectral Space and Feature Space Concept

Landgrebe (1997) based his work on hyperspectral data analysis on the signal theory and the principles of signal processing.

Hyperspectral data can be represented in three principal ways:

1. Image Space: data are shown as a 2 dimensional raster image. This applies only for imaging spectrometer data where every spectrum has a spatial location.
2. Spectral Space: the data are shown as spectra, i.e. as the reflectance response per wavelength
3. Feature Space: the data consist of vectors, which define points in an N-dimensional space

Figure 3 illustrates the concepts of spectral space and feature space for the example of three different spectra. Spectral space shows their reflectance values. In feature space, three classes are shown, defined by vector positions in a 2 dimensional space.

In the given example, the feature space was formed by choosing a subset of 2 components out of the possible N components that make up the signal vectors. The vector components are equivalent to the reflectance values at wavelengths 600nm and 1000nm respectively:

$$\vec{x}_i = \begin{vmatrix} \rho_i(600) \\ \rho_i(1000) \end{vmatrix}$$

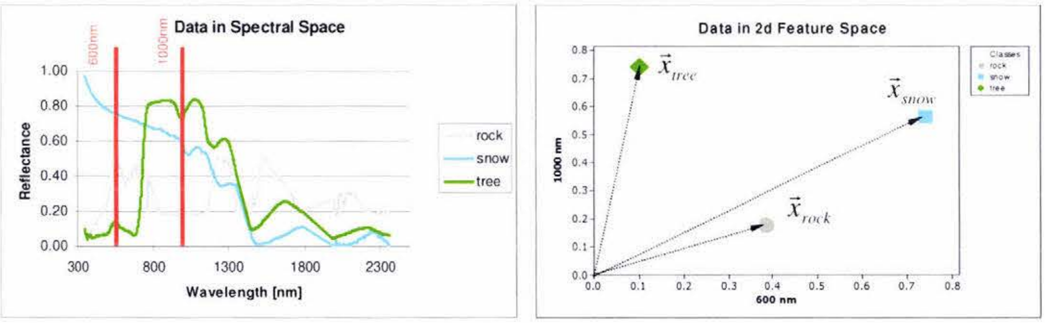


Figure 3: Examples for spectral space and feature space (Data from a preliminary stage of this study)

2.3.1.2 Data Distributions

Objects of the same type have reflectance vectors that lie close to each other in feature space. Object types are usually referred to as classes, e.g. snow, vegetation and rock. Vectors of class objects form clusters in feature space.

Classes in remote sensing applications are assumed to be of Gaussian distribution. An illustration of such distributions is given in Figure 4.

The mean position and distribution (shape) are defined by the mean vector and covariance matrix respectively. The covariance is one of the most important mathematical concepts in the analysis of multispectral (and hyperspectral) remote sensing data (Richards, 1993).

It must be noted that for the sake of visualization only 2 dimensional examples are shown. Real data distributions will have many more dimensions.

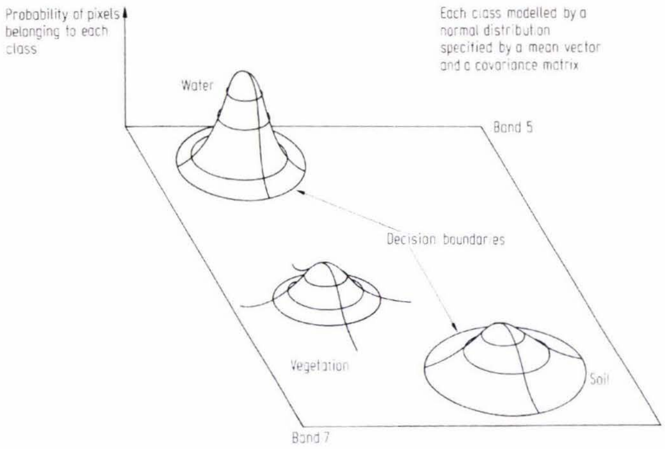

Probability of pixels belonging to each class
Each class modelled by a normal distribution specified by a mean vector and a covariance matrix
Water
Band 5
Decision boundaries
Vegetation
Soil
Band 7

Figure 4: Probability distributions in a 2d feature space (Richards, 1993)

The mean position of a class consisting of K samples with their respective vectors x_i in feature space is given by the mean vector:

$$\vec{m} = \frac{1}{K}\sum_{i=1}^{K}\vec{x}_i$$

The shape of the distribution is given by the covariance:

$$\sum_{x} = \frac{1}{K-1}\sum_{i=1}^{K}(\vec{x}_i - \vec{m})\cdot(\vec{x}_i - \vec{m})^t$$

Figure 5 illustrates the concept of mean vectors and covariances. The data distribution is described by the covariance matrix (represented by the scatter cloud in the figure), while the mean value is a single point in space. The oval shape of the cluster shows that the two dimensions are correlated.

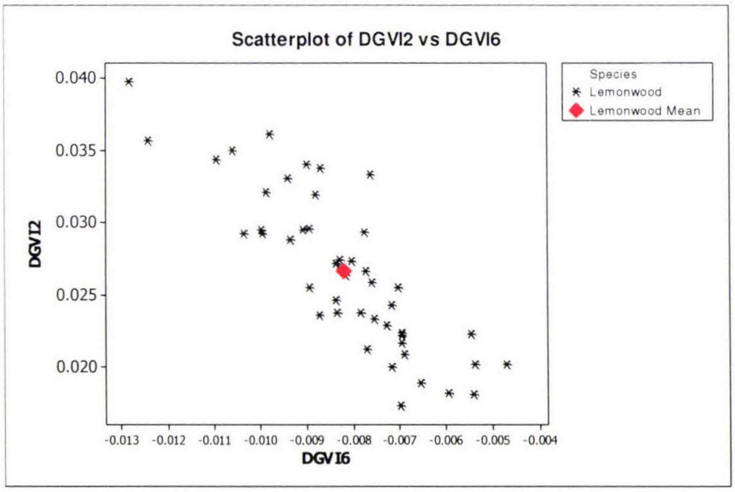


Figure 5: An example of a data distribution in a 2d feature space, showing independent samples of a class and their mean

The correlation between dimensions can be found by interpreting the covariance matrix (Richards, 1993):

- □ If there is little correlation between the axes of a feature space, the off-diagonal elements of the covariance matrix are close to 0.
- □ If there is a correlation, the off-diagonal elements will be large by comparison to the diagonal elements

The following two covariance matrices are examples of little correlation (a) and high correlation (b). This is show graphically in Figure 6.

$$\sum_a = \begin{bmatrix} 2.40 & 0 \\ 0 & 1.87 \end{bmatrix} \qquad \sum_b = \begin{bmatrix} 1.900 & 1.100 \\ 1.100 & 1.100 \end{bmatrix}$$

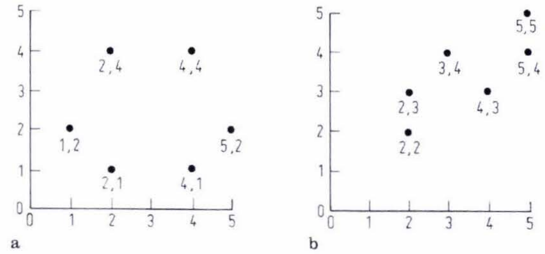


Figure 6: Two dimensional data with little correlation (a) and high correlation (b) (Richards, 1993)

2.3.1.3 On the Importance of 2nd Order Statistics

The use of average values may be useful in some circumstances, however, Landgrebe (1997) notes that the reduction of data to mean values results in a loss of information.

Second order statistics contain vital information about the distribution of data in spectral or feature space. An example of the loss of data is shown in Figure 7. If only the mean values are used, it seems that the classes could be discriminated without any problem. But the scatterplot which shows the variability of the classes reveals an overlap between the classes Lemonwood and Ngaio, thus indicating that a 100% separability of these classes is less likely if this 2 dimensional feature space is used. The discrimination could be increased by utilizing more dimensions.

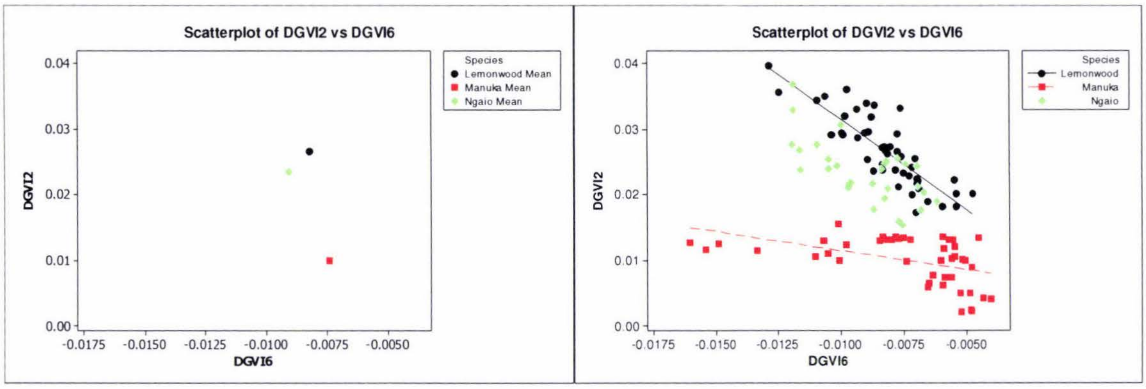


Figure 7: Data reduced to mean values (left) and data including 2nd order statistics information and showing regression lines for each class (right)

2.3.2 Data Processing

2.3.2.1 General Structure

The procedures used in several studies of hyperspectral data show a general, discernible structure which is:

1. Removal of noisy or uncalibrated bands
2. Atmospheric corrections (applies only to airborne and spaceborne sensor data)
3. Data smoothing
4. Reduction of dimensionality
5. Transformation into feature space
6. Analysis techniques

All of these steps are not always necessary. They are described hereafter in detail.

2.3.2.2 Removal of Noisy or Uncalibrated Bands

This step eliminates bands which are either uncalibrated or give no useful signal because of a low signal to noise ratio.

Uncalibrated bands occur when a sensor contains known, faulty sensor elements. An example is the Hyperion sensor, where certain bands are listed as non-calibrated. The removal of uncalibrated bands requires detailed information of the sensor in use.

Low signal to noise ratios are found naturally in some wavelength ranges due to atmospheric interference, e.g. water vapour at 1350-1440nm, 1790-1990nm and 2360-2500nm (Thenkabail et al., 2004a).

Water vapour causes the most noise found in field spectroscopy data. Only when the distance between sensor and sensed object is minimized (e.g. if a contact probe is used) will the influence of the atmosphere be practically non existent.

An example of a spectrum showing water band noise is shown in Figure 8.

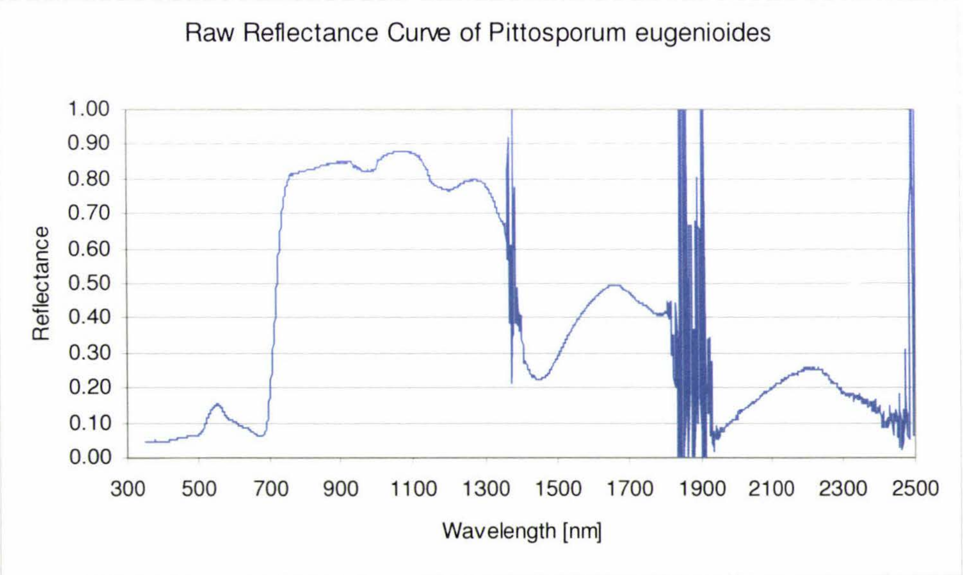


Figure 8: An example of a spectrum showing water band noise in 3 wavelength ranges

2.3.2.3 Atmospheric Corrections

Atmospheric correction applies only to airborne and spaceborne sensor data. Field sensor data do not need to be atmospherically corrected due to the small distance between sensor and object (usually only a few metres maximum).

Atmospheric correction of hyperspectral data is essential and extremely complex (Thenkabail et al., 2004a) and must be carried out if hyperspectral imagery is to be compared with spectral ground data or with other, temporally or spatially different hyperspectral imagery (Lillesand et al., 2004). Thus in the context of this research some form of atmospheric correction will be needed to relate ground spectra to Hyperion data.

Numerous techniques for atmospheric correction exist, amongst which are:

- □ Flat Field (FF) Calibration: the data is normalized against a spectrally flat, uniform area with known spectral reflectance (Vane and Goetz, 1988; Research Systems Inc., 2004)
- □ Empirical Line (EL) Correction: a linear fit between ground reflectance data and raw spectral data is calculated and then applied to the raw data. Ground data can be collected simultaneously with the satellite overpass (Ramsey and Nelson, 2005) or non simultaneously (Martin and Aber, 1997; Ben-Dor and Levin, 2000).
- □ Internal Average Relative Reflectance (IARR): the raw data is normalised against the average spectrum of the image (Research Systems Inc., 2004).
- □ Model based methods: a radiative transfer model is used to calculate surface reflection from raw data. The model requires the amount of water vapour, distribution of aerosols and scene visibility. Due to the contiguous, narrow band spectral data, water vapour information can be extracted from every pixel. Several software packages with this functionality exist: FLAASH, ATREM and ACORN (Kruse, 2004)

The FF and IARR Calibrations are both normalization processes and generally produce the poorest results. The model based methods often produce better results than the other corrections but they need atmospheric information true for the time of data acquisition which can be difficult to obtain. The EL calibration requires information about ground targets and can produce acceptable results within a few percent of true reflectance (Smith and Milton, 1999).

2.3.2.4 Data Smoothing

Hyperspectral data acquired by field, airborne or spaceborne sensors exhibit a certain degree of random noise. The combination of high spectral and relatively high spatial resolution renders imaging spectrometers sensitive to noise (Landgrebe, 1997). Field sensors tend to have even narrower bandwidths than airborne or spaceborne sensors and are sensitive to noise even when close to the object. The reduction of this noise is especially crucial when derivative analysis is to be employed (Tsai and Philpot, 1998). Explicit data smoothing can be omitted if the dimensionality of the data is reduced by a method that implicitly applies a smoothing function (see details in section 2.3.2.5).

The goal of every filtering function must be to reduce the noise while preserving the original features. Some smoothing techniques are reviewed hereafter.

One commonly used operation is the convolution. Here, a convolution function is moved over the data points and the mid point of this moving window is the data point to be smoothed. One of the best known convolution functions is the average (Savitzky and Golay, 1964).

The convolution process is described by:

$$Y_j* = \frac{\sum_{i=-m}^{i=+m} C_i Y_{j+i}}{N}$$

where

Y_j* = smoothed data point

C_i = convolution coefficient

Y_{j+i} = original data point

N = moving window size (-m...+m)

For the average, all coefficients are 1 and N is the number of convolution coefficients.

One of the most popular smoothing functions applied to hyperspectral data is the Savitzky-Golay filter (Tsai and Philpot, 1998). It uses linear least squares regression to smooth the data; a polynomial of a certain order is fitted to N data points, where N is defined by the filter size. An advantage of this filter is the ability to calculate smoothed derivative data in one operation.

Savitzky and Golay (1964) provided tables with the convolution coefficients for different combinations of filter sizes, derivative orders and approximating polynomial orders. While these lookup tables served well to increase the computing speed of the machines available when this technique was developed, filters are limited to the filter size/polynomial order/derivative order available in these tables. Modern implementations therefore calculate the required coefficients at run time (Tsai and Philpot, 1998; Press et al., 2002).

Tsai and Philpot (1998) noted that the filter size was the principal factor that affected the results of derivative analysis.

A study conducted by Schmidt and Skidmore (2004) investigated several smoothing techniques (Mean, Median, Savitzky-Golay, Discrete Wavelet Transformation (DWT), Non-decimated DWT and Cubic Spline) for noise reduction of vegetation data. It suggested that the wavelet transformations were superior to the other methods.

Another well-known filtering technique is the Fourier transformation, but it has been shown that wavelet transformations preserve local features better because they are locally adaptive (Press et al., 2002; Schmidt and Skidmore, 2004).

Piecewise multiplicative scatter correction (PMSC) is based on linear regressions when fitting against a standard spectrum. It is used to correct for nonlinear additive and multiplicative scatter (Fyfe, 2003).

2.3.2.5 Reduction of Dimensionality

It has been shown that neighbouring wavebands have a high degree of correlation, i.e. they contain redundant data (Thenkabail et al., 2004a). This redundancy is created by oversampling, i.e. the spectral signal is sampled at small enough steps to describe very narrow features that could be discriminating

(Shaw and Manolakis, 2002). This oversampling is what caused several studies to research the ideal wavebands needed for certain applications. Such knowledge could then lead to specialised sensors that capture optimal bands and thus reduce the data redundancy (Thenkabail et al., 2000; Thenkabail et al., 2004a).

By using appropriate techniques it is possible to reduce the dimensionality significantly while retaining most of the information.

The most widely used algorithm is the principal component transformation (PCT) (Shaw and Manolakis, 2002). Principal component analysis performs an eigen-decomposition, the resulting eigenvectors are used to build a transformation matrix, which is then applied to the original data. The PCT is a zero correlation, rotational transformation. The components are ordered by their power to describe the variation found in the data. The first few components explain the most variation, while the later components usually contain noise (Richards, 1993). By choosing a subset of the available eigenvectors to build the transformation matrix, the data dimensionality can be reduced drastically while retaining most information.

The Maximum Noise Fraction (MNF) Transform (Green et al., 1988) is similar to the PCT but addresses the weakness of the latter when the noise variance is not uniform over all bands of the dataset. MNF is a linear transformation and orders the resulting components by their signal-to-noise ratio. MNF, also known as NAPC (Noise-Adjusted Principal Components), is therefore a useful technique to reduce the dimensionality of a dataset while retaining most information and minimizing the noise at the same time (Lee et al., 1990).

Some researchers have reduced the data by simply selecting every tenth waveband, thus reducing the data by factor ten (Shepherd and Walsh, 2002). This approach should be treated with caution, as it may violate the sampling theorem by Shannon (1949). The sampling theorem states that the discrete samples are a complete representation of the signal if the bandwidth is less than half the sampling rate. Shannon's sampling theory is applicable whenever the input function is band-limited. When this is not the case, the standard signal-processing practice is to apply a low-pass filter prior to sampling in order to suppress aliasing (Unser, 2000). The process of filtering followed by downsampling is referred to as decimation (Fliege, 1994). Thus, in the context of spectral data, the application of a smoothing function which is effectively a low-pass filter, may be advisable prior to a downsampling.

Another possibility of dimensionality reduction is the simulation of other hyperspectral sensors having fewer bands than the original sensor.

Thenkabail et al (2004a) transformed ASD spectroradiometer data to Hyperion data by using 10nm bandpasses. The filtering function of the bandpass was not detailed.

A different study also simulated Hyperion data by averaging every ten bands of the ASD data (Mathur et al., 2002). The use of the average function seems questionable, however, as the sensor response function of the Hyperion sensor is of Gaussian nature (Zanoni et al., 2002). It would therefore seem logical to use a Gaussian instead of an average function for the band convolution process.

The simulation of other sensor responses from given data is an important operation, e.g. for the performance evaluation of new sensors (Zanoni et al., 2002).

As an example, Landsat7 ETM+ was chosen in this study because (a) Landsat imagery is widely available as the Landsat program has already run for decades (b) many studies of have produced successful results, e.g. a study on New Zealand vegetation (Dymond and Shepherd, 2004) and (c) the atmospheric correction of New Zealand Landsat imagery has been carried out at Landcare Research and Landsat simulated signatures can therefore provide valuable information when trying to correct Hyperion data.

2.3.2.6 Feature Space Transformation

In feature space, signals are treated as vectors in a multidimensional space (Landgrebe, 1997). Technically, it suffices to arrange the reflectance bands of a spectrum in vector form to achieve the transformation into feature space.

From the example given in section 2.3.1.1 where a 2 dimensional feature space was shown, it becomes clear that a feature space must not be of dimension N if the spectrum was sampled in N bands. A feature space can be built so that it maximises the discrimination between classes.

The real power of the feature space lies in the possibilities for information extraction. A wealth of stochastic methods exists that can be applied to vector data (e.g. Minimum Distance to Means or Maximum Likelihood) (Landgrebe, 1997). Many studies make use of the feature space concept, although it is usually not explicitly mentioned. A few examples of feature space transformations are given hereafter.

Principal Components Transformation (PCT) (see also section 2.3.2.5) is widely employed. It is a linear algebra method and as such operates in feature space. PCT transformed data represent an example of an optimised feature space as their axes are theoretically uncorrelated.

The calculation of indices also performs a transformation of spectral data into a feature space. Indices are mathematical combinations of reflectance band data. The simplest index is the difference between the reflectances of two bands:

$$I = \rho(b_x) - \rho(b_y)$$

The influence of illumination conditions, surface slope, aspect and other factors on the indices can be reduced by normalization (Lillesand et al., 2004):

$$NI = \frac{\rho(b_x) - \rho(b_y)}{\rho(b_x) + \rho(b_y)}$$

E.g. by calculating a NDVI (Normalized Difference Vegetation Index), the spectral data is automatically transformed into a 1 dimensional feature space.

Derivative Greenness Vegetation Indices (DGVI) (Elvidge and Chen, 1995; Thenkabail et al., 2004a; Thenkabail et al., 2004b) make use of many hyperspectral bands. They describe changes in slopes by summing up the differences of first derivatives over defined waveband regions.

$$DGVI = \sum_{i=m}^{n} \frac{\rho'(b_{i-1}b_i) - \rho'(b_i b_{i+1})}{\Delta b_i}$$

where

$\rho'(b_{i-1}b_i)$ = first derivative of reflectance curve between b_{i-1} and b_i

m..n = start and end band number of DGVI area

b_i = centre wavelength of band i

i: band number

Δb_i = step width: $b_{i+1} - b_{i-1}$

These DGVI regions are: 515-535 nm (DGVI1), 540-560 nm (DGVI2), 560-580 nm (DGVI3), 650-670 nm (DGVI4), 700-740 nm (DGVI5), 626-795 nm (DGVI6), 1500-1650 nm (DGVI7), 2080-2350 nm (DGVI8) (Thenkabail et al., 2004a) and 428-906 nm (DGVI9), 428-2355 nm (DGVI10) (Thenkabail et al., 2004b).

Thenkabail et al. (2004a) carried out nonparametric least significant tests on the mean DGVIs of different vegetation. The most discriminating DGVI was found to be DGVI5, followed by DGVI6, DGVI7, DGVI3, DGVI4 and DGVI8.

Thus, feature spaces can be built of:

□ Any subset of original reflectance bands

□ Any subset of zero correlation transforms (e.g. PCT)

□ Any number of indices

2.3.3 Analysis

The typical analyses carried out can be grouped into the following broad categories:

□ Basic research into discrimination, best bands, etc.

□ Predictive correlation studies to develop measures that can predict physical properties from spectral data

□ Spectral unmixing: estimation of the abundance of endmembers

In practice, these categories often overlap or are combined to get the best results.

2.3.3.1 Discrimination

Theoretically, every material should have a unique spectral signature. The study of the discrimination of materials forms a basis for the classification of individual signatures or hyperspectral imagery. Landgrebe (1997) notes two approaches to the problem of classification: spectral matching and analysis in feature space.

2.3.3.1.1 Spectral Matching

Spectral matching regards the data to be classified as spectra, i.e. a continuous curve over a defined wavelength range. Here, an unknown spectrum is compared to known spectra. A match thus identifies the unknown.

An early example is the two code binary vector, consisting of amplitude and slope information. A spectral match is determined by assigning the unknown spectrum to the reference spectrum which has the minimum Hamming distance (Mazer et al., 1988). An implementation of this algorithm, called Binary Encoding, can be found in ENVI (Research Systems Inc., 2005).

Other ways to match spectra are: a distance calculation based on the root mean square difference between two spectra over a wavelength region (Price, 1994) or a least squares fit against reference spectra, termed Spectral Feature Fitting (SFF) (Research Systems Inc., 2005). SFF uses continuum removal and thus identifies the absorbtion band centres. It has been successfully used in a preliminary study to detect two types of aquatic vegetation species (Williams et al., 2002).

A recent development in the domain of spectral matching is the USGS Tetracorder expert system which uses continuum removal and least squares fitting to match library and unknown spectra. The resulting correlation values r are termed the fit value. In a study that mapped the landcover of the Yellowstone National Park, an additional raster image containing these fit values was produced, showing the degree of confidence of the match. The accuracies achieved were 91% for 4 forest classes and 85% for 4 non-forest classes. It also showed a classification accuracy of 40.3% - 93% for 5 growth stages of one species (Kokaly et al., 2003).

The spectral matching makes use of known spectra, usually compiled in spectral libraries, such as the USGS spectral library (Clark et al., 1993). Such libraries tend to contain only a limited number of spectra per material type, in many cases just one single, representative spectrum which means that only 1st order statistical data is available.

The concept of the Shape Space (Cochrane, 2000) tried to overcome this limitation by defining the variability of classes by the upper and lower bounds of spectral reflectance. Classification is then achieved by a process termed shape filtering. The classification criterion is the fit or overlap of the unknown spectrum with the shape of a class. Although Cochrane reports quite high classification accuracies, the way in which the reflectance of branches and trees was estimated using leaf spectra suggests that the results of the shape space approach should not be regarded as conclusive.

2.3.3.1.2 Feature Space Representation

The Feature Space representation, on the other hand, models classes as clusters in a multidimensional space and as such offers the possibility of using 1st and 2nd order statistics more easily.

The problem of assigning an unknown vector to one of several clusters in a multidimensional space can be solved by using discriminant analysis, also called supervised pattern recognition. A training set is used to find a discriminant function (linear or quadratic). This function is subsequently applied to new objects to allocate it to a group (Miller and Miller, 2005).

Like discriminant analysis, partial least squares regression (PLS) is a method of multivariate analysis. Although PLS is normally used to model continuous data, it has been successfully used to predict group memberships for two or three groups by assigning numerical codes to the groups (Richardson et al., 2003).

A study of classifiers using either mean or covariance or combining both showed that the best classification results are achieved when a classifier makes use of both statistics (Landgrebe, 1997).

A few examples of studies based on the feature space concept are given hereafter.

Younan et al (2004) studied the discrimination of 8 different sample types (bare soil, soybean, mixed weeds, combination of soybean and weeds, and 3 types of weed). Half of the samples were used as a training set and the other half was classified against the training set. Six different nearest neighbour calculations gave overall classification accuracies between 33% and 68%. A further classification also used nearest neighbour as the discriminant function but the input data were wavelet coefficients obtained from a wavelet decomposition of the spectra. Wavelet decomposition represents a signal by approximation and detail vectors. It is mostly used in signal de-noising and image compression. The concept of the wavelets can also be extended to feature extraction and classification. The wavelet based classification resulted in 45% accuracy.

Several points are noteworthy: the species vectors were made up of all sampled bands, thus many components would be found redundant. As no smoothing was applied, the data was still noisy. The discrimination of eight surface types, out of which one (the soil) was very different to the vegetative types, should have yielded quite good results. The tendency of the classification accuracy when adding more surface classes remains unanswered, but one would expect that more classes result in a reduced accuracy.

The separability of classes in hyperspectral space can be determined by using a distance analysis. Two such measures are the Jeffries-Matusita (JM) and the Bhattacharya (B) distance (Schmidt and Skidmore, 2003). The JM distance is asymptotic to 2, i.e. a value of 2.0 would equal a 100% separability of the two classes (Richards, 1993). A value of 1.9 indicates a good separability (Research Systems Inc., 2004).

Schmidt and Skidmore used the JM and B distances in a study into spectral discrimination of vegetation types. They reported JM and B distances between 27 classes using 6 wavebands. The JM distances were between 0.81 and 2.0 with the majority of the distances around 1.8. They concluded that for overlapping classes, other information such as elevation could aid the distinction.

Instead of measuring the distance between two vectors in space, their separability can also be determined by calculating the angle between the two vectors (Price, 1994). This measure is called Spectral Angle Mapper (SAM) (Landgrebe, 2003) and is part of the ENVI software (Research Systems Inc., 2005). Again, this is a method that works on single vectors and not on clusters and therefore only uses 1st order statistics. This limitation can be partly overcome by adding numerous variations of the same endmember to the spectral library. One advantage of SAM is its insensitivity to changes in signal strength, i.e. object albedo. A lower/higher albedo should only change the length of the signal vector but not its direction.

Mundt et al. (2005) used SAM for the discrimination of an invasive plant species (hoary cress) in airborne hyperspectral imagery. Reported classification accuracies were around 80% for areas of more than 30% infestation. Signatures of the target species were selected from the imagery after overlaying ground survey data. Two endmembers (mesic and xeric) were then formed by averaging the selected regions. No spectral signature examples that compare target to non-targets were given in the article. However, the imagery and field surveys took place during bloom. Hoary cress exhibits dense white flowers, the plant forms flat, mat like covers. Thus, one could assume that hoary cress signatures are significantly different from the surrounding landcover signatures.

Clark et al. (2005) compared the accuracy of SAM, Maximum Likelihood (ML) and Linear Discriminant Analysis (LDA) when classifying tropical rainforest trees at leaf and crown scale. Generally the performance of SAM was lower than 53.7% while LDA and ML reached a maximum of 100% and 87.3% respectively. It was concluded that the poor result of SAM was due to the interspecies variability which is not entered into the model as SAM uses no 2nd order statistics.

2.3.3.2 Best Bands

Best bands are a subset of the original bands that maximise the separability of the classes.

As mentioned before, hyperspectral data is usually highly correlated. Thus the search for the best bands should also identify non correlated bands. The combination of such bands then forms a feature space with an optimised discrimination. A few examples follow.

The discriminative power of single wavebands (i.e. dimensions) can be tested using statistical methods. The Mann-Whitney U-test determines if two populations are statistically significantly different. By applying this test to all species combinations at every waveband and counting the cases where the populations differ, a histogram is computed that shows the important wavelengths in terms of discrimination (Schmidt and Skidmore, 2003; van Till et al., 2004). A process called 'single-factor analysis of variance' (Fyfe, 2003) renders the same information as the Mann-Whitney U-test.

The results of the Mann-Whitney U-test with a significance level of 0.01 applied to saltmarsh vegetation showed that the most discriminating wavebands occurred in the NIR and SWIR regions (740-1820 nm) of the spectra (Schmidt and Skidmore, 2003). The wavebands in these regions were >83% statistically different between species, i.e. the p-values of these tests were less than the chosen significance level for at least 83% of all cases. Wavebands between 1970 and 2450 nm were >77% statistically different.

Lambda-Lambda R^2 models (LL R^2 M) are a data mining technique that identifies band combinations of low correlation (Thenkabail et al., 2004a). By calculating the correlation matrix of a number of given spectral vectors, a correlation factor r is obtained for every possible band combination. After the conversion of r to R^2 the matrix can be plotted as raster or contour image highlighting the least correlated band combinations. Thenkabail et al. (2004a) used LL R^2 M in a study into waveband performance, applied to samples of crops and weeds. The most frequently occurring, non redundant wavebands were: red, FSWIR (far short wave infrared; 1901-2500 nm), ESWIR (early short wave infrared; 1301-1900 nm) and late NIR. They suggested that LL R^2 models are most useful when testing species where spectral similarities are likely to be close.

Principal Components Analysis (PCA) can also yield information about possible best bands. The influence of the original bands on the data variability is given by the factor loadings, i.e. components of the eigenvectors (Thenkabail et al., 2004a). Thenkabail et al. (2004a) carried out PCA on weed and crop spectra. The first five principal components (PCs) explained 93-95% of the variability. The original 168 bands could therefore be reduced to 5 new bands, resulting in a reduction of the data volume by about 97%. The ESWIR bands had the highest factor loadings in the first PC which explained 65% of the variability. The second PC was dominated by the red wavelengths and explained 20% of the variability. PC3-PC5 had the highest factor loadings in the FSWIR. This indicated the importance of red and SWIR wavebands for the discrimination of vegetation.

Stepwise Discriminant Analysis (SDA) is a multivariate technique that tries to identify an optimal set of predictors (bands) by a stepwise selection (Thenkabail et al., 2004a). One of the outputs of SDA is the Wilk's Lambda. The smaller Wilk's Lambda, the better the discrimination. Thenkabail et al. (2004a) applied SDA to shrub, grass, weed and crop spectra. The most frequently selected wavebands that yielded optimal Wilk's Lambda values for shrubs, grasses, weeds and crops were centred at 1215, 730, 1245 and 1285 nm respectively. This indicated that discriminating bands are situated in the NIR.

A method termed best feature selection (Mathur et al., 2002) constructs a feature vector based on the area under a ROC (Receiver Operating Characteristic) curve. The area under the ROC curve is related to the amount of histogram overlap of two classes. The feature vector elements are then used as the weights in a linear discriminant analysis. Mathur et al. used this method to classify grass species into two classes: weed (1 species) and non-weed (5 species). The field spectra were first convolved into Hyperion sensor bands and subsequently put through the best feature selection process. Nearest Neighbour was used as the discriminant function and classification accuracies of 85.47-97.98% were reported.

2.3.3.3 Predictive Correlation

Hyperspectral data offer new, non-destructive and efficient ways of estimating physical properties of objects. E.g. estimation of biomass, leaf area index (LAI) or prediction of crop yield.

The challenge is to identify spectral features that correlate with physical measurements.

Four approaches to this problem are discernable: (a) to use knowledge about the electron transition or the bond vibration of chemicals at certain wavelengths or (b) to rely on mathematical tools or (c) to visually assess the spectral reflectance curves to identify high correlations between predictors (the reflectances) and the responses (the physical data) or (d) to use or modify indices provided by other studies.

2.3.3.3.1 Absorbtion/Reflectance of Chemical Bonds

Because of the frequent overlap of spectral characteristics of compounds, the interpretation of plant spectra using compound absorbance is difficult at best (Richardson et al., 2003). The absorbtion features (position, depth and width) of chemical bonds are however frequently and successfully used in mineral or chemical applications. The analysis of absorption usually involves continuum removal (Kokaly and Clark, 1999) as a preceding operation.

However, knowledge about reflectance/absorbance characteristics of compounds can help to understand the shape of spectra. An example is the low reflectance of plants in the visible wavelengths due to chlorophyll absorbtion.

Analysis of absorbtion features has been used for the successful estimation of foliar nitrogen (coefficient of determination $r^2 = 0.85$) and most of the known nitrogen absorption features could be identified (Huang et al., 2004).

2.3.3.3.2 Mathematical Tools

From a statistical viewpoint, the collection of hyperspectral reflectance yields multivariate data. Multivariate analysis is a branch of statistics that can deal with multiple measured variables per object. This paragraph is based on Miller et al (2005) and Minitab (2003).

Multiple Linear Regression (MLR) finds regression equations of the form

$$c_i = b_{0i} + b_{1i}A_1 + ... + b_{ni}A_n$$

where

A_n = predictor

c_i = response

b_{ni} = coefficient to be determined

The number of sampled specimens must be greater than the number of predictors n. This limits the use of MLR when hundreds of predictors (i.e. wavebands) are available. The determination of the best combination of all possible predictor combinations would take a long time. Here, the evaluation of the best bands can yield useful results. Also, MLR can not handle high colinearity of the predictors. Using too many predictors can result in an overfit.

MLR was used in a study that tried to estimate foliar nitrogen content of Eucalyptus species from Hyperion data (Coops et al., 2003). In order to prevent overfitting, three bands were used that explained the most variation in N at the sample plots: 458, 2264 and 2294 nm. An R^2 value of 0.84 was achieved, although this result must be treated with caution as no jack-knifing procedures were applied.

One solution to the colinearity and overfitting problem of MLR is to apply a principal component transformation (PCT) to the data first and then carry out a MLR. PCT is a multivariate technique that can reduce data if the variables are correlated (see section 2.3.2.5).

The combination of PCT and MLR is known as principal components regression (PCR).

Partial least squares (PLS) regression is in its concept similar to PCR as it uses linear combinations of the predictor variables. However, in contrast to PCR, PLS does not try to maximise the variation of the predictors but gives extra weight to predictors that are highly correlated with the responses.

Coops et al. (2003) used PLS for the above mentioned Nitrogen estimation and reported an R^2 value of 0.68 using cross validation (R^2 0.95 without cross validation). They also noted that outliers can negatively affect the accuracy of PLS while MLR tends to be more robust.

Stepwise regression is one more way of multivariate regression. Stepwise regression tries to build an optimal subset of predictors that maximises the regression correlation. A number of algorithms exist to derive this best set: add and remove, forward selection or backward elimination (Minitab Inc., 2003).

Thus, it seems that PCR, PLS and stepwise regression are all well suited tools for correlation studies using hyperspectral data.

2.3.3.3.3 Visual Assessment

Visual assessment of spectral plots may help to identify regions where discrimination seems likely.

This technique was used in a study of sugarcane disease where both the magnitude of the difference between band reflectances and the direction of relationship (i.e. divided band reflectances) were assessed. Results of visual assessment were then combined with statistical information to create new indices (Apan et al., 2003).

2.3.3.3.4 Use and Modification of Existing Indices

To improve existing indices has been the goal of many studies over the past years, e.g. (Thenkabail et al., 2002; Apan et al., 2003; Haboudane et al., 2004).

In plant studies, one of the most widely used indices is the NDVI (Normalized Difference Vegetation Index). It is formed by contrasting red band with near infrared (NIR) band reflectance (Elvidge and Chen, 1995):

$$NDVI = \frac{\rho(IR) - \rho(VIS)}{\rho(IR) + \rho(VIS)}$$

It is known that the red band is dominated by chlorophyll absorbtion while the NIR has a high reflectance due to the internal leaf structure (Lusch, 1989). The NDVI has been used successfully for large area vegetation monitoring using AVHRR (Advanced Very High Resolution Radiometer) data. Vegetated areas usually yield a high NDVI value while non-vegetated areas tend to have negative values.

In terms of existing indices, it must be noted that quite a number were developed for use with broadband sensors. Their direct application to hyperspectral data does not necessarily exploit the higher information content of these data. In a study that correlated several indices with LAI, it was found that newly developed, narrow band indices were superior to existing broadband indices, even when the bandwidth of the latter was reduced (Elvidge and Chen, 1995).

Data mining techniques are also applicable to two band indices to identify the best band combinations, resulting in an R^2 plot similar to the Lambda-Lambda R^2 band correlation plot (Thenkabail et al., 2002).

2.3.3.4 Spectral Unmixing

Spectral mixing occurs for two reasons: (a) the spatial coverage of the sensor includes more than one endmember or (b) the material being sampled is in fact a homogenous mixture of two or more endmembers (Keshava and Mustard, 2002). Endmembers are materials that are pure, i.e. mixtures are made up of endmembers.

For imaging spectrometers, the spectral mixing results in mixed pixels, also called mixels. Although the process of unmixing is usually applied to raster images, it can conceptually be applied to field spectroradiometer data as well.

Spectral unmixing is the procedure that yields the abundances of the involved endmembers.

Two models that describe the mixing exist: Linear Mixing and Nonlinear Mixing.

2.3.3.4.1 Linear Mixing Model (LMM)

Linear mixing assumes that the surface consists of distinct materials (the endmembers) and incident energy only interacts with these pure materials. The reflectance that arrives at the sensor consists of all endmember signals in the field of view. If only one endmember takes up the field of view, its abundance is 100%, if more than one endmember make up the field of view, their abundance is equal to the proportion of the area they occupy. Figure 9 illustrates the concept of linear mixing: of the three occurring endmembers A, B and C, A and B have a fractional abundance of 0.25 while C has an abundance of 0.5.

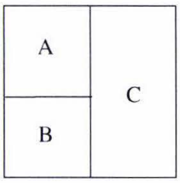


Figure 9: An example of a mixed pixel (linear mixture model)

Mathematically, the linear mixture model can be written as

$$X = \sum_{i=1}^{M} a_i \cdot s_i + w = S \cdot a + w$$

where X is the N x 1 signal vector received by the sensor, S is a N x M matrix, consisting of M endmember vectors and w is a N x 1 additive noise vector.

Two further conditions must be satisfied for the unknown abundance vector *a*: the full additivity constraint requires that the sum of all abundances must be 1 and the nonnegative constraint requires that all abundances must be positive:

$$\sum_{i=1}^{M} a_i = 1 \quad , \qquad a_i \geq 1, i = 1..M$$

The unmixing consists of three consecutive procedures (Keshava and Mustard, 2002):

1. Dimension reduction (optional, reduces computation effort)
2. Endmember determination
3. Inversion (a least squares solution)

The selection of the endmembers is critical. The abundance estimation accuracy is highest when the exact number of endmembers are used in the model. If too few endmembers are used the estimated fractions will include the abundance of the missing endmembers. This is termed fraction error. If too many endmembers are used the model will be sensitive to instrumental noise, atmospheric contamination and natural variability in spectra, resulting again in fraction errors (Roberts et al., 1998). Additionally, not only the number of endmembers in the model influences the result but the correct endmembers should be selected that are present in the scene. A technique called multiple endmember spectral mixture analysis (MESMA) tries to address these two issues. Based on a collection of endmembers, sets of endmember mixture models are created. These models are then applied to each pixel in the image. For every model the root mean square error (RMSE) between modelled spectrum and observed spectrum is calculated. The model that minimizes the RMSE is chosen (Roberts et al., 1998).

If it is assumed that the endmembers are pure substances then their spectra should reside along the hull of a multidimensional space. Thus, mixed spectra occupy the interior of the space (Keshava and Mustard, 2002). If two endmembers and their mixtures are plotted in spectral space, the endmembers take up the highest and lowest reflectance values while their mixtures show reflectances in between, i.e. the endmember spectra enclose the mixtures. This creates a problem if an endmember C lies totally in between two other endmembers A and B. In this case it is not possible to distinguish the endmember C from mixtures of A and B (Price, 1994). In these circumstances spectral unmixing is unlikely to yield useful results.

2.3.3.4.2 Nonlinear Mixing Model

In contrast to the LMM, the nonlinear mixing model does not assume that the endmembers appear in segregated areas but can be mixed at spatial scales smaller than the path length of photons. Sand grains made up of different compositions are an example of such a surface type. Due to multiple scattering between the grains, the resulting signal is a nonlinear mixture.

A solution to the nonlinear mixing is the development of models for particulate surfaces. At present it is still unclear whether spectral signatures of mixed pixels are dominated by linear or nonlinear mixing. If linear unmixing is applied to nonlinear mixtures, the absolute errors can be up to 30% (Keshava and Mustard, 2002).

2.4 Spectral Libraries and Spectral Databases

A main focus of hyperspectral remote sensing research is basic research or correlation studies as mentioned above. In the context of such studies, spectral libraries are frequently used but rarely are they explained in detail. Spectral database is a term heard of even less.

At first, it may seem that the difference between spectral libraries and databases is subtle but, as will be shown in this section, this is not the case.

2.4.1 Spectral Libraries

Spectral libraries are best described as a collection of representative spectra of a variety of materials. As such, they are crucial for identification of unknown spectra and aid the correction and classification of remote sensing data by providing endmember spectra.

Some of these libraries are accumulated during a specific study, e.g. a spectral library for urban materials containing non averaged data from a ground survey (Herold et al., 2004) or land cover types being averaged AVIRIS pixel signatures (Kokaly et al., 2003).

None of the above studies detailed how the library was organised or what metadata was assembled.

Price (1994) studied the variability found in crops. The accuracy of spectral matching against library spectra led to the conclusion that the accuracy could be increased if libraries contained a larger number of cases (i.e. spectra showing the variability of a given material).

A well known public domain spectral library is provided by the USGS. It is focused mainly on laboratory spectra of rocks and minerals but includes a few vegetation spectra as well. It contains 498 spectra of 444 samples (i.e. different materials). As such, mostly only one representative, high quality spectrum is available for each material. Consequently, no second order statistics are held in this library.

Technically, the library is one binary file with a record data structure. Apart from reflectance data, each record holds information such as: record number, title, date of acquisition and length of data set. Also included in this file is information about the spectrometer used, wavelength range, resolution and spectral purity (Clark et al., 1993).

The majority of the publicly available spectral libraries are distributed as physical files. This has drawbacks such as low flexibility and low query performance (Bojinski et al., 2003).

Milton (2001) lists metadata that should be contained in a spectral library of field data such as: location of site, time/date, sky conditions, instrument details, viewing geometry, height of sensor above ground and band information.

It is unclear if any libraries have been assembled that include metadata as suggested by Milton. Missing metadata can render spectral information useless as the circumstances of the capturing event are unknown. Only a complete metadata allows the researcher to gain confidence that the spectra are indeed representative for the intended use.

It is concluded that spectral libraries contain vital information but their organisation is unclear in many cases. It would not be surprising to find some libraries that are merely a collection of single reflectance files residing in a folder.

2.4.2 Spectral Databases

'Data are unstructured facts and figures. When they have been organised or processed, they become information' after Williams and Summers (2004).

As pointed out above, the organisation of spectral libraries is rarely an issue.

Generally, the organisation of spectral data collected during studies is never detailed.

Typically, after having conducted several field or laboratory sampling campaigns one can expect to end up with thousands of files plus associated metadata.

The time and effort that are spent in collecting spectral data, combined with the characteristically large number of files, makes it clear that spectral data should be well organised. Otherwise valuable data can be lost or lose their value because of missing metadata.

Considering the above, it seems logical to employ a database to store spectral data in a suitable form.

Only one example of such a database has been found: SPECCHIO (Bojinski et al., 2003) contains spectral metadata ordered by campaigns, information about sensors, instrument models, landuse type of the sampled area, spatial position and descriptions of the target. A relational database management system (DBMS) is used to hold the above data in several tables. The actual reflectance data is not stored in the DB but held on a dedicated file server and the spectral database links the metadata to the reflectance file via a file path.

A web based interface is used to interact with the system. The database can be queried to show e.g. information about field campaigns, locations, target types and land cover. Researchers can subsequently download required spectral data to their workstations.

The centralised database approach of the described system facilitates the sharing of field data of different studies and ensures the integrity of the data.

Despite the fact that modern database systems can handle huge volumes of data easily, a study (Bell and Baranoski, 2004) has been undertaken to investigate the possibility of reducing the dimensionality, and as such the data amount, of plant spectral databases.

The data size can be minimized while still retaining much of the information by applying a principal component transformation to the spectra. The number of utilised principal components influences the accuracy of the reconstructed spectra. The database needs to store the transformation matrix V or a subset of V. The decomposition of the observation matrix M is done by applying the singular value decomposition (SVD) $M = USV^T$ or by performing an eigen-decomposition of the covariance (or correlation matrix) of M.

A spectrum x is then transformed by

$$y = x \cdot V$$

respectively reconstructed to a certain accuracy given by the number of components by:

$$x = y \cdot V^T$$

2.5 Intermediate Conclusions

As demonstrated above, there exists a certain chain of processes that may be applied to hyperspectral data in order to derive useful information. For all these stages, different techniques and philosophies exist. In order to gain a sound knowledge of hyperspectral data acquisition and processing, the most suitable and promising methods should be applied to real data.

The review of spectral libraries and databases reveals an open field where not much work has been done yet. In terms of organisation and storage of spectral data, the concept of spectral databases seems to be the best solution.

3 Methods

3.1 Acquisition and Storage of Field Data

3.1.1 Dataflow Overview

The dataflow adopted for this study is illustrated in Figure 10. An ASD FieldspecPro spectroradiometer was used to capture the radiance and calculate the reflectance of field objects. A GPS was connected to the field laptop for most of the field data acquisition and the spatial position of the field object was added to the metadata, which also included user comments and date/time of capture. Reflectance and metadata were automatically saved in a binary file for every reading taken.

These binary files were transferred to a laboratory computer where they were read by customised software and stored in the relevant tables in the spectral database.

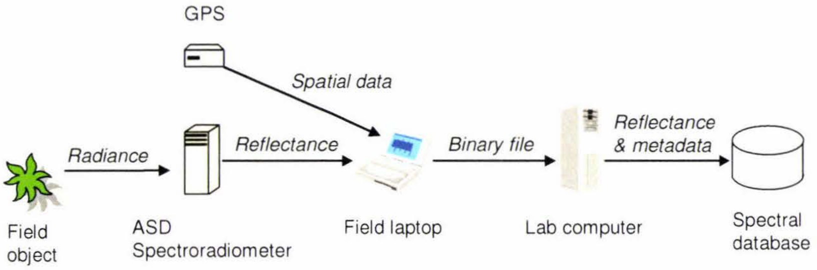


Figure 10: Dataflow and involved hardware

3.1.2 ASD FieldSpecPro

The Institute for Natural Resources had recently acquired a FieldSpecPro spectroradiometer (Analytical Spectral Devices Inc.). This instrument records spectra from 350-2500nm and samples at intervals of 1.4nm for the region 350-1000nm and 2nm for the region 1000-2500nm. These known data points are then interpolated by cubic splines to produce 1nm spaced data points. The sampling unit is comprised of three separate spectrometers: VNIR (Visible and Near Infrared), SWIR1 and SWIR2 (SWIR = Short Wave Infrared). The data of the three elements are spliced at 1000nm (VNIR – SWIR1) and 1800nm (SWIR1 – SWIR2). The light is fed into the system by a 3 metre fibre optic.

3.1.3 Study Sites

Spectra of native plants were collected at four different sites on the North Island:

- □ Massey University Turitea campus, Palmerston North
- □ Foothills of the Tararua Range, catchment of Turitea stream
- □ Along the Mountain Road between Ohakune and Turoa Skifield, Tongariro National Park
- □ Queen Elizabeth II Nature Trust(QE II Trust) Land near Otorohanga, King Country

The first two sites were selected due to their proximity to the institute's location. The Mountain Road, Tongariro National Park was chosen to (a) capture different species that are found in mountainous areas only, (b) provide easy accessibility by car and (c) collect ground data to be used in connection with a

Hyperion image covering the Tongariro National Park that had been acquired. The QE II Trust was used due to the good accessibility and the variation of podocarp species found.

3.1.4 Structure of Field Data

Storing the binary files in an organised manner helped to keep control of the data and enabled an automated import into the database at a later stage.

A hierarchical data structure that reflects the real world and the setup of sampling campaigns was designed. This structure was derived from the following conditions:

1. Reflectances of several different species are captured
2. In order to describe the in-species variation, several specimens of a species are sampled
3. The variability of the specimens is described by several measurements per specimen

The spatial extent where a specimen is sampled was termed a sample site, thus a species contained a number of sample sites. The sites were numbered in the order of their capturing. At each site, several readings were taken to capture the variation exhibited by the specimen in question. A site therefore contained a number of spectra. This led to a hierarchical directory structure (Figure 11).

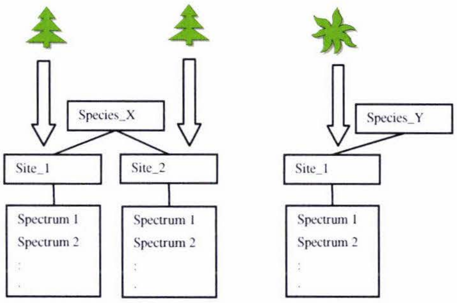


Figure 11: Hierarchical directory structure

3.1.5 Acquisition of Field Data

Spectra of New Zealand native plants were acquired in the field using an ASD FieldSpec Pro spectroradiometer.

Standards for the collection of field data were:

- □ Only cloudless conditions were used
- □ Readings were taken from nadir
- □ Data for each specimen were stored as separate site
- □ White references were taken every few readings
- □ An average number of 10 samples were collected per site
- □ The samples were averaged over 10 readings internally by the spectroradiometer
- □ Collection of spectra took place between 11am and 1pm (data collected during winter)
- □ A bare fibre optic with a 25° field of view was used
- □ Homogenous targets were selected to provide the best endmembers possible
- □ The height above the targets was kept approximately 0.5 metres. The resulting FOV was 22 cm in diameter

In order to take nadir samples of shrubs and small trees, the fibre optic was mounted on a swinging head, which was itself fitted to the end of a pole. This ensured the nadir view of the probe and proved to be a valuable means of collecting spectral data of taller objects in the field.

As capture date and time were contained in the metadata as well as in the creation time of the binary files, no additional logs had to be kept to keep track of the field data collection process.

In some cases the capture of leaf litter, soil or other vegetation could not be avoided due to the sparse foliage structure of some species, e.g. Manuka (*Leptospermum scoparium).*

The number of spectra captured per site varied slightly with the size or variation exhibited by the target plant, i.e. more samples were taken from some larger objects to describe them more thoroughly.

3.1.6 Species

Spectra of a total of 39 different species were collected (see Table 2). The species assembled at this point are by no means sufficient to describe the variety found in New Zealand bush. However, as a first step the number and variety collected suffices for the purpose of assessing the spectral separability and classification of New Zealand native vegetation.

Table 2: Collected species

Latin name	Common name	Maori name	No of spectra
Agathis australis	Kauri	Kauri	18
Brachyglottis repanda	Rangiora	Rangiora	15
Chionochloa rubra	Red tussock		10
Coprosma robusta	Karamu	Karamu	33
Cordyline australis	Cabbage tree	Ti kouka	31
Cordyline indivisa	Mountain cabbage tree	Toii	9
Cortaderia richardii	Toetoe	Toetoe	27
Corynocarpus laevigatus	Karaka	Karaka	26
Cyathea dealbata	Silver fern	Ponga	35
Cyathea medullaris	Black tree fern	Mamaku	42
Dacrycarpus dacrydioides	White Pine	Kahikatea	20
Dacrydium cupressinum	Red Pine	Rimu	20
Dicksonia squarrosa	Rough Tree Fern	Wheki	19
Dracophylum subulatum	Monoao	Monoao	9
Gleichenia dicarpa var. alpina	Tangle fern	Waewaekaka	18
Griselinia littoralis	Broadleaf	Papauma	18
Halocarpus biformis	Pink pine, yellow pine		27
Hebe stricta	Koromiko	Koromiko	52
Hedycarya arborea	Pigeonwood	Porokaiwhiri	23
Knightia excelsa	New Zealand honeysuckle	Rewarewa	21
Leptospermum ericoides	Kanuka	Kanuka	10
Leptospermum scoparium	Manuka	Manuka	73
Libocedrus bidwillii	Kaikawaka	Kaikawaka, Pahautea	21
Macropiper excelsum	Kawakawa	Kawakawa	43
Melicytus ramiflorus	Whiteywood	Mahoe	60
Metrosideros excelsa	Pohutukawa	Pohutukawa	40
Metrosideros robusta	Rata	Rata	11

Myoporum laetum	Ngaio	Ngaio	48
Myrsine australis	Mapou	Mapou	45
Nothofagus menziesii	Silver Beech	Tawhai	27
Nothofagus solandri	Mountain beech	Tawhairauriki	26
Nothofagus truncata	Hard beech	Tawhairaunui	37
Olearia paniculata	Akiraho	Akiraho	10
Phormium tenax	New Zealand flax	Harakeke	45
Phylocladus alpinus	Mountain toatoa	Toatoa	18
Pimelea buxifolia	Tall pinatoro	Pinatoro	18
Pittosporum eugenioides	Lemonwood	Tarata	58
Podocarpus totara	Totara	Totara	42
Pseudopanax arboreus	Five-finger	Puahou	10

3.2 Spectral Database

3.2.1 Spectral Database Model

This section describes the entities that make up the spectral database model. For an overview of this model showing all entities and their relations please refer to Figure 12.

The spectral database was designed as a relational database. The presented table structure is in third normal form.

The database was primarily designed to hold spectral data of vegetative studies. Therefore it started with a simple structure that could hold spectral data sorted into sites and species. The presented model was iteratively developed during the study, mainly driven by upcoming requirements.

The desired feature list of a spectral database according to the requirements identified in this study is as follows:

- □ Implements the same hierarchical structure as used for the field data to store species, site and spectrum data
- □ Multiple studies: can hold spectral data of different field/laboratory campaigns
- □ Reflectance storage: stores the reflectance data in the database in its original form
- □ Processing parameters: holds parameters that are needed for the processing of the data
- □ Statistics: holds 1st and 2nd order statistics to enable classification, discriminant analysis and separability measurements to be carried out efficiently

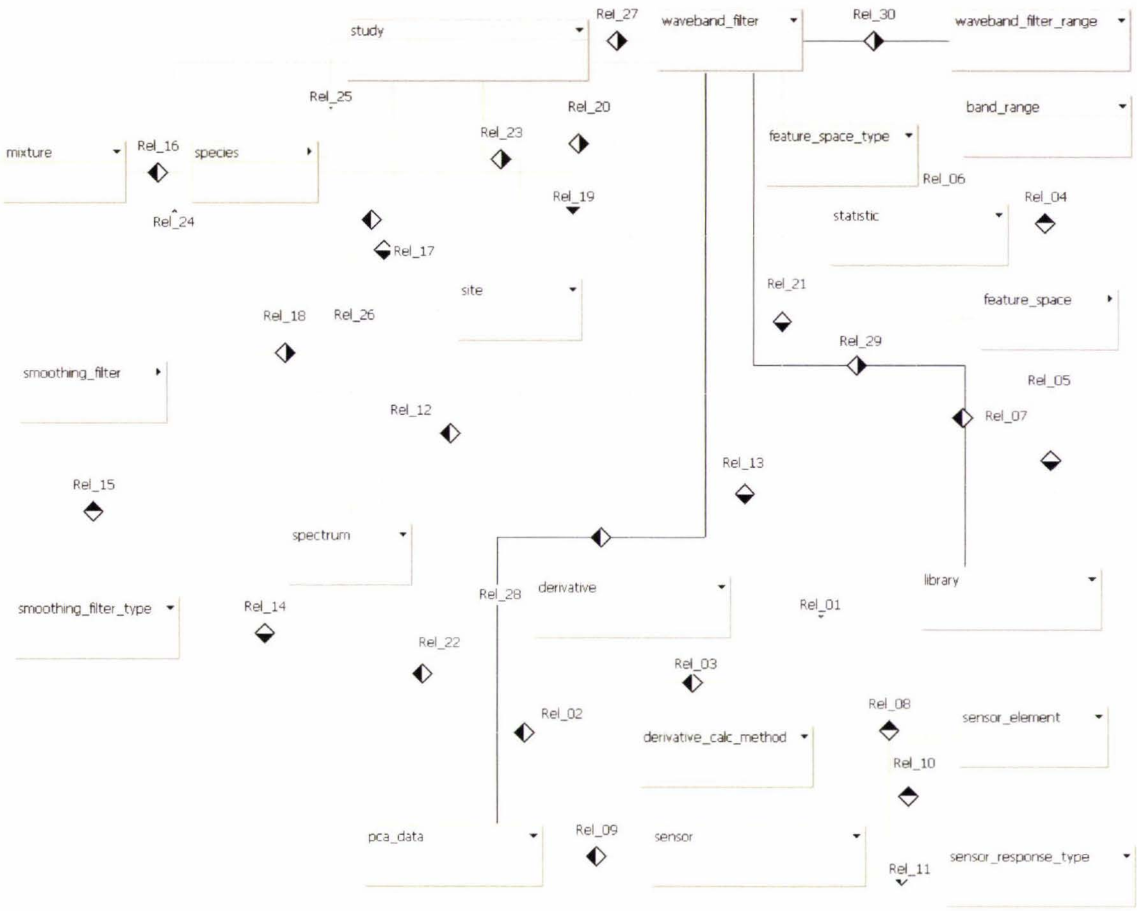


Figure 12: Database model overview at entity level

3.2.1.1 Study, Species, Site and Spectrum Entities

The entities species, site and spectrum reflect the hierarchical structure that was introduced previously (see 3.1.4). The study entity was added to the top of this structure to enable the storage of data belonging to different studies in the same database (Figure 13).

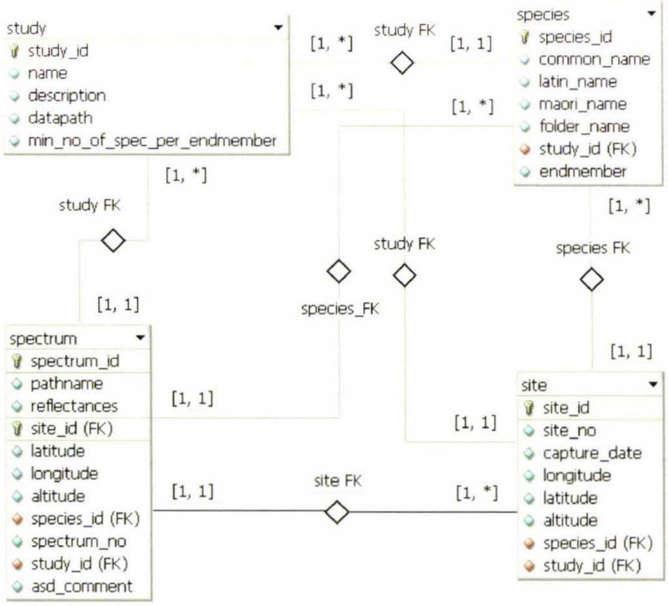


Figure 13: ERD of the entities study, species, site and spectrum

Study attributes	
Attribute	Description
study_id	Primary key
name	Name of the study
description	Description of the study
datapath	Path to the directory that holds the species folders of this study. This directory is the start of the hierarchical data structure. The datapath is used to automatically read the spectra into the database.
min_no_of_spec_per_endmember	This number defines how many spectra a species needs as a minimum to be included in the creation of statistics. The reason for this is that the covariance does not describe the shape of the data adequately enough if only a few samples are used in its calculation.

Species attributes	
Attribute	Description
species_id	Primary key
common_name	The common, i.e. English name of the species
latin_name	The latin, i.e. scientific name of the species
maori_name	The maori, i.e. native name of the species
folder_name	Name of the physical folder that holds spectral data of this species
endmember	A boolean value. This facilitates data export if only endmembers are to be exported. It also is used in spectral mixture studies to designate the endmembers.
study_id	Reference to the study this species belongs to

Site attributes	
Attribute	Description
site_id	Primary key
site_no	The number of this site
capture_date	Date when the site was captured
longitude	Longitude of the spatial position of this site
latitude	Latitude of the spatial position of this site
altitude	Altitude of the spatial position of this site
study_id	Reference to the study this site belongs to
species_id	Reference to the species this site belongs to

Spectrum attributes	
Attribute	Description
spectrum_id	Primary key
pathname	The full pathname of the binary ASD file
reflectances	The reflectance data stored as binary object
latitude	Longitude of the spatial position of this spectrum
longitude	Latitude of the spatial position of this spectrum
altitude	Altitude of the spatial position of this spectrum
spectrum_no	The number that is auto-assigned to this spectrum by the ASD controller software
asd_comment	User comment as entered in the ASD controller software
study_id	Reference to the study this spectrum belongs to
species_id	Reference to the species this spectrum belongs to
site_id	Reference to the site this spectrum belongs to

3.2.1.2 Waveband_filter and Waveband_filter_range

The waveband_filter and waveband_filter_range entities hold data that are needed for the removal of noisy or uncalibrated bands from the spectra. They were defined at the study level because every study might have different requirements for the data filtering (Figure 14). E.g. a study that contains data collected by a contact probe will not need to remove water bands as the influence of the atmosphere is practically non existent. Similarly, if a study wishes to concentrate on a certain part of the spectrum only, the unused wavebands can be removed by entering them into the filter structure. The design is thus able to accommodate not only vegetation data collected under field conditions with solar illumination but can deal with contact probe data as well.

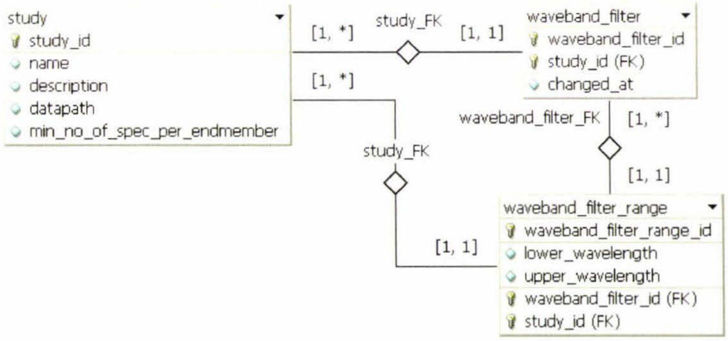


Figure 14: ERD of the entities study and waveband_filter and waveband_filter_range

Waveband_filter attributes	
Attribute	Description
waveband_filter_id	Primary key
changed_at	The date when this filter was last modified
study_id	Reference to the study this filter belongs to

Waveband_filter_range attributes	
Attribute	Description
waveband_filter_range_id	Primary key
lower_wavelength	The wavelength in nanometres where the filter starts
upper_wavelength	The wavelength in nanometres where the filter ends
waveband_filter_id	Reference to waveband_filter
study_id	Reference to the study this filter range belongs to

3.2.1.3 Library, Statistic, Feature Space, Sensor, Smoothing, Derivative and associated tables

The library can be thought of as a collection of data that is needed to look up unknown signatures. A library is built for certain settings of the data processing chain, namely waveband filtering, smoothing, sensor convolution, derivative calculation and feature space transformation. The resulting library is setup for classification of data that is processed in exactly the same way. In other words, before a classification can be carried out on a dataset, its library must be built.

A library therefore references the entities smoothing_filter, sensor, derivative and feature_space (see Figure 15). For a library to be valid its build date must be newer than the dates of modification of the entities waveband_filter, smoothing_filter, derivative and feature_space.

The actual data needed for a classification is held in the statistic entity in form of a mean vector and a covariance matrix for every species.

The smoothing_filter entity holds data needed for the smoothing by a Savitzky-Golay filter.

The sensor entity contains data for the synthesizing of sensor responses. Two general classes of sensors exist, defined by the description of the response type of their elements:

1. Gaussian: each sensor element response is modelled by a Gaussian function. The Gaussian curve is defined by the average wavelength and the full width at half the maximum (FWHM).
2. Ratio: each sensor element response is modelled by ratios applied to narrow band data over a certain range of wavelengths.

The entity sensor_element holds both Gaussian and Ratio settings, depending on the type of sensor. In the case of Gaussian sensors, one sensor_element entry describes one sensor band. For Ratio sensors, many sensor_element entries may be needed to describe one sensor band.

The derivative entity holds data for the calculation of derivatives either by an iterative, finite difference method or by Savitzky-Golay coefficients.

The feature_space entity holds or refers to data needed for the feature space transformation.

A feature space belongs to a type of feature space. The type of feature space defines the way in which the transformation is calculated.

Three types of feature space were considered to be useful, although more possibilities exist:

1. Derivative Indices (DI): a feature space is formed by calculating several DIs. The band ranges for these indices are held in the band_range entity.
2. Normalized Two Band Indices (NTBI): a feature space is formed by calculating several NTBIs. The two bands that define one index are held in the band_range entity
3. PCT: a feature space is formed by calculating a certain number of components. The transformation matrix is held in the pca_data entity. The number of components to be calculated is equal to the dimension of the feature space.

Similar to the library, the pca_data is calculated for a certain setup of waveband_filter, smoothing, sensor synthesizing and derivative calculation.

Library attributes	
Attribute	Description
library_id	Primary key
build_date	The date when this library was last compiled
feature_space_id	Reference to feature space used when building library
smoothing_filter_id	Reference to smoothing filter used when building library
sensor_id	Reference to sensor used when building library
derivative_id	Reference to derivative used when building library
waveband_filter_id	Reference to waveband filter used when building library
study_id	Reference to study this library belongs to

Statistic attributes	
Attribute	Description
statistic_id	Primary key
no_of_samples	Number of samples that were used in the statistic calculation
mean	The mean vector stored as binary object
cov	The covariance matrix stored as binary object
library_id	Reference to library this statistic belongs to
species_id	Reference to species this statistic belongs to

Feature_space attributes	
Attribute	Description
feature_space_id	Primary key
fs_type_id	Reference to the type of feature space
dimension	The dimension of the feature space
name	Name of this feature space
description	Description
build_date	Date when feature space was created or modified

Feature_space_type attributes	
Attribute	Description
fs_type_id	Primary key
name	Name of this feature space type (i.e. DI, NTBI or PCT)
type	A numeric coding for the type. Identical to the numbers used in the processing software.

PCA_data attributes	
Attribute	Description
pca_data_id	Primary key
eigenvectors	The eigenvector matrix of the principal components analysis stored in binary format
eigenvalues	The eigenvalue matrix of the principal components analysis stored in binary format
dim	The dimension of the above matrices
build_date	Date when the eigenanalysis was carried out
smoothing_filter_id	Reference to smoothing filter used when performing PCA
sensor_id	Reference to sensor used when performing PCA
derivative_id	Reference to derivative used when performing PCA
waveband_filter_id	Reference to waveband filter used when performing PCA
study_id	Reference to study on which PCA was performed

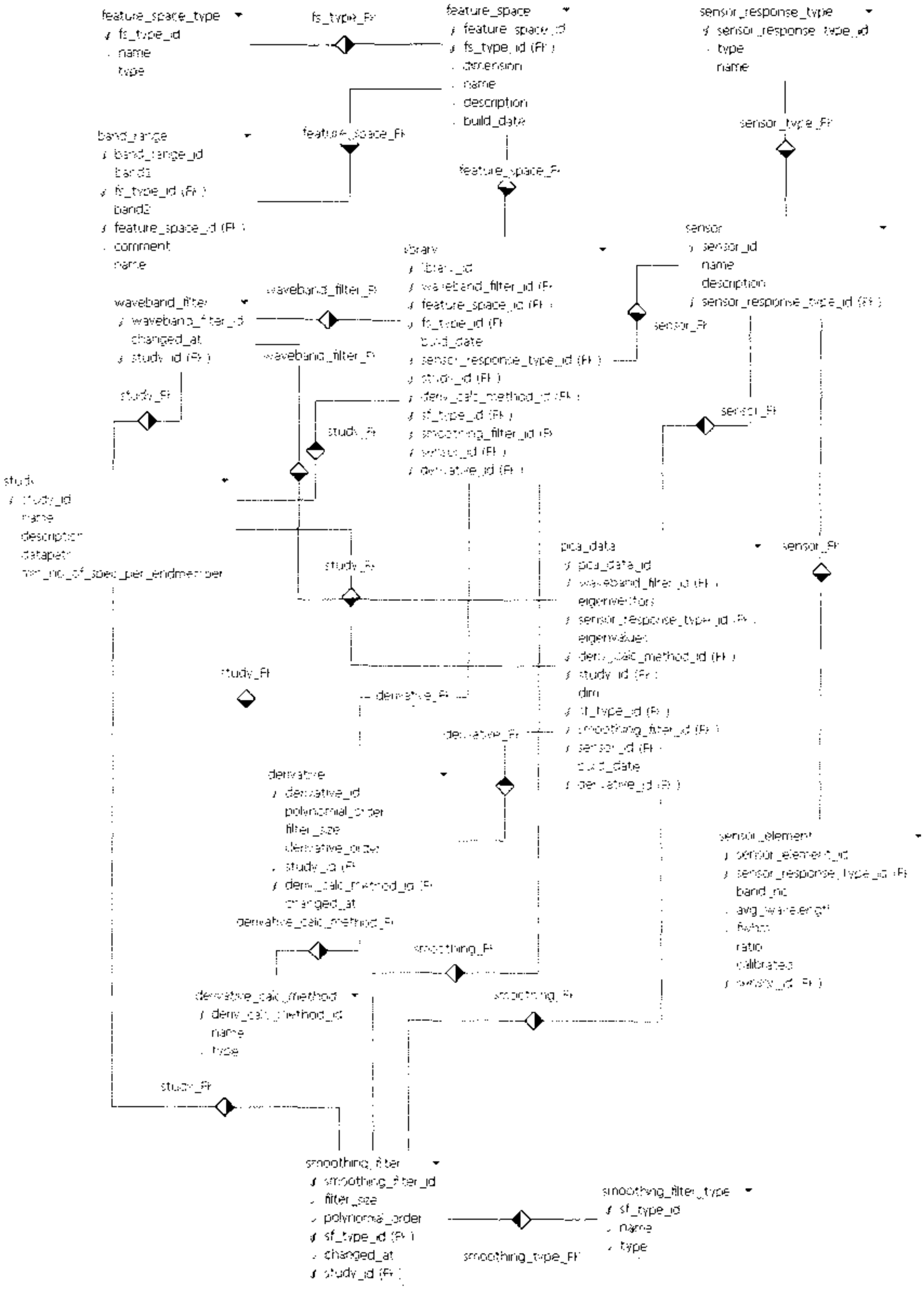


Figure 15: ERD of library, statistic, feature_space, sensor and associated entities

Band_range attributes	
Attribute	Description
band_range_id	Primary key
band1	First sensor band to be used for NTBI or start band of DI band range
band2	Second sensor band to be used for NTBI or end band of DI band range
comment	Free user comment on this band range
name	Name of this band range. This is used as column name when exporting feature space data.
feature_space_id	Reference to feature space this band range belongs to

Sensor attributes	
Attribute	Description
sensor_id	Primary key
name	Name of the sensor
description	Description of the sensor
sensor_response_type_id	Reference to sensor type

Sensor_response_type attributes	
Attribute	Description
sensor_response_type_id	Primary key
type	A numeric coding for the type. Identical to the numbers used in the processing software.
name	Name of this sensor response type (i.e. Gaussian or Ratio)

Sensor_element_attributes	
Attribute	Description
sensor_element_id	Primary key
band_no	Band number of the sensor
avg_wavelength	The average wavelength of the sensor element (for Gaussian sensors) or the wavelength of the input band to be ratio-ed.
fwhm	Full width at half the maximum. Essentially defines the shape of the Gaussian response curve. Only for Gaussian sensors.
ratio	The ratio to be applied to the input band (defined by the avg_wavelength). Only for ratio sensors.
calibrated	Boolean, defines if the band is calibrated or not. Uncalibrated bands will not be used in the processing. Some sensors have certain bands defined as uncalibrated (e.g. Hyperion) and it may be desired to store this information in the database.
sensor_id	Reference to sensor this element belongs to

Derivative attributes	
Attribute	Description
derivative_id	Primary key
polynomial_order	Polynomial order (for Savitzky-Golay derivative calculation only)
filter_size	Size of the filter (for Savitzky-Golay derivative calculation only)
derivative_order	Order of derivative
changed_at	Date when this derivative setup was changed
study_id	Reference to study this derivative setup belongs to
deriv_calc_method_id	Reference to the calculation method

Derivative_type_method attributes	
Attribute	Description
deriv_calc_method_id	Primary key
type	A numeric coding for the type. Identical to the numbers used in the processing software.
name	Name of this calculation method

Smoothing_filter attributes	
Attribute	Description
smoothing_filter_id	Primary key
filter_size	Size of the filter
polynomial_order	Polynomial order
changed_at	Date when this smoothing filter was changed
sf_type_id	Reference to filter type
study_id	Reference to study this smoothing filter belongs to

Smoothing_filter_type attributes	
Attribute	Description
sf_type_id	Primary key
type	A numeric coding for the type. Identical to the numbers used in the processing software.
name	Name of this filter type

3.2.1.4 Mixture

The mixture entity (Figure 16) is used to describe mixtures where the abundance is known, such as in laboratory experiments. The abundance settings are then used to display error statistics after the unmixing process. Several entries in the mixture entity are needed to describe a mixture, e.g. if a mixture consists of three endmembers, then three mixture records are required to describe the mixture.

Mixture attributes	
Attribute	Description
mixture_id	Primary key
abundance	The fractional abundance of the endmember in this species
endmember_id	The species_id of the endmember
species_id	The species_id of the mixture.

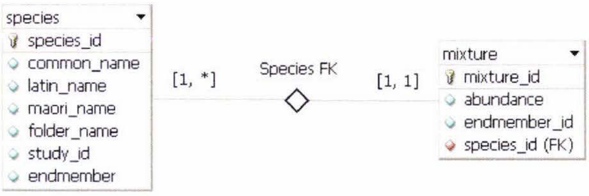


Figure 16: ERD of the entities species and mixture

3.2.2 Spectral Database Implementation

The database was implemented in MySQL (MySQL AB, 2005), a GNU open source software. MySQL is a relational database management system that can handle large amounts of data, allows data access via standard SQL commands, provides multi-user access over TCP/IP and supports several APIs (Application Programming Interfaces) amongst which is C/C++.

3.3 A Spectral Data Management and Processing Software

A spectral database as described above is not of much use on its own. Data must be fed into the database and data extraction routines must exist in order to exploit the benefits of a spectral database. The technical requirements for such a system were identified as follows:

- □ Graphical user interface to the database
- □ Functions for loading spectral data into the database
- □ Data pre-processing functions
- □ Data analysis functions
- □ File export functions to allow data analysis and plotting in 3rd party packages

The resulting software was called SpectraProc. The software architecture is described in section 3.3.2, the concepts and algorithms used in the spectral data processing and analysis functions are described in the sections 3.4 and 3.5. For a screenshot of the graphical user interface and according description please refer to the Appendix.

3.3.1 Programming Language, Libraries and Environment

The software was developed for the Microsoft Windows environment using Microsoft Visual C++ V6.0. The graphical user interface was based on Microsoft Foundation Classes (MFC), using a simple Document-View architecture with one document and one associated view. MySQL C API (Application Programmer Interface) was used for the database access from C++ code. Matrix calculations were based on the excellent C++ matrix library NewMat V10B (Davies, 2002) which is available freely on the internet.

3.3.2 Software Architecture

3.3.2.1 File System Interfaces

SpectraProc provides input and output interfaces to the file system (see Figure 17). Input file formats are: ASD binary file as produced by the ASD FieldSpecPro Spectroradiometer, ENVI Z-Profiles that are signatures extracted from hyperspectral imagery in ENVI and sensor specifications in a proprietary, tabulator separated format. ASD files can be imported into the database as part of a study or loaded into memory for classification against a study dataset. ENVI Z-Profiles can be loaded for classification only. Sensor specification files are a way of defining new sensors in the database.

Output can be written in three data formats: (1) CSV (Comma Separated Values) for import into various 3rd party applications like spreadsheets or statistic packages, (2) ENVI Spectral Library for import into ENVI and subsequent use for e.g. signature matching and (3) ARFF which is a special format used by WEKA (University of Waikato, 2005). WEKA is a collection of machine learning algorithms for data mining tasks.

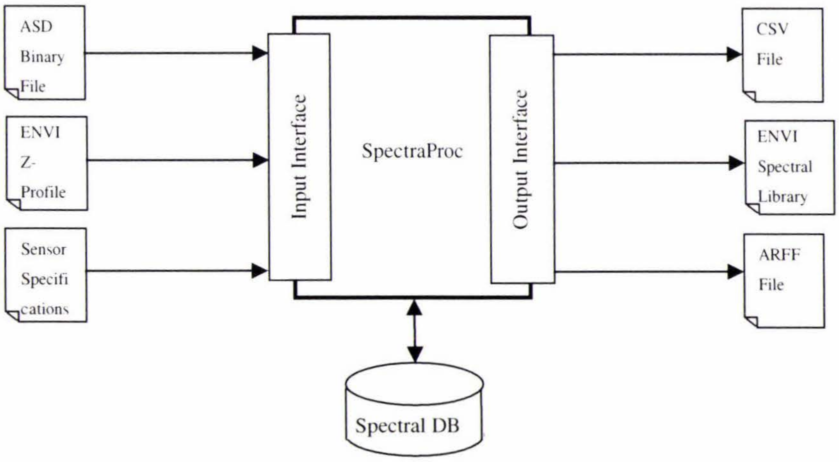


Figure 17: File system interfaces

3.3.2.2 Class Overview

SpectraProc was designed as an object oriented program. Many of the SpectraProc classes were derived from MFC classes as they form part of the graphical user interface.

Table 3 lists all classes derived from MFC including a short description of the purpose. The non-MFC classes are described in Table 4 and graphically presented as an UML (Unified Modelling Language) diagram in Figure 18.

Table 3: Short description of MFC derived classes

Class Name	Derived from	Description
CSpectraProcDoc	CDocument	The document of the document-view architecture. Holds the runtime objects of: library, spectra_factory and spec_proc_data
CSpectraProcView	CFormView	The main form of the application. Manages all dialogs and handles Windows messages.
Dlg_abundance_setting_class	CDialog	Defines the known abundances of endmembers in known mixtures
Dlg_accuracy_check_class	CDialog	Selection dialog to choose another study to be used as independent dataset
Dlg_endmember_selection_class	CDialog	Used to set the endmembers in a given dataset
Dlg_feature_space_edit_class	CDialog	Create and modify feature space definitions
Dlg_file_export_class	CDialog	Choices for file export
Dlg_filterband_def_class	CDialog	Defines a lower and upper wavelength for a waveband filter range
Dlg_import_sensor_class	CDialog	Import dialog for sensor files
Dlg_new_study_class	CDialog	Creation of new studies
Dlg_progress_class	CDialog	Progress bar, used by several processes
Dlg_site_accuracy_class	CDialog	Dialog to select a species for classification accuracy check on site level
Dlg_Smoothing_filter_settings_class	CDialog	Set the smoothing parameters for Savitzky-Golay filters
Dlg_Waveband_Def_class	CDialog	Defines two wavebands, used for entering new indices or waveband regions for feature spaces
Dlg_waveband_filter_setup_class	CDialog	Creation/Modification of waveband filters

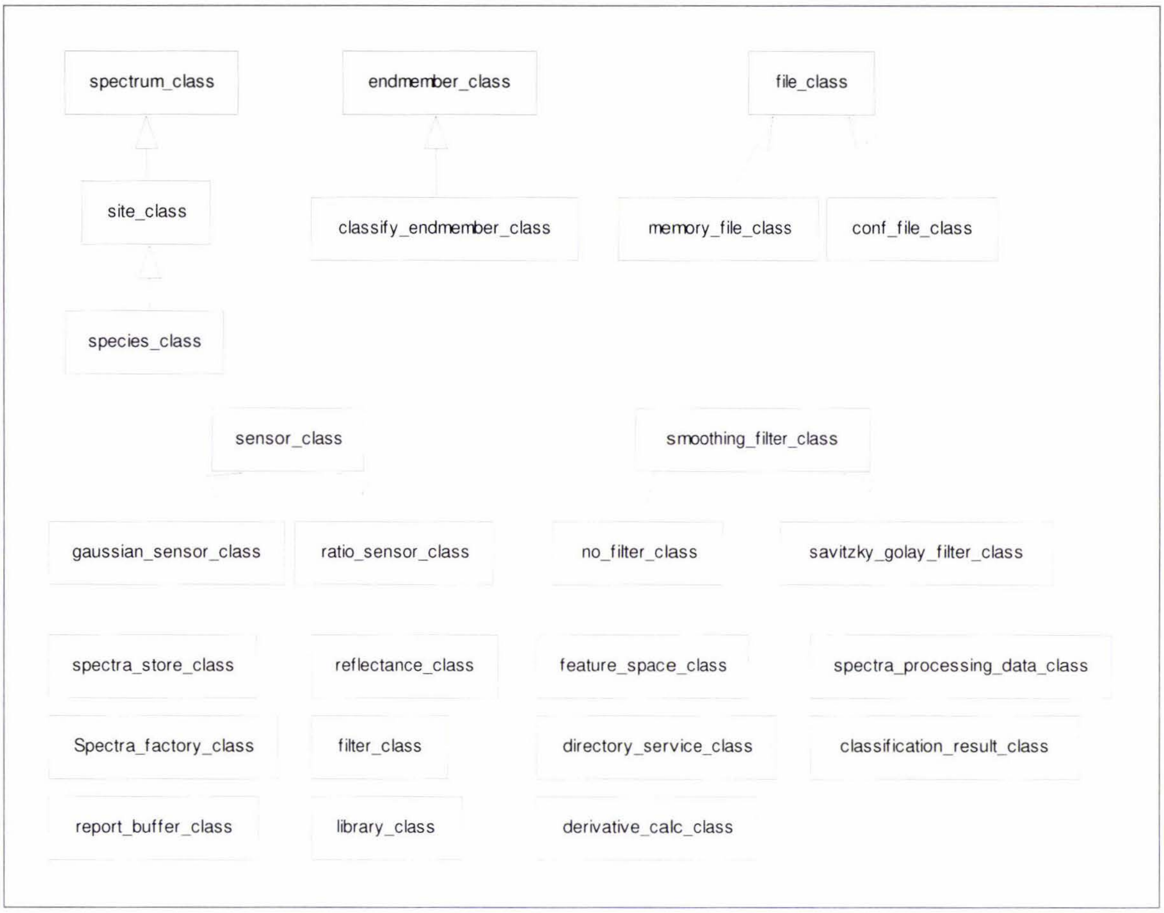


Figure 18: Non-MFC classes

Table 4: Short description of non-MFC derived classes

Class Name	**Derived from**	**Description**
classification_result_class	-	Storage and manipulation of classification results
endmember_class	-	Represents an endmember with a mean vector and a covariance matrix
classify_endmember_class	endmember_class	Used as endmember in classifications
conf_file_class	file_class	Reads and writes configuration files
derivative_calc_class	-	Calculate derivatives
directory_service_class	-	Creates lists of files and subdirectories of a file system directory
feature_space_class	-	Represents a feature space with its settings loaded from database
file_class	-	File input and output
filter_class	-	Class for waveband filtering
gaussian_sensor_class	sensor_class	Represents a sensor with elements of Gaussian response function

Class Name	**Derived from**	**Description**
library_class	-	Holds functions for: classification, separability report, building spectral libraries, eigenanalysis and spectral unmixing
memory_file_class	file_class	Allows the transposing of structured files by holding the file structure in memory before writing to a file.
no_filter_class	smoothing_filter_class	Performs no smoothing but copies the data directly into the next reflectance structure
ratio_sensor_class	sensor_class	Represents a sensor with elements of ratio response function
reflectance_class	-	Holds spectral data: band number, reflectance and average wavelength
report_buffer_class	-	Buffer class for handling the text output in the main window
savitzky_golay_filter_class	smoothing_filter_class	Smoothes the data using Savitzky-Golay coefficients
sensor_class	-	Base class for Gaussian and ratio sensors
site_class	spectrum_class	Represents a site
smoothing_filter_class	-	Base class for smoothing filter types
species_class	site_class	Represents a species
Spectra_factory_class	-	Used for loading spectra from files or database, inserting into the database and outputting data to text files
spectra_processing_data_class	-	Central data pool for processing settings such as: waveband filter, smoothing filter, sensor, derivative calculator and diverse waveband filters for the removal of smoothing or derivative artefacts.
spectra_store_class	-	A list that holds spectra. Used when spectra are loaded directly from file into memory.
spectrum_class	CObject	Represents a spectrum

3.3.2.3 Spectral Processing Concept

The spectral database only stores the raw spectral data. Further processing of the spectra is performed at runtime and the results are held in memory. Once a spectrum is loaded from the database it is put through a cascade of operations as shown in Figure 19. The result of every stage is saved in a separate data structure in memory. These data structures and processing functions are attributes or methods respectively of every object of the spectrum class. An instance of the spectrum class offers a method that returns the data of a certain stage of processing and will internally execute all preceding steps needed for that stage. This allows the easy file export of spectral data at any processing step.

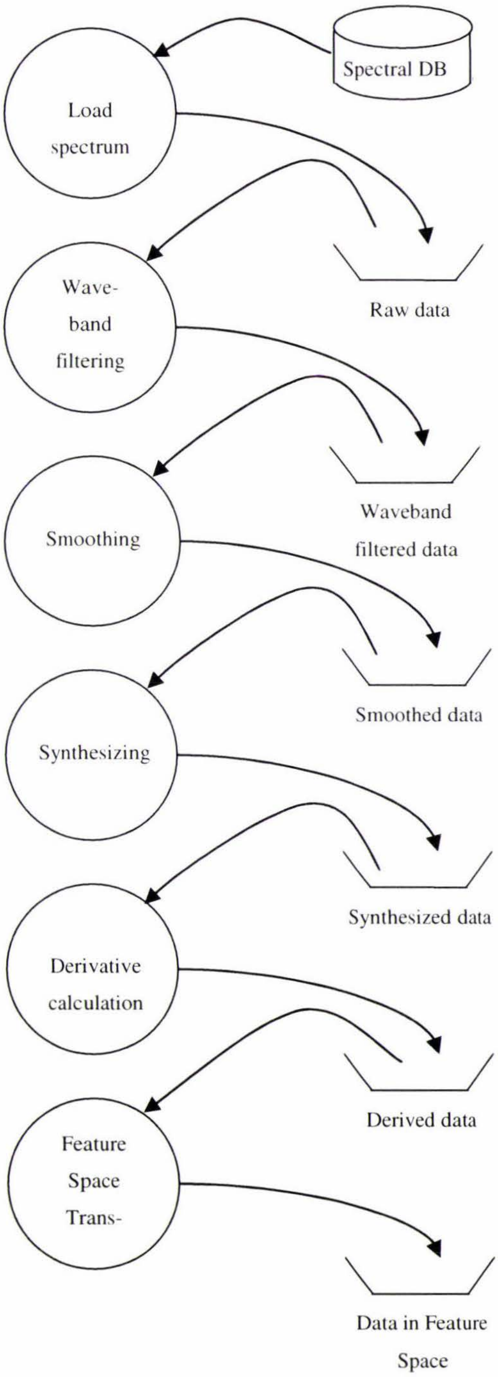


Figure 19: Spectral data processing cascade

3.4 Data Processing

The data processing was divided into the following stages:

1. Waveband Filtering
2. Smoothing
3. Sensor Synthesizing / Downsampling
4. Derivative Calculation
5. Feature Space Transformation

The underlying algorithms of these stages are described hereafter.

3.4.1 Waveband Filtering

The data in the following band ranges were seriously affected by atmospheric absorbtion and had to be removed from the spectra: 1350-1440 nm, 1790-1980 nm and 2360-2500 nm (see Figure 20).

Technically this was done by setting the reflectance values in the filtered regions to -1. The later processing steps then just ignored these values.

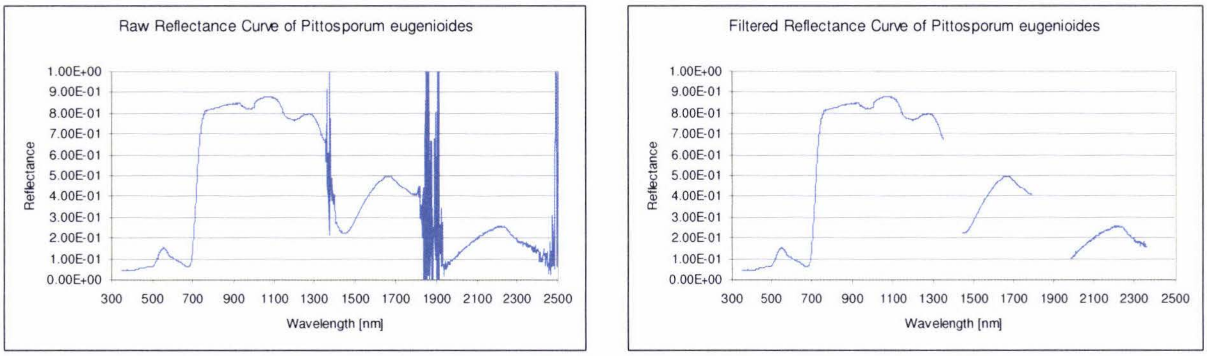


Figure 20: An example of pre and post filtering of noise bands

3.4.2 Smoothing

The Savitzky-Golay filter was chosen because of its reported good performance and the relatively simple algorithm involved. As mentioned in the review of smoothing methods, the filter coefficients are calculated at run time instead of read from lookup tables. The chosen implementation is based on Press et al. (2002).

In a first step a design matrix is created that holds the polynomial equations:

$$A_{ij} = i^j \qquad i = -n_L, \ldots, +n_R \qquad j = 0, \ldots, M$$

where

n_L, n_R = left hand, right hand filter size

M = polynomial order

The coefficients are then calculated by

$$c_i = \left\{ \left(A^T \cdot A\right)^{-1} \right\}_0 \cdot \vec{n}$$

where

$\vec{n}$ = vector with elements $n_j = i^j \qquad i = -n_L, \ldots, +n_R \qquad j = 0, \ldots, M$

Note that for the calculation of the coefficients only the first row of the inversed matrix is used.

Two possibilities exist for the smoothing of the data using the Savitzky-Golay coefficients: convolution by a moving window filter or by multiplication in frequency space.

The moving window function calculates the smoothed value for every band by:

$$Y_j^* = \frac{\sum_{i=-m}^{i=+m} C_i Y_{j+i}}{N}$$

where

Y_j^* = smoothed data point

C_i = convolution coefficient

Y_{j+i} = original data point

N = moving window size (-m...+m)

It can be shown that a convolution in time space is equal to a multiplication in frequency space. Fast Fourier Transformation (FFT) of both the smoothing function and the signal transforms them into frequency space.

$s * r = S \cdot R$

where

s = signal

r = response function (smoothing function)

S, R = FFT(s) resp. FFT(r)

The smoothed signal in time space is then the inverse FFT:

$s_{smooth} = invFFT(S \cdot R)$

In both cases the result must be filtered to remove artefacts that appear at the start and end of every valid waveband segment (Figure 21). The new valid segment sizes are calculated by:

$\lambda u_{smoothed} = \lambda u - pos\_filter\_size$

$\lambda l_{smoothed} = \lambda l - neg\_filter\_size$

where

λl, λu = lower and upper segment wavelengths

Thus every segment looses information of filter_size – 1.

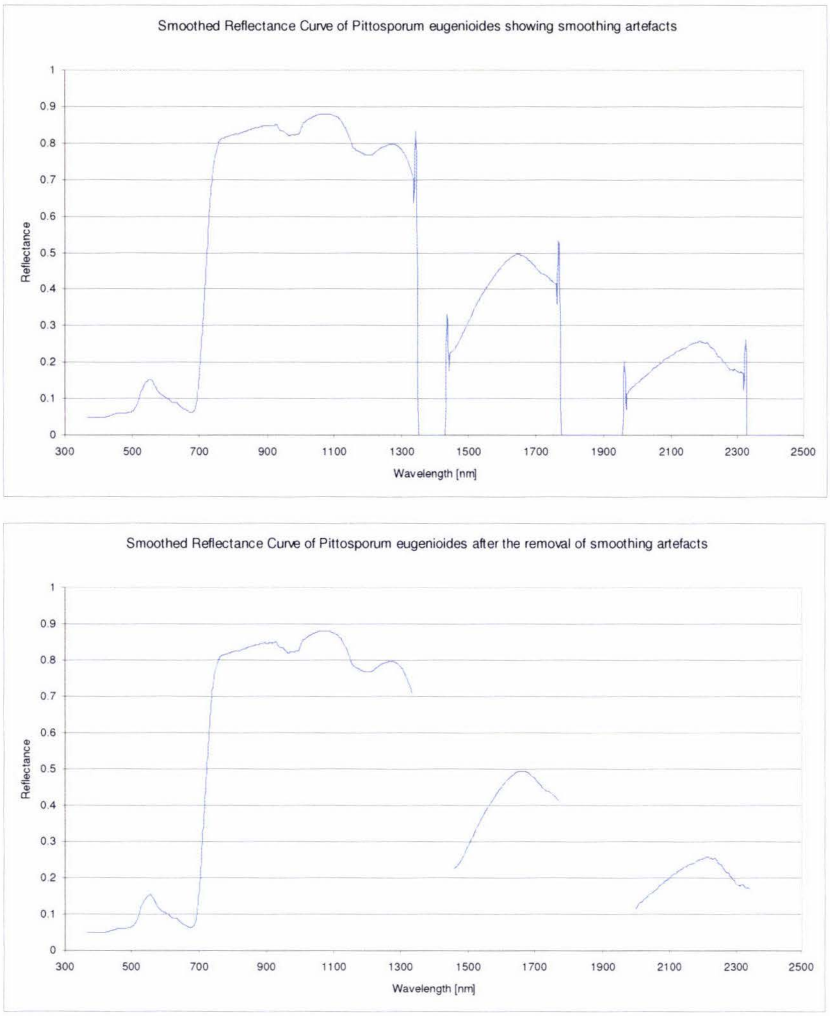


Figure 21: A smoothed signature of Pittosporum eugenoides before and after the removal of smoothing artefacts

3.4.3 Synthesizing of other Sensors

The synthesizing of other sensor responses using ASD data is useful due to several reasons:

1. Reduction of dimensionality
2. Direct comparison of airborne/spaceborne sensor and ground data
3. Implicit smoothing of the data
4. Prediction and assessment of the usefulness of a certain sensor

The synthesizing of other sensor bands is also called band convolution. The process used is principally a convolution operation as described in 2.3.2.4. A filter is moved over the data and used to calculate the band values of the sensor to be synthesized. The process of spectral band synthesis is based on the algorithm used by Zanoni (2002).

The simulation of Hyperion and Landsat7 ETM+ were of interest for this specific research. These sensor types can however be generalized and thus a generic synthesizing operation can be designed that allows the simulation of any sensor that falls into the following two classes: Ratio and Gaussian.

3.4.3.1 Ratio Sensors

The sensor element function of these sensors is modelled by a number of known coefficients, thus the synthesizing operation is simply a convolution of a defined wavelength region using these coefficients (ratios).

An example of such ratios is shown for Landsat7 ETM+ band 1 (Figure 22). The ratios for Landsat7 used in this study were made available by Dr J. Shepherd of Landcare Research. The ratios were given at 1nm steps, thus they could be directly applied to the ASD band reflectances.

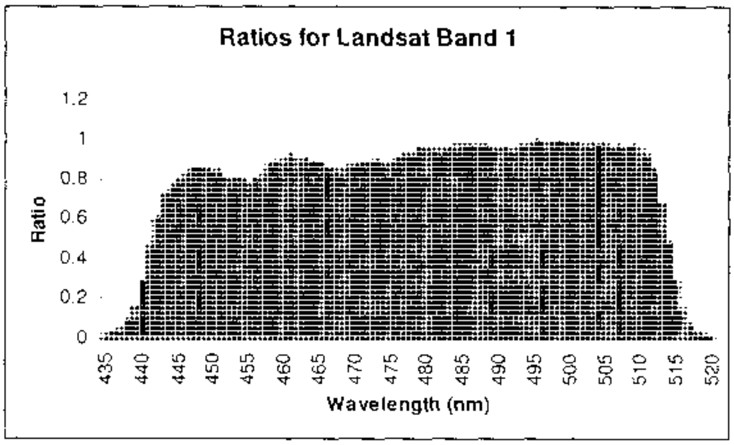


Figure 22: Ratios for Landsat7 ETM+ band 1

The convolution is calculated by:

$$r_j = \frac{\sum_{i=lw\_j}^{uw\_j} c_i \cdot r_i}{\sum_{i=lw\_j}^{uw\_j} c_i}$$

where

r_j = the synthesized reflectance value of the j-th synthesized band

c_i = the coefficient for wavelength i

r_i = reflectance value of i-th ASD band

lw_j = lower wavelength of the j-th band

uw_j = upper wavelength of the j-th band

3.4.3.2 Gaussian Sensors

The sensor element response function of these sensors is best approximated by a Gaussian function. The sensor elements are technically defined by the middle wavelength and the full width at half the maximum (FWHM) (see Figure 23).

The Gaussian function is defined by:

$$f(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where

σ = standard deviation

μ = mean value

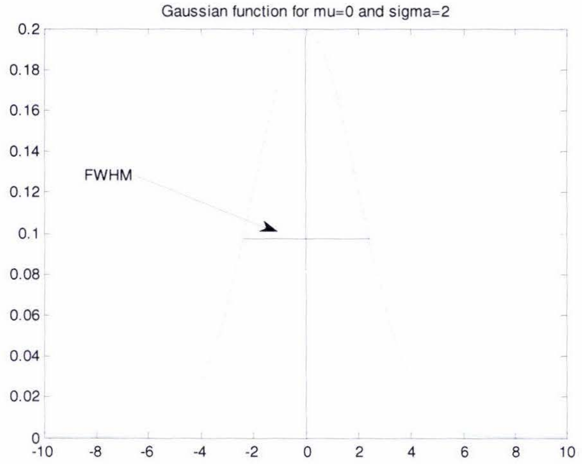


Figure 23: Gaussian curve illustrating the FWHM measure

The maximum of a Gaussian function is always at the mean value μ and the function is symmetrical to the mean. Thus for x = μ the Gaussian function becomes:

$$f(\mu) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma}} \cdot e^{-\frac{1}{2}\left(\frac{\mu-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma}}$$

The curve becomes more sharply defined for smaller values of σ and wider for bigger values of σ.
The standard deviation σ can be calculated from the FWHM as described hereafter.
As the Gaussian curve assumes half the maximum at the points defined by the FWHM, the function can be written as:

$$f(\mu) = 2 \cdot f\left(\mu \pm \frac{FWHM}{2}\right) = 2 \cdot f(\mu \pm d)$$

$$\frac{1}{\sqrt{2 \cdot \pi \cdot \sigma}} = 2 \cdot f(\mu \pm d)$$

where d = FWHM/2

As the curve is symmetric, the above equation can be solved for either x = μ + d or x = μ – d.

$$\frac{1}{\sqrt{2\cdot\pi\cdot\sigma}} = 2\frac{1}{\sqrt{2\cdot\pi\cdot\sigma}}\cdot e^{-\frac{1}{2}\left(\frac{(\mu+d)-\mu}{\sigma}\right)^2}$$

$$\frac{1}{2} = e^{-\frac{1}{2}\left(\frac{d}{\sigma}\right)^2}$$

$$\ln(0.5) = -\frac{1}{2}\left(\frac{d}{\sigma}\right)^2$$

$$\sigma^2 = -\frac{1}{2}\frac{(d)^2}{\ln(0.5)}$$

$$\sigma = -0.8493218003\cdot\mathrm{d}, 0.8493218003\cdot\mathrm{d}$$

The coefficients used for the convolution operation are given by the Gaussian function:

$$c_i = f(wavelength\_band\_i) = \frac{1}{\sqrt{2\cdot\pi\cdot\sigma}}\cdot e^{-\frac{1}{2}\left(\frac{wavelength\_band\_i-wavelenth\_band\_j}{\sigma}\right)^2}$$

where

$wavelength\_band\_i$ = wavelength of the i-th ASD band

$wavelength\_band\_j$ = wavelength of the j-th band of the sensor to be synthesized

c_i = the i-th coefficient for the convolution operation

The band convolution is calculated by

$$r_j = \frac{\sum_{i=\mu-range}^{\mu+range} c_i\cdot r_i}{\sum_{i=\mu-range}^{\mu+range} c_i}$$

where

r_j = the synthesized reflectance value of the j-th synthesized band

c_i = the coefficient determined by the Gaussian function for the wavelength of the i-th ASD band

r_i = reflectance value of i-th ASD band

$range$ = defines the range of values to be used for the band convolution symmetrically to the middle ASD waveband. The middle ASD waveband is the one closest to the average wavelength of the j-th synthesized band.

3.4.3.3 Hyperion

The Hyperion sensor captures data from 400 to 2400 nm with bandwidths of 10nm. This spectral range and resolution of the Hyperion sensor is a generalization. Waveband centres do not lie at whole number frequencies, bandwidths are not sharply defined and the sensitivity of the sensor is not uniform over the bandwidth. The sensor characteristics are available in Microsoft Excel format from the United States Geological Survey (USGS, 2005).

The Hyperion sensor is an example of a Gaussian sensor. The spectral response function of the sensor elements is well approximated with a Gaussian function (Liao and Jarecke, undated).
See Figure 24 for an example of the sensor response functions of two neighbouring sensor elements.

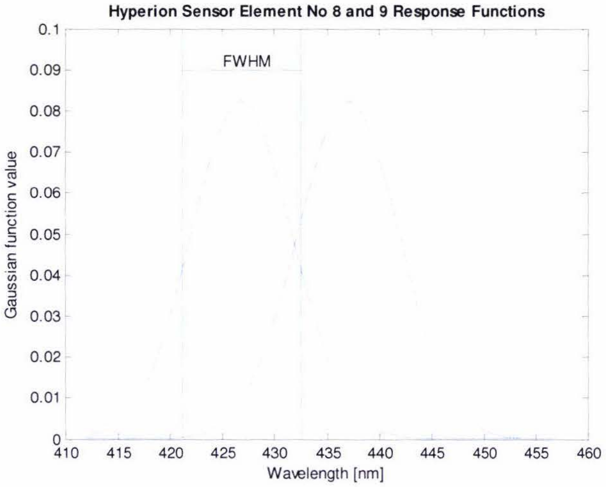


Figure 24: Sensor response functions for Hyperion sensor elements 8 and 9 and the FWHM of band 8

Hyperion synthesizing was used in this study for three purposes: (a) create a spectral library that could be used to classify Hyperion imagery at a later stage, (b) reduce the dimensionality of the data to simplify the data analysis and (c) implicitly remove noise from the spectra by the smoothing effect of the synthesizing operation.
The range of values used for the convolution was set to 3 times the standard deviation: $range = 3 \cdot \sigma$, i.e. 99.74% of all contributing values are used (Papula, 1994).
This range was practically not usable for all bands because some wavelengths had been filtered previously. In these situations, the range was symmetrically reduced to avoid filtered areas.
By convoluting the ASD data to Hyperion-like bands, the dimensionality was reduced by approximately a factor of 10.
The actual synthesizing process was carried out as described under Gaussian Sensors.
The Hyperion band creation resulted in 166 new bands.

3.4.3.4 Downsampling

The downsampling sensor is a hypothetical ratio sensor. Bands are spaced a certain wavelength apart with a ratio of 1. A convolution of data using a downsampling sensor with a band spacing of 10nm results in a downsampling of the data by factor ten, i.e. every tenth waveband is chosen. It may be advisable to apply the downsampling only to smoothed data in order to avoid aliasing. The combination of smoothing and downsampling is called decimation.

3.4.4 Derivative Calculation

Two approaches to the calculation of derivatives were identified.
An explicit calculation of the derivative for a given wavelength by the finite difference method by:

$$\rho'(b_i b_{i+1}) = \frac{\rho(b_{i+1}) - \rho(b_i)}{\lambda(b_{i+1}) - \lambda(b_i)}$$

where

$\rho(b_i)$ = reflectance of band i

$\lambda(b_i)$ = wavelength of band i

$\rho'(b_i b_{i+1})$ = first derivative of linear curve segment between reflectances of band i and band i+1

The n-th derivative is thus calculated by applying the above formula n times.

An implicit calculation is possible by Savitzky-Golay coefficients by simply selecting the n-th row of the inverse matrix and multiplying by n!:

$$c_i = \left(\left\{\left(A^T \cdot A\right)^{-1}\right\}_{order} \cdot \vec{n}\right) \cdot order!$$

However, this method performs automatically a smoothing of the data which may not be needed or wanted after the sensor synthesizing.

Both methods lose n data points per valid segment. The explicit calculation loses one data point per iteration. The Savitzky-Golay filtered data lose data points due to the removal of artefacts. The number of points lost depends on the filter size.

For the calculation of the filter coefficients and correct derivatives the following conditions must be met:

$polynomial\_order \geq derivative\_order$

$filter\_size \geq \max(polynomial\_order + 1, derivative\_order + 1)$

Thus the minimal filter size depends on both the polynomial and the derivative order. A minimal filter size of (derivative_order+1) will result in the removal of n = derivative_order number of points.

3.4.5 Feature Space Transformation

Three types of feature spaces were implemented:

- □ Derivative Indices (DI)
- □ Normalized Two Band Indices (NTBI)
- □ Principal Component Transformation (PCT)

3.4.5.1 DI (DGVI)

DGVIs (Derivative Greenness Vegetation Indices) are examples of DIs. These indices are effectively describing the shape of the reflectance curve. The DGVI calculation was based on the equations used by Thenkabail et al. (2004a) and Elvidge and Chen (1995). The derivatives of the reflectance were computed by using the slopes of linear interpolations between the discrete reflectance band values (Figure 25). The first derivative of reflectance is therefore:

$$\rho'(b_i b_{i+1}) = \frac{\rho_{i+1} - \rho_i}{\lambda(b_{i+1}) - \lambda(b_i)}$$

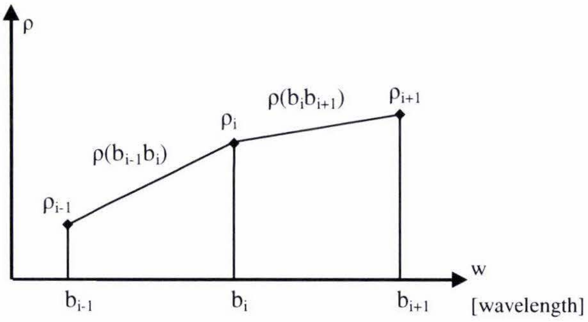


Figure 25: Illustration of discrete reflectance values ρ and interpolated linear curves to form a continuous reflectance curve

The DGVIs were calculated using the equation:

$$DGVI = \sum_{i=m}^{n} \frac{\rho'(b_{i-1}b_i) - \rho'(b_i b_{i+1})}{\Delta b_i}$$

where

$\rho'(b_{i-1}b_i)$ = first derivative of reflectance curve between b_{i-1} and b_i

m..n = start and end band number of DGVI area

b_i = centre wavelength of band i

i = band number

Δb_i = step width: $b_{i+1} - b_{i-1}$

In detail this meant that for the calculation of the DGVI value for one band, the reflectance of three bands was needed:

$$DGVI(b_i) = \frac{\rho'(b_i) - \rho'(b_{i+1})}{\Delta b} = \frac{\dfrac{\rho_i - \rho_{i-1}}{b_i - b_{i-1}} - \dfrac{\rho_{i+1} - \rho_i}{b_{i+1} - b_i}}{\Delta b}$$

The above implies that for the calculation of the DGVI over a region of n bands, n+2 bands are needed.

3.4.5.2 NTBI

Normalized two band indices were calculated by:

$$NTBI = \frac{\rho(b_x) - \rho(b_y)}{\rho(b_x) + \rho(b_y)}$$

3.4.5.3 PCT

Principal components transformation requires as input the eigenvectors of a given dataset. The eigenvalues are given by the solution of the characteristic equation:

$$\chi^{(\Sigma_x)} = \det(\Sigma_x - \lambda \cdot I)$$

where

$\chi^{(\Sigma_x)}$ = characteristic polynomial of Σ_x

Σ_x = covariance matrix of the dataset

I = identity matrix

The eigenvectors are given by the solutions to the equation

$$\left|\sum_x - \lambda_i \cdot I\right| \underline{x}_i = 0$$

where

$\sum_x$ = covariance matrix of the dataset

I = identity matrix

λ_i = eigenvalue i

$\underline{x}_i$ = eigenvector i

The size of the eigenvalues is an indication of the correlation of the data. A rapid fall off in the size of the eigenvalues indicates a high correlation. The eigenvalues can be plotted as a scree plot which shows the drop off graphically. The proportion of variability explained by each component is given by:

$$proportion_i = \frac{eigenvalue_i}{sum(eigenvalues)}$$

The cumulative proportion is given by:

$$cum\_proportion_i = cum\_proportion_{i-1} + proportion_i$$

If the eigenvalues are contained in a matrix their sum is given by the trace of the eigenvalue matrix. The proportion and cumulative proportion indicate how many components the PCT should utilize.

The eigenvectors are then found by solving the characteristic equation for every eigenvalue.

The transformation matrix G is the transposed eigenvector matrix. The original data x is transformed into a new feature space by:

$$y = G \cdot x$$

If only the first n components are to be used, the transformation matrix is a sub-matrix of G, consisting of the first n components. The dimension of the resulting feature space is therefore n:

$$\underset{n\times 1}{y} = \underset{n\times m}{G} \cdot \underset{m\times 1}{x}$$

where

m = original size of data space

n = new size of data space, equal to the number of selected components

It must be noted that the eigen-decomposition of a dataset should be recomputed if new data is added to a dataset.

3.5 Statistical Analysis

3.5.1 Classification

Classification is the process of assigning an unknown object to a given class. Classifiers are the algorithms that are applied to the data during classification. Classifiers can be defined by discriminant functions. A simple example of a discriminant function is the distance to mean. The discriminant function of every class is applied to an unknown vector. The classifier then selects the class whose discriminant function produced the least distance between the unknown and the mean of this class. The resulting classifier is called 'Minimum Distance to Mean'.

Three different classifiers were implemented. Their discriminant functions are as follows:

Quadratic (Gaussian) distance (Richards, 1993):

$$g_i(x) = \ln\left|\sum{}_i\right| + (x - m_i)^t \sum{}_i^{-1} (x - m_i)$$

General squared distance (Minitab Inc., 2003):

$$g_i(x) = -2\left(m_i^t \sum{}_i^{-1} x - 0.5 \cdot m_i^t \sum{}_i^{-1} m_i\right) + x^t \sum{}_i^{-1} x$$

Spectral Angle Mapper (SAM) (Landgrebe, 2003):

$$g_i(x) = \cos^{-1}\left(\frac{x^t m_i}{\sqrt{x^t x}\sqrt{m_i^t m_i}}\right)$$

where

x = unknown vector

i = species i

$\sum_i$ = covariance matrix of species i

m_i = mean vector of species i

3.5.2 Discriminant Analysis

Discriminant analysis is similar to classification but the input consists of data of known classes. Therefore, discriminant analysis can be used to test the classifier as well as the discriminating power of the feature space. DA outputs not only a classification accuracy but also errors of omission and commission.

The output of the DA was an error matrix built as follows: the columns were the spectra to be classified ordered by species and the rows were the species that made up the library, i.e. the known classes. All spectra of each species were classified using one of the discriminant functions, the results were then written into the corresponding column and row, i.e. the number of correctly classified spectra ended up in the diagonal elements while the omission errors were stored in the off diagonal elements. The column and row total was the sum of all column or row elements respectively. The total number of classified spectra was the sum of all column and row totals. The overall accuracy was calculated by dividing the sum of all diagonal elements by the total number of spectra (Lillesand et al., 2004):

$$overall\_accuracy = \frac{Trace(error\_matrix)}{Total\ number\ of\ spectra}$$

The producer and user accuracies were then given by dividing the diagonal elements with the total of the respective column or row.

3.5.3 Separability Analysis

Separability measures were calculated for all species pair combinations that had enough spectra to form a well defined distribution in an n-dimensional feature space.

The Jeffries-Matusita (JM) and the Bhattacharya (B) distances were chosen for this task (Richards, 1993):

$$J_{ij} = 2\left(1 - e^{-B}\right)$$

in which

$$B = \frac{1}{8}\left(m_i - m_j\right)^T \left(\frac{\sum_i + \sum_j}{2}\right)^{-1}\left(m_i - m_j\right) + \frac{1}{2}\ln\left(\frac{\left|\frac{1}{2}\left(\sum_i + \sum_j\right)\right|}{\left|\sum_i\right|^{1/2}\left|\sum_j\right|^{1/2}}\right)$$

where

i, j = species i, resp. j

$\sum_i$ = covariance matrix of species i

m_i = mean vector of species i

3.5.4 Most Discriminating Bands

The discrimination potential of the bands was tested using the Mann-Whitney test, also known as the two sample Wilcoxon test. The Wilcoxon test was applied to all possible species pairs. Only library relevant i.e. species with at least 15 spectra were included in this test. The number of possible species pairings is given by the binominal coefficient:

$$\binom{N}{2} = \frac{N!}{(N-2)!\cdot 2!}$$

Thus the 32 species formed 496 combinations. For all possible species pairs the Wilcoxon test was carried out for every band:

$$p_{XY\_i} = wilcoxon\left(\vec{r_{X\_i}}, \vec{r_{Y\_i}}\right)$$

where

r_{X_i} = vector containing the reflectances of all spectra of species X at the band i

r_{Y_i} = vector containing the reflectances of all spectra of species Y at the band i

p_{XY_i} = probability of the null hypothesis assuming that the samples supplied in the vectors were drawn from the same population. The smaller the value of p_{XY_i} the stronger the evidence that band i will discriminate.

A significance level of 0.01 was used to decide if the tested species were significantly different for the given band. The number of species pairs that were significantly different was counted for each waveband. This process was implemented in R (Venables et al., 2005) and applied to (a) raw data, (b) Hyperion synthesized data and (c) first derivative of Hyperion synthesized data.

3.6 Mixed Spectral Signatures

Spectral unmixing is usually applied to imagery. It is however conceivable that spectroradiometer field data can also be unmixed. A few experiments were conducted as described hereafter to produce spectral mixtures under a controlled environment.

The general setup for all these experiments is shown in Figure 26. A circular area was illuminated by a tungsten lamp set at an angle of 45° and sampled by a 25° bare fibre fore optic.

The diameter of the sampling area was chosen as 140mm, i.e. radius r = 70mm. The height h of the optic above the sampling area was calculated by:

$$h = \frac{r}{\tan(12.5°)} = 315mm$$

The distance of the Spectralon panel for the taking of white references was similarly calculated and set to 200mm.

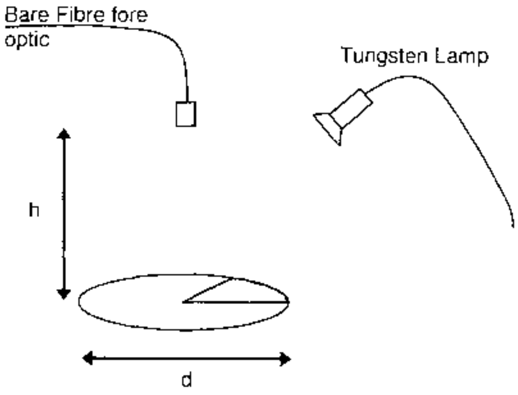


Figure 26: General mixing setup

3.6.1.1 Paper/Plant Mixture

A set of mixtures of white printing paper and kawakawa leaves (*Macropiper excelsum*) was sampled. The mixtures were defined by the angle of coverage, leading to the abundances shown in Table 5. The step size between mixtures was 30 as shown in Figure 27.

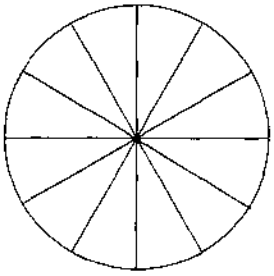


Figure 27: Mixture segments

Table 5: Mixtures of paper and kawakawa

Paper angle	Kawakawa angle	Paper abundance	Kawakawa abundance
360	0	1.00	0.00
330	30	0.92	0.08
300	60	0.83	0.17
270	90	0.75	0.25
240	120	0.67	0.33
210	150	0.58	0.42
180	180	0.50	0.50
150	210	0.42	0.58
120	240	0.33	0.67
90	270	0.25	0.75
60	300	0.17	0.83
30	330	0.08	0.92
0	360	0.00	1.00

3.6.1.2 Paper/Plastic/Plant Mixture

Similar to the above mixture, several combinations of mixtures of three endmembers were produced. The materials involved were: white printing paper, green plastic from fast binding folders and kawakawa (*Macropiper excelsum*) leaves. As 30° steps would have led to too many possible combinations with 3 materials, the step size was increased to 90° which led to the mixtures listed in Table 6.

Table 6: Mixtures of paper, plastic and kawakawa

Paper angle	Plastic angle	Kawakawa angle	Paper abundance	Plastic abundance	Kawakawa abundance
0	0	360	0.00	0.00	1.00
0	90	270	0.00	0.25	0.75
0	180	180	0.00	0.50	0.50
0	270	90	0.00	0.75	0.25
0	360	0	0.00	1.00	0.00
90	270	0	0.25	0.75	0.00
180	180	0	0.50	0.50	0.00
270	90	0	0.75	0.25	0.00
360	0	0	1.00	0.00	0.00
90	90	180	0.25	0.25	0.50
90	180	90	0.25	0.50	0.25
180	90	90	0.50	0.25	0.25
270	0	90	0.75	0.00	0.25
180	0	180	0.50	0.00	0.50
90	0	270	0.25	0.00	0.75

3.6.1.3 Three plant mixture

Similar to the paper/plastic/plant mixture experiment, mixtures of three plants were setup: kawakawa (*Macropiper excelsum*), lemonwood (*Pittosporum eugenioides*) and karaka (*Corynocarpus laevigatus*) (see Table 7). The experiment was conducted outdoors with the sun as light source.

Table 7: Mixtures of kawakawa, lemonwood and karaka

Kawakawa angle	Lemonwood angle	Karaka angle	Kawakawa abundance	Lemonwood abundance	Karaka abundance
0	0	360	0.00	0.00	1.00
0	90	270	0.00	0.25	0.75
0	180	180	0.00	0.50	0.50
0	270	90	0.00	0.75	0.25
0	360	0	0.00	1.00	0.00
90	270	0	0.25	0.75	0.00
180	180	0	0.50	0.50	0.00
270	90	0	0.75	0.25	0.00
360	0	0	1.00	0.00	0.00
90	90	180	0.25	0.25	0.50
90	180	90	0.25	0.50	0.25
180	90	90	0.50	0.25	0.25
270	0	90	0.75	0.00	0.25
180	0	180	0.50	0.00	0.50
90	0	270	0.25	0.00	0.75

3.6.1.4 Positional Dependence of Paper/Plastic Mixtures

In order to establish if the position of a segment on the mixing circle had any influence on the resulting signature, a positional dependence experiment was conducted. To cancel out effects that might be due to the illuminating tungsten lamp, the experiment was carried out once using the lamp in the laboratory and once using sunlight outdoors. Three mixtures were used with each mixture being setup in four different positions as shown in Table 8. The illumination was from the right hand side.

Table 8: Paper/plastic mixtures and positions

Paper angle	Plastic angle	Paper abundance	Plastic abundance	Position 1	Position 2	Position 3	Position 4
270	90	0.75	0.25				
180	180	0.50	0.50				
90	270	0.25	0.75				

3.6.1.5 Unmixing

The unmixing was implemented in Matlab (The MathWorks Inc., 2004) using a linear mixing model with full additivity constraint (Keshava and Mustard, 2002):

$$\hat{a}_U = \left(S^T S\right)^{-1} S^T x$$

$$\hat{a}_F = \hat{a}_U - \left(S^T S\right)^{-1} Z^T \left(Z\left(S^T S\right)^{-1} Z^T\right)^{-1} \left(Z\hat{a}_U - b\right)$$

where:

x = spectrum vector to be unmixed (L x 1)

S = endmember matrix (L x M) consisting of M endmembers with the columns being the endmember spectra vectors

$\hat{a}_U$ = the unconstrained least squares solution for the abundances of the endmembers in the spectrum vector x

Z = a 1 x M row vector having all ones

b = set to 1 for the enforcement of the full additivity constraint $Za = b$

$\hat{a}_F$ = full additivity solution of the abundance of the given endmembers in x

The negativity constraint was not used for the unmixing procedure due to the complexity of the involved implementation.

3.6.1.6 Probe Rotation

This experiment was designed to establish if the bare fibre optic was sampling the field of view homogenously. A fifty-fifty mixture of white printing paper and green plastic was sampled outdoors to remove any influence of the tungsten lamp at four 90° rotational positions of the probe as shown in Figure 28.

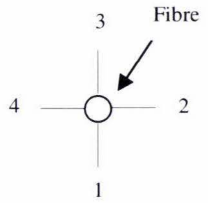


Figure 28: Rotational positions of the bare fibre

4 Results

4.1 Spectral Properties of New Zealand Native Plants

The spectra of the collected species show the typical features of vegetation (see Figure 29): a low reflection in the visible with a noticeable peak in the green around 570nm for most species. Exceptions are *Chionochloa rubra* (Red Tussock) and *Dracophylum subulatum* (Monoao) that are both of a brownish colour and therefore show a slope rising from blue to red. The red edge is found around 690nm where a steep rise begins that starts to level out at the NIR shoulder around 780nm. The first NIR absorbtion feature lies around 990nm, followed by the 1st NIR peak (~1090nm), the 2nd NIR absorbtion feature (~1220nm) and the 2nd NIR peak (~1290nm). The shortwave infrared (SWIR) shows two peaks at ~1690nm and ~2220nm respectively.

Figure 30 shows the mean spectra per species of all collected species. The waterband noise was removed before carrying out a Hyperion synthesizing.

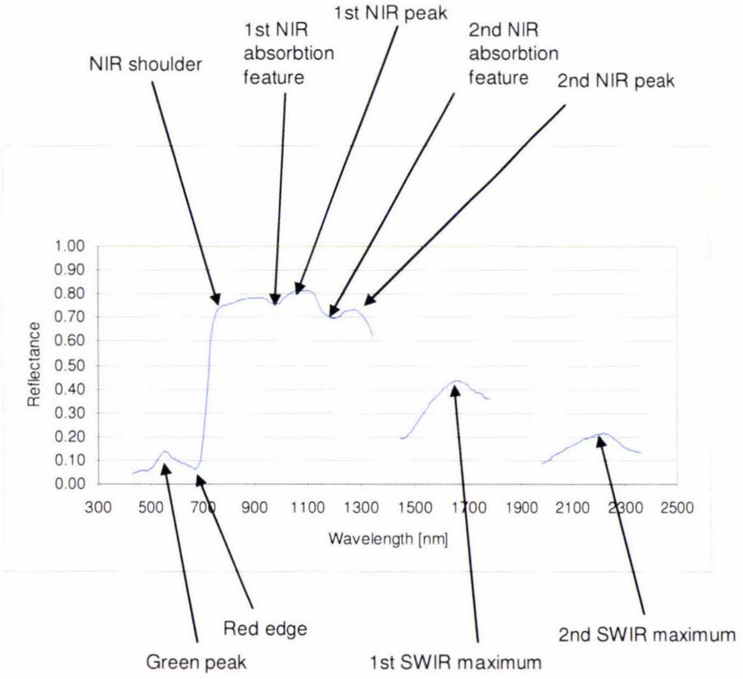


Figure 29: Features of a vegetation curve

A visual discrimination of the species by their reflectances alone must be regarded as difficult for most species. As the feature space concept was chosen for this study, pre-processing steps including a feature space transformation had to be applied before a multivariate discrimination could be carried out. The best settings for the pre-processing had to be established first. The results of these steps are described in the following sections.

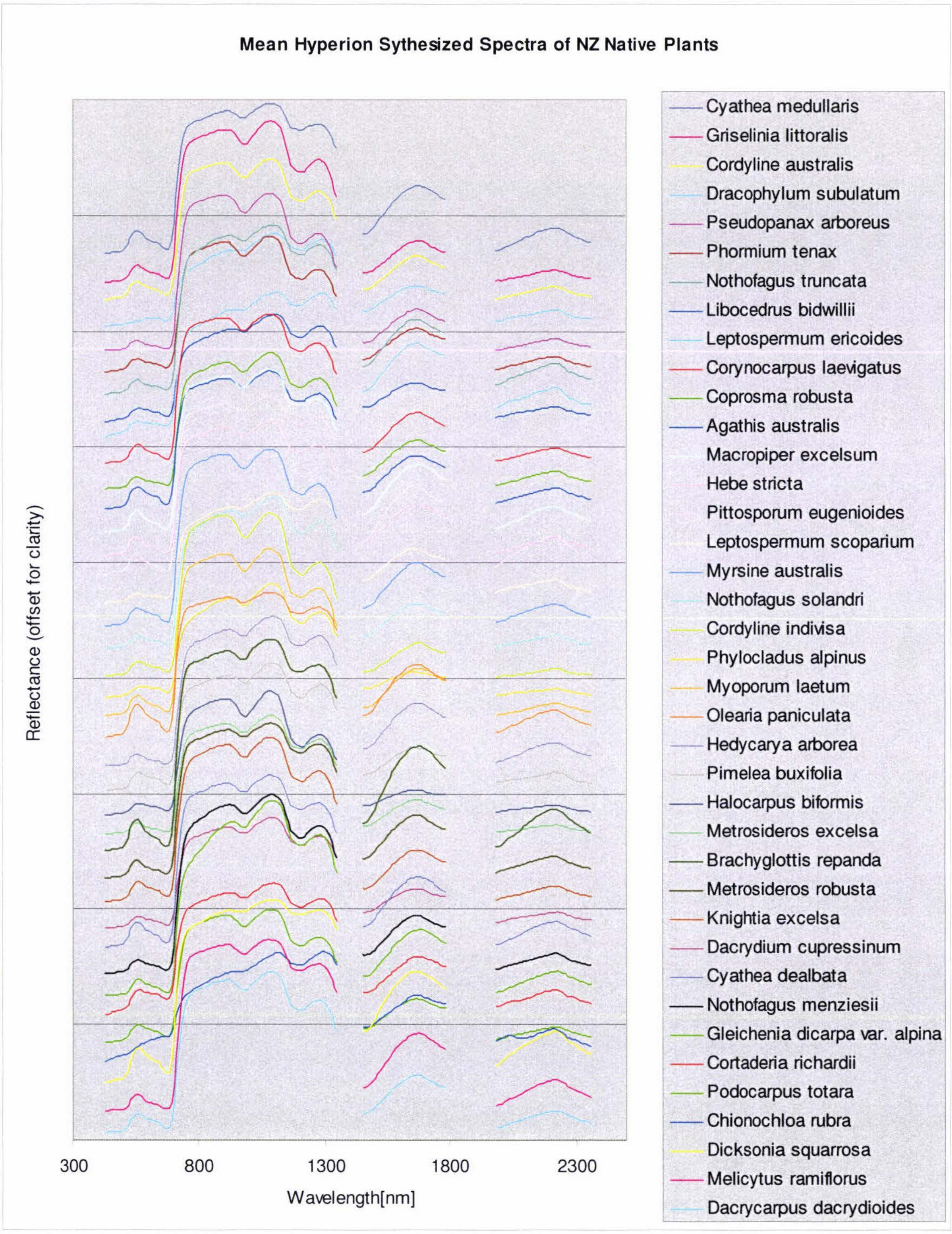


Figure 30: Mean Hyperion synthesized spectra of NZ native plants

4.1.1 Smoothing

A Savitzky-Golay filter was applied to the data. The resulting smoothed data is a function of the filter size and the polynomial order. The determination of the best parameters that, in the best case, remove all noise but retain all information is however not straightforward as mentioned by Schmidt and Skidmore (2004). The main difficulty is that a non-noisy reference spectrum, against which the efficiency of the smoothing filter could be measured, does not exist. The remaining options to assess the smoothing result were visual inspections; either of raw and smoothed data together or of the noise spectra calculated by subtracting the smoothed from the raw spectra.

To show the trends of the smoothing operation, filter sizes in steps of 10 between 11 and 51 were combined with polynomial orders 3, 4 and 6. The reason for leaving out order 5 is due to the fact that the smoothing coefficients for orders 4 and 5 are identical, as are those for orders 2 and 3. All tests were carried out on a spectrum of *Pittosporum eugenioides*. The resulting spectra are shown in Figure 31.

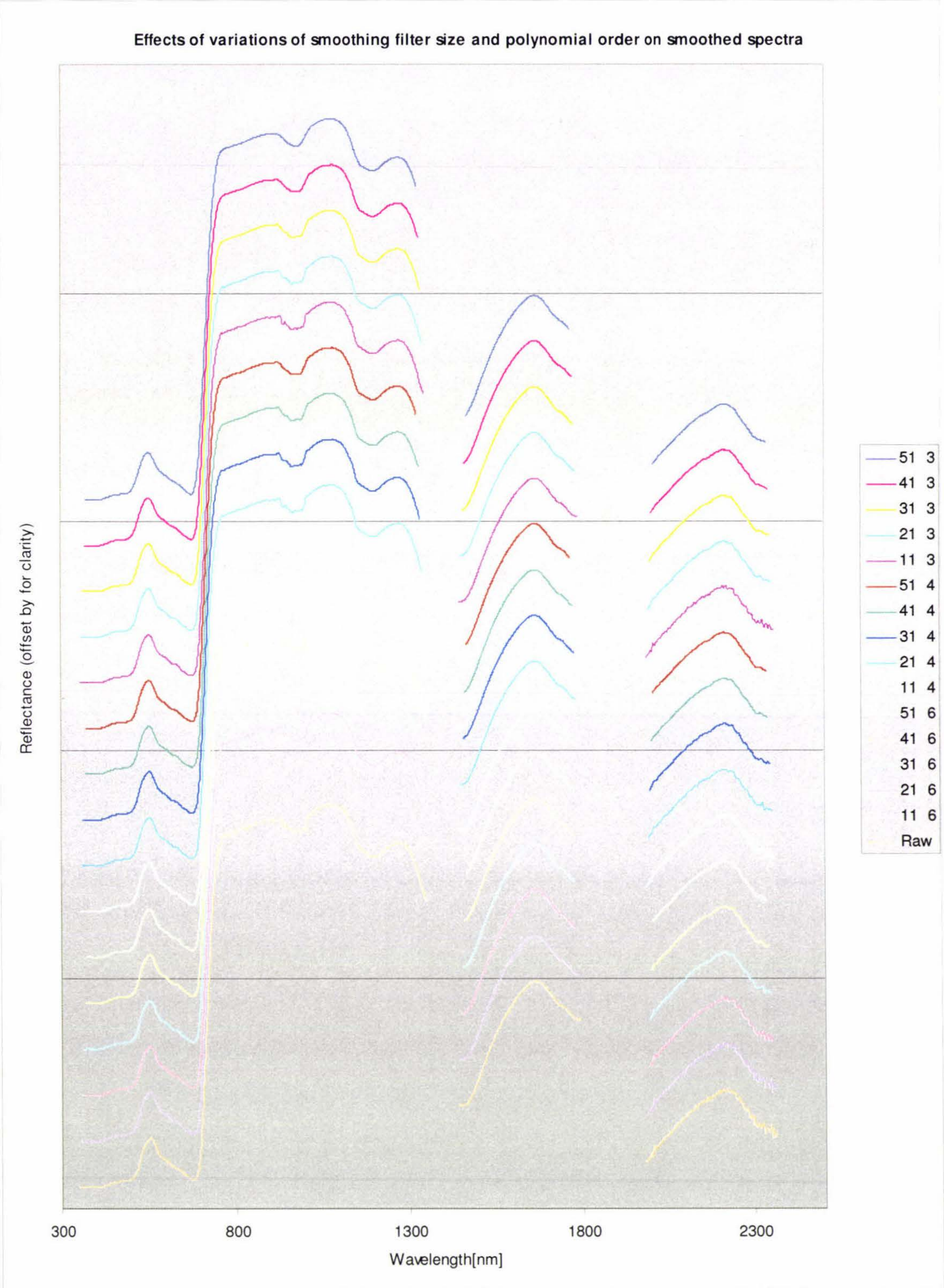


Figure 31: Effects of variations of smoothing filter size and polynomial order on smoothed spectra

Regardless of the polynomial order the bigger filter sizes remove more noise, best seen in the region around 2300nm. The effect of the order of the fitted polynomial is most noticeable in the NIR at the start of the first absorbtion feature (~950nm). Orders 4 and 6 still preserve some subtle changes in this region, even with a 51 filter size while order 3 (filter size 51) almost totally removes these undulations and produces a smooth curve.

The noise spectra (Figure 33) show that invariably the region between 2000nm and 2300nm is the noisiest followed by the red-NIR region (700nm-1180nm). As expected, the biggest filter combined with the smallest order performs the most smoothing. However, the noise spectra just show the removed data regardless of whether they are noise or valuable information. One would expect noise to be randomly distributed. A close inspection of the noise spectra in the red-NIR region (see Figure 34) gives an indication that valid data are removed by the filters of order 3. Regions that seem to consist of valid features and not random noise are: 700nm-770nm, 900nm-1020nm and 1130nm-1170nm. One can also observe the jump at 1000nm that occurs where the visible and SWIR1 spectrometer data are spliced.

These findings suggest that polynomial orders of 3 or lower are likely to remove valid information while order 6 filters might retain too much noise by over-accurate curve fitting. Filter sizes around 31 combined with polynomial order 4 seem to be a good trade off between retention of spectral features and smoothing being attained.

An analysis of the RMSE of raw minus smoothed spectra (Figure 32) shows that with order 3 the filtered noise grows with increasing filter size while for order 5 the RMSE for filter size 51 is actually lower than for size 41. The phenomenon of decreasing RMSE with increasing filter size for higher orders is because of (a) a generally lower overall noise due to more accurate curve fitting and (b) a complete loss of information at every segment due to increasing filter sizes.

Ultimately, the best smoothing filter will be the one that produces the best results in the analysis stage.

Two versions of smoothing using Savitzky-Golay coefficients were implemented: moving window and convolution by fast Fourier transformed (FFT) data. It was found that the moving window calculation performed faster than FFT.

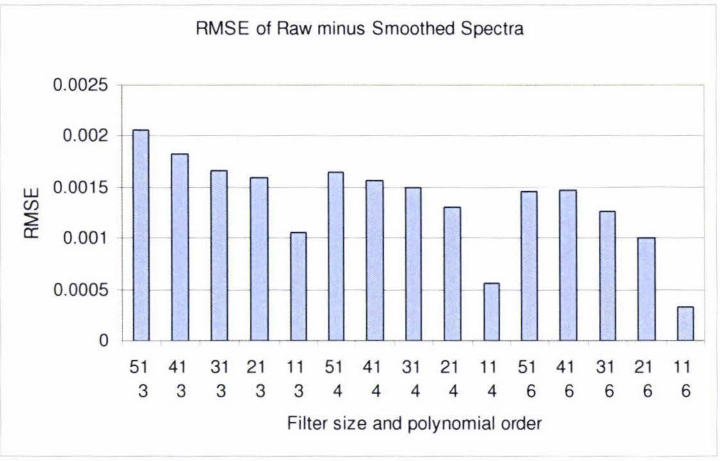


Figure 32: RMSE of raw minus smoothed spectra

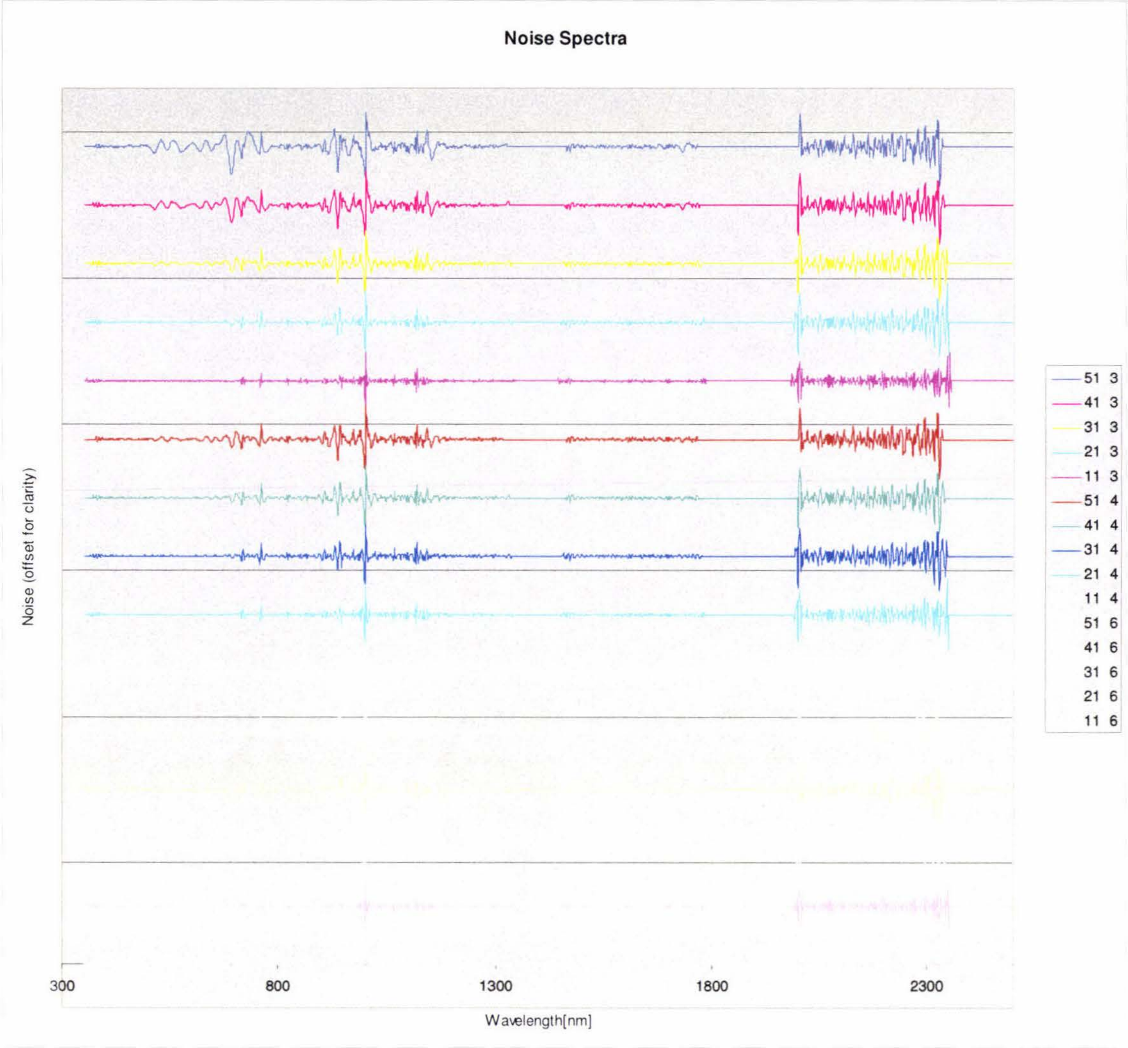


Figure 33: Noise spectra of different Savitzky-Golay filter settings (raw minus filtered spectra)

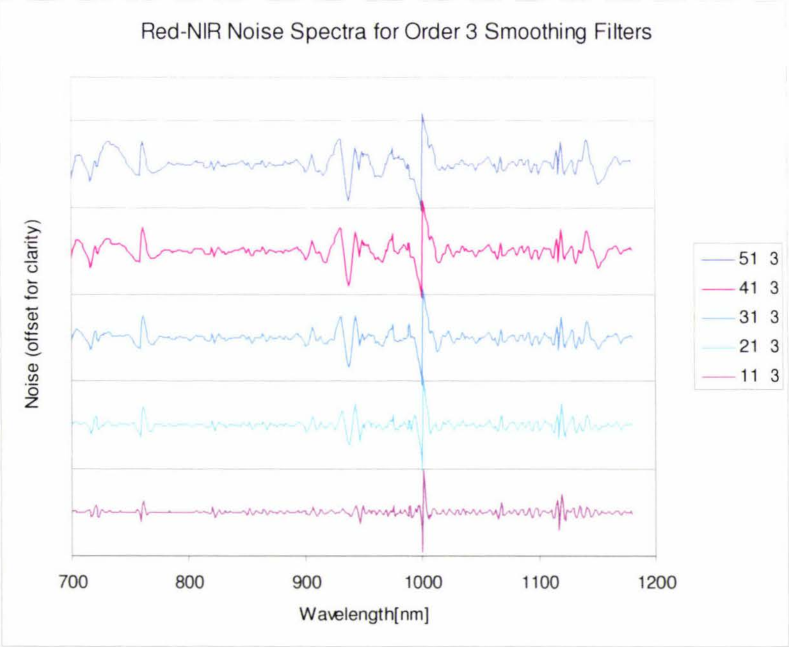


Figure 34: Red-NIR region of the noise spectra after filtering with order 3 smoothing filters (raw minus filtered spectra)

4.1.2 Sensor Synthesizing

Sensor responses for Hyperion, downsampling and Landsat7 ETM+ were calculated. The noise spectra shown were calculated by first interpolating the sensor bands to ASD bands and then subtracting the interpolated ASD data from the raw ASD data. The interpolation function was chosen as a linear curve, based on the fact that straight lines were also used for the explicit calculation of derivatives. These straight segments between the sensor reflectance values were defined by:

$$f(\lambda) = \alpha + \lambda \cdot \beta$$

$$\beta = \frac{\rho(b_j) - \rho(b_i)}{b_j - b_i}$$

$$\alpha = f(b_i) - b_i \cdot \beta$$

where

$f(\lambda)$ = interpolation function for a curve segment

b_i, b_j = wavelengths of consecutive sensor bands

The resulting RMSE could not be directly compared with the RMSE obtained from the different smoothing parameters. The data reduction of the synthesizing and subsequent interpolation by straight segments naturally results in a higher loss of data.

4.1.2.1 Hyperion

The Hyperion synthesizing resulted in 166 new bands. The smoothing function of the synthesizing process proved to be good enough to apply the synthesizing to the raw data without any previous smoothing step. Figure 35 shows the full raw and Hyperion synthesized spectra of *Pittosporum eugenioides*. One can observe that the Hyperion synthesizing results in a good fit of the raw data at least visually. Figure 36 shows the NIR and SWIR2 parts where the most data were removed by the smoothing operation. The NIR part again shows the jump at 1000nm due to the internal spectrometer switchover.

The noise spectrum of raw minus interpolated Hyperion synthesized data with an RMSE of 0.002254 is shown in Figure 37. The negative and positive noise peaks around 700nm are the effect of the data loss at the red edge. The spectral curve rises from 0.07 to 0.81 in only about 80nm, i.e. the ASD sensor samples 80 data points while the Hyperion sensor models this curve segment with only 9 data points. The average vertical difference between data points is 0.00925 for ASD and 0.08 for Hyperion. The exact shape of the curve is thus lost in this region.

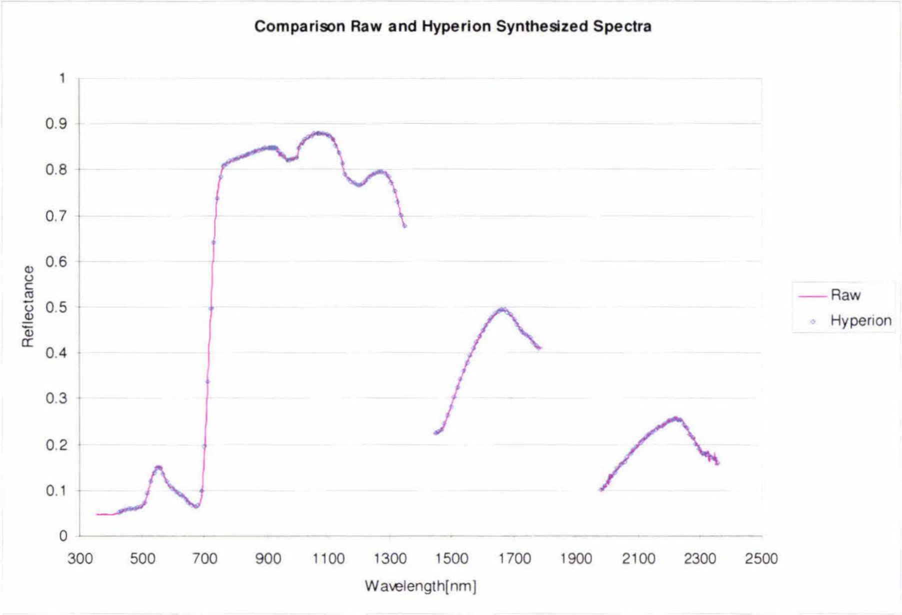


Figure 35: Raw and Hyperion synthesized spectra of Pittosporum eugenioides

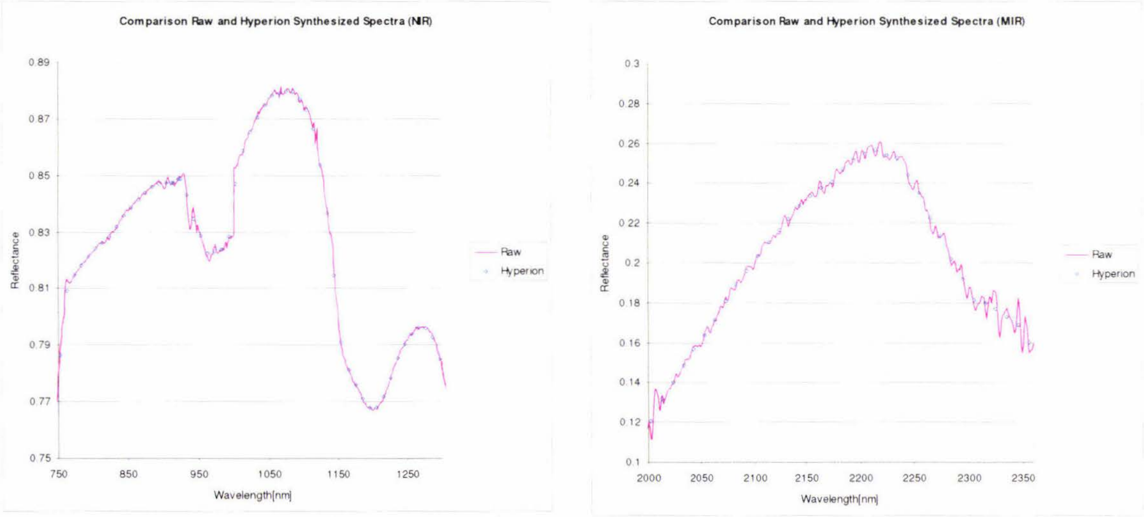


Figure 36: Raw and Hyperion synthesized in NIR and SWIR2 parts of the spectrum

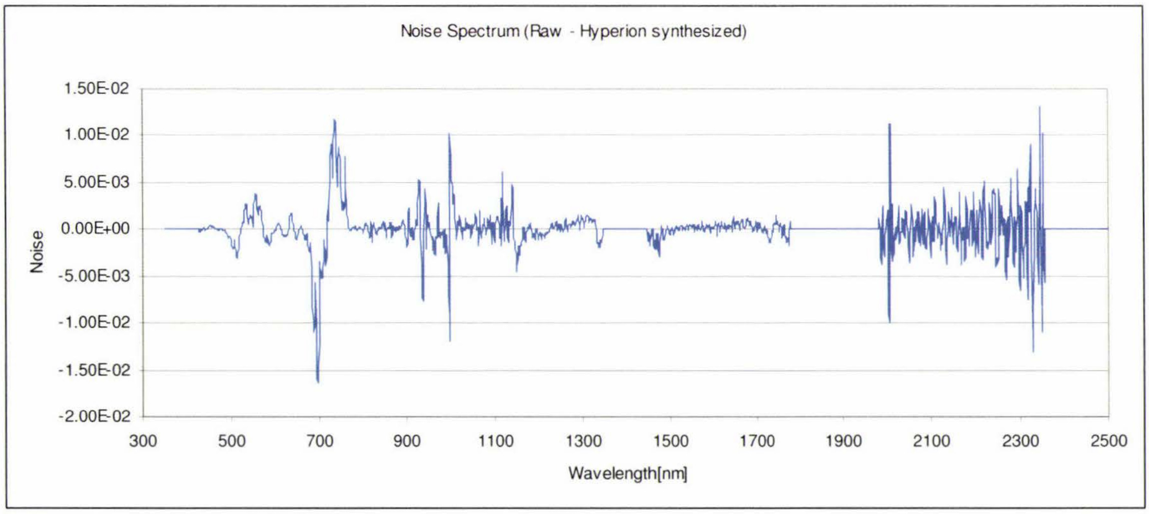


Figure 37: Noise spectrum of Pittosporum eugenioides (raw minus Hyperion synthesized)

4.1.2.2 Downsampling

All downsampling was preceded by a smoothing operation using a filter size of 31 and an order of 4. Thus the results presented here are the output of a decimation function.

Two different downsampling rates were implemented: factor 5 and factor 10.

A graphical comparison shows that decimation by 5 is superior in retaining details of the spectral curve (see Figures 38-40). This does not necessarily imply that analysis based on decimation 5 will yield better results as the retained spectral features might just as easily be noise.

The noise spectra (see Figures 41 and 42) show that decimation by 5 removes less noise than decimation by 10. This is most obvious in the red edge (700-770nm) where the closer sampling interval of the decimation by 5 models the curve shape more accurately. The RMSEs were 0.001291 for decimation by 5 and 0.00161 for decimation by 10.

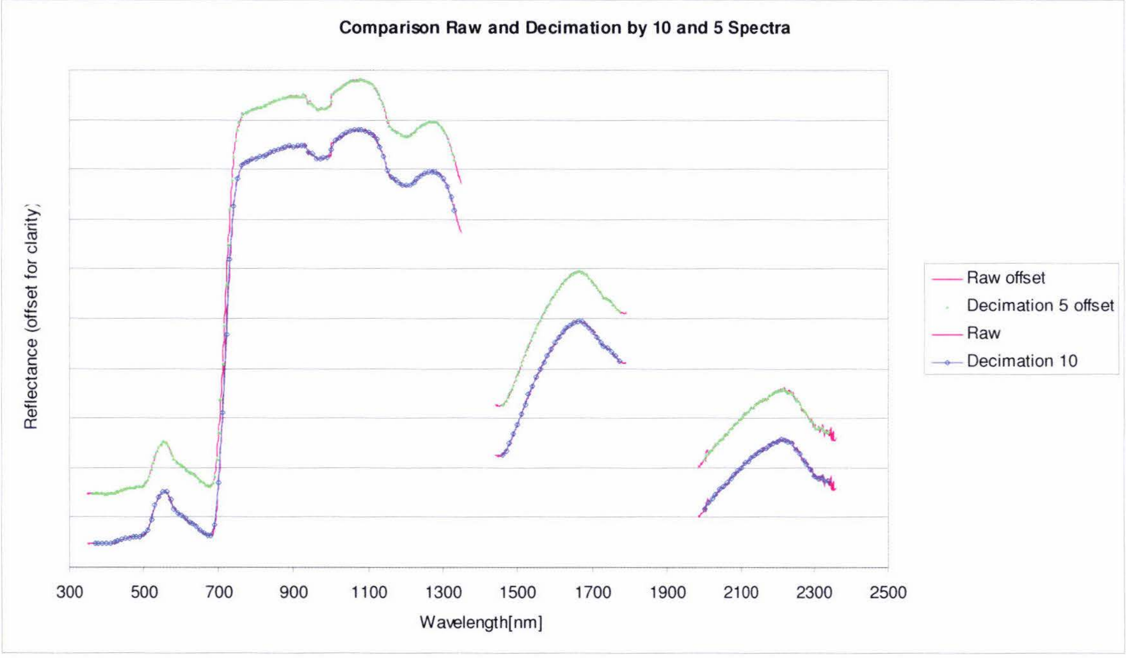


Figure 38: Raw and decimated by factor 10 and 5 spectra of Pittosporum eugenioides (offset for clarity)

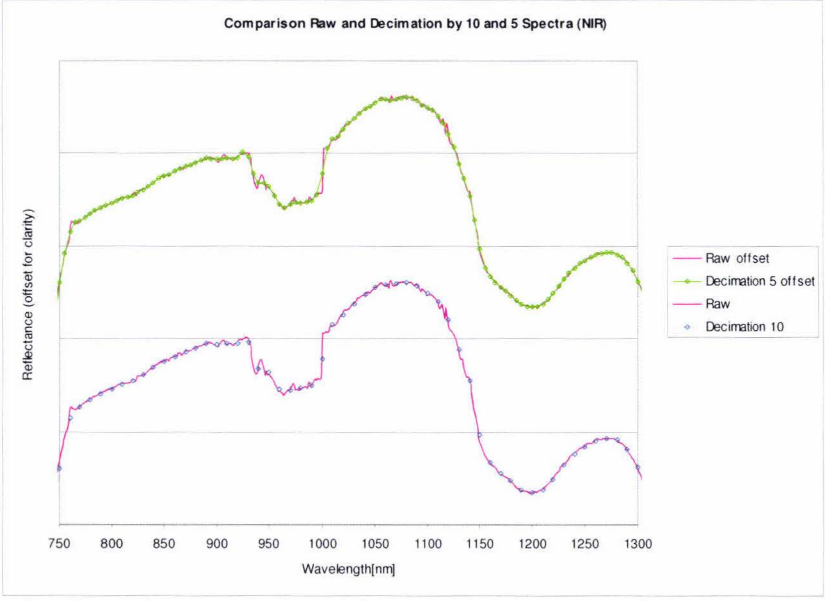


Figure 39: Raw and decimated by factor 10 and 5 (NIR part of the spectrum)

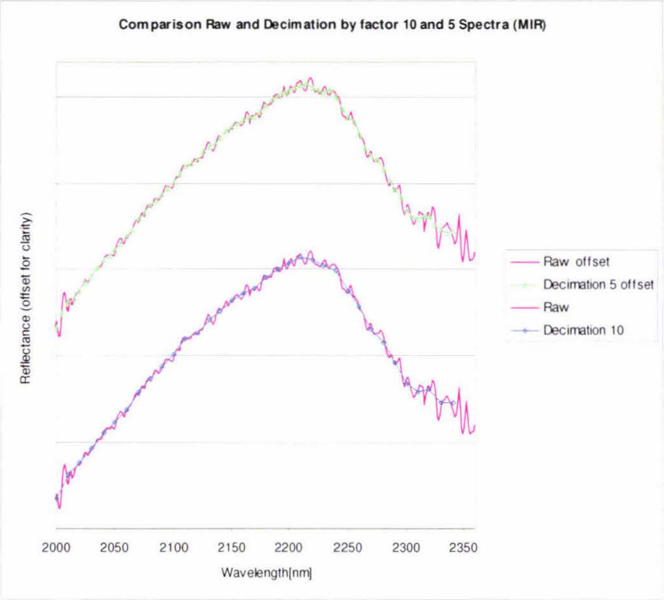


Figure 40: Raw and decimated by factor 10 and 5 (SWIR2 part of the spectrum)

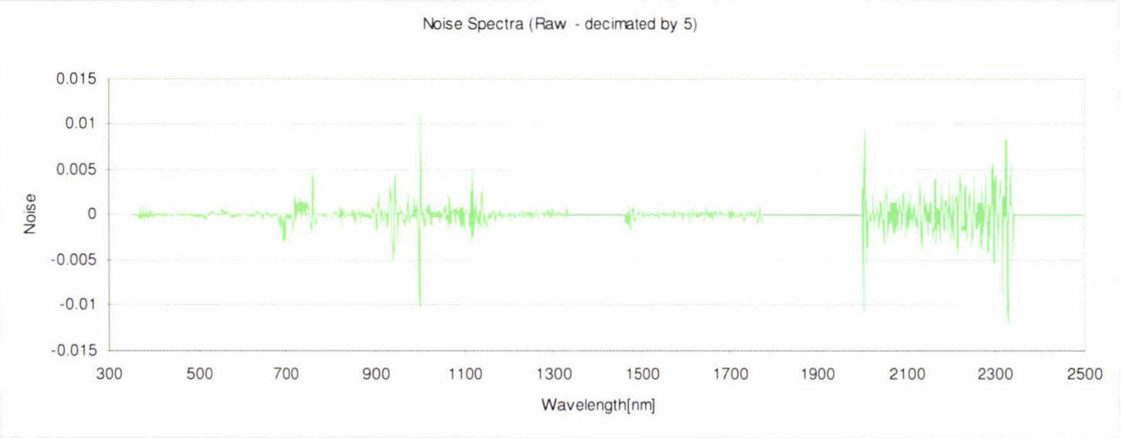


Figure 41: Noise spectrum of Pittosporum eugenioides (Raw minus decimated by factor 5)

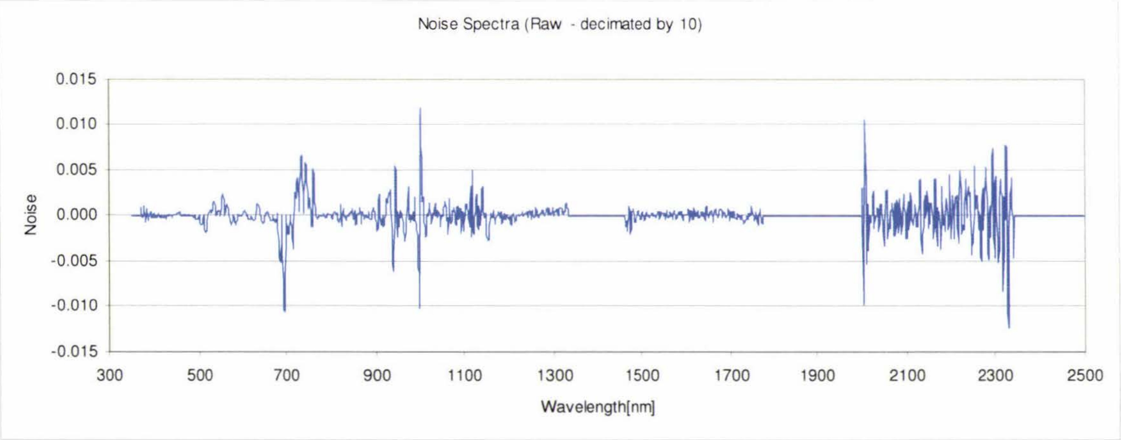


Figure 42: Noise spectrum of Pittosporum eugenioides (Raw minus decimated by factor 10)

4.1.2.3 Comparison of Hyperion Synthesizing and Decimation

A comparison of the RMSE of the noise spectra of raw minus Hyperion synthesized and decimation by 10 and 5 shows that Hyperion synthesizing removes the most noise (see Figure 43). Again, the optimal sensor synthesizing for a subsequent analysis task should be chosen based on the analysis results as it is not easy to distinguish between removed noise and valid spectral features.

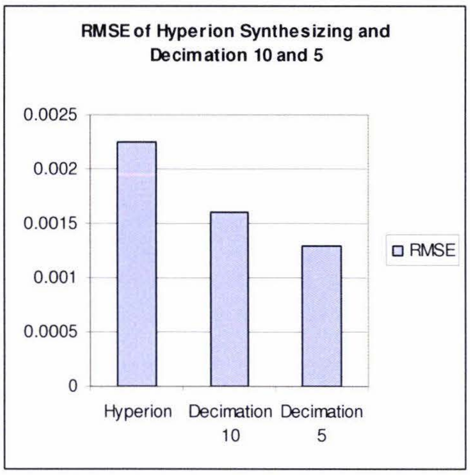


Figure 43: RMSE of Hyperion synthesizing and Decimation 10 and 5

4.1.2.4 Landsat 7 ETM+

Landsat7 ETM+ synthesizing resulted in a drastic data reduction, creating 6 new bands (Landsat bands 1-5 and 7). The wavelengths chosen for the Landsat bands were the middle wavelengths of the individual sensor elements. Figure 44 compares the synthesized signatures of *Pittosporum eugenioides* for Landsat7 ETM+ and Hyperion. While an identification of species using Landsat7 ETM+ data would undoubtedly be more difficult than using Hyperion, it does feature datapoints in the blue, green, red, NIR and SWIR, meaning that data for vegetation studies is available as has been demonstrated by many studies using Landsat data.

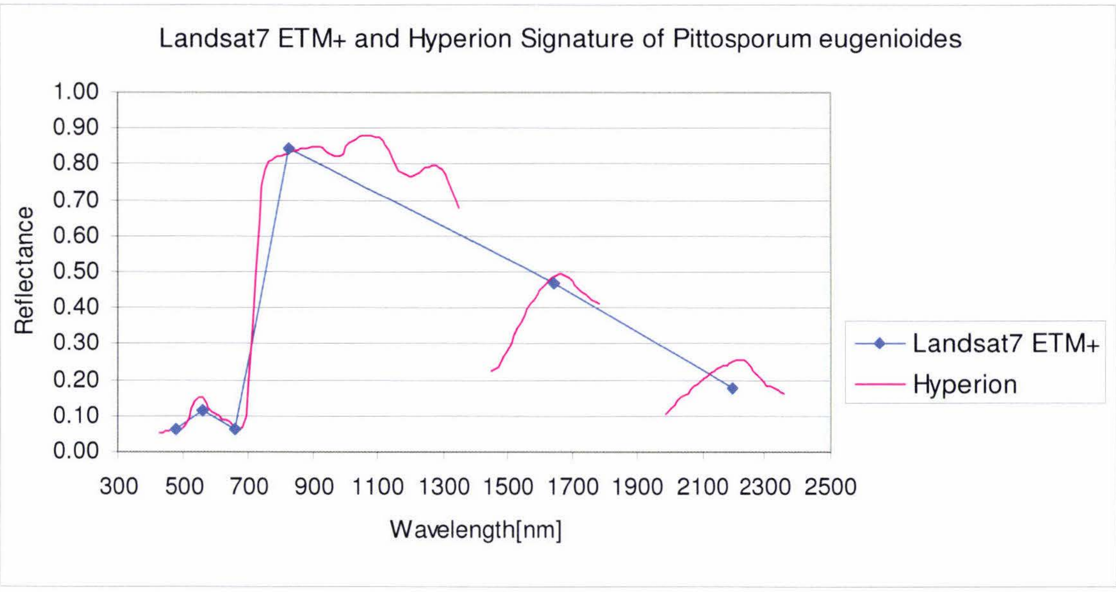


Figure 44: Landsat7 ETM+ and Hyperion signatures

4.1.3 Derivative Calculation

The derivative calculation was found to be very dependent on the pre-processing of the data; even a slight noise in the input data resulted in high noise in the derived data. To illustrate this, six different derivatives of *Pittosporum eugenioides* are shown in Figure 45. The noisiest derivative was calculated from the raw data. The sharp spike at 1000nm is exactly at the position of the sensor overlap, i.e. it is an artefact of the machine. These steps can appear due to a lack of warm up time of the ASD instrument. Information to support or reject this possibility was not available in this particular case. The derivative of the smoothed data (Savitzky-Golay smoothed with filter size 31 and order 4) appears much smoother than the derivative of the raw data, especially so in the region 800-1100nm that includes the sensor overlap and above 1800nm where the raw data show high noise. The noise was further minimized by smoothing the data using again a Savitzky-Filter followed by a derivative calculation using Savitzky-Golay coefficients for a first derivative with a filter size of 31 and polynomial order 4 (see curve named '1st derivative (SavGol) of smoothed data'). The data was thus essentially smoothed twice. The resulting derivative was smoother than the one that had been smoothed once only. The smoothest derivative was obtained from Hyperion synthesized data, followed by data decimated by factor 10 and factor 5. The decimations were calculated by first smoothing with a Savitzky-Golay filter of size 31 and order 4 and then downsampling by afore mentioned factors.

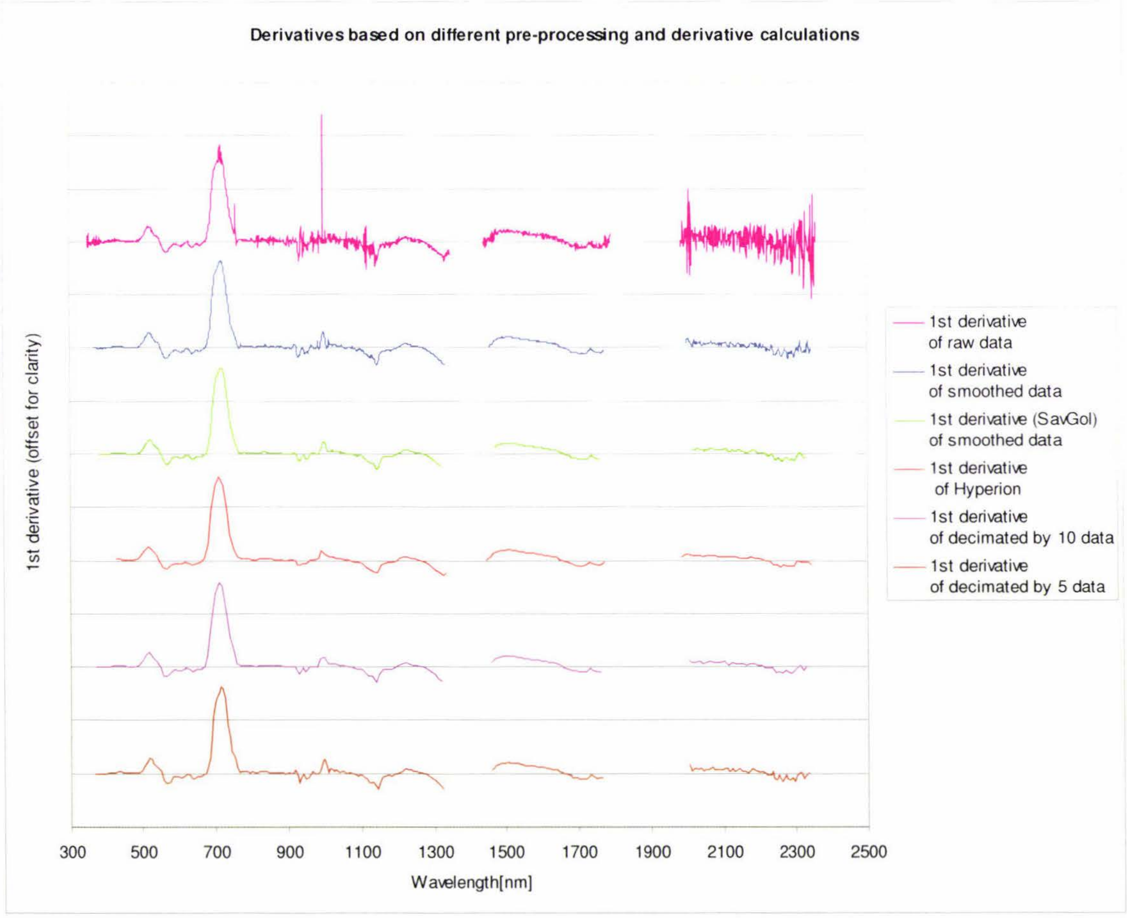


Figure 45: Derivatives based on different pre-processing and derivative calculations

4.1.4 Feature Space Transformation

4.1.4.1 DGVI

The DGVI regions were based on the wavelengths used by Thenkabail et al. (2004a; , 2004b) 515-535 nm (DGVI1), 540-560 nm (DGVI2), 560-580 nm (DGVI3), 650-670 nm (DGVI4), 700-740 nm (DGVI5), 626-795 nm (DGVI6), 1500-1650 nm (DGVI7), 2080-2350 nm (DGVI8) and 428-906 nm (DGVI9), 428-2355 nm (DGVI10). These regions were then slightly modified to render them useful for Hyperion synthesized data as some Hyperion band wavelengths were just outside the original regions. These modified regions were: DGVI1 (508-539nm), DGVI4 (650-672nm), DGVI5 (700-743nm). DGVI8 and DGVI10 were cut short to avoid the highest noise in the SWIR2 segment: DGVI8 (2080-2336nm), DGVI10 (428-2336nm).

For clarity these regions are shown in Figure 46. DGVIs 1-4 were narrow (~20nm), DGVIs 5, 7 and 8 were broader (40 – 270nm) and DGVIs 6, 9 and 10 were very broad and included other DGVI regions.

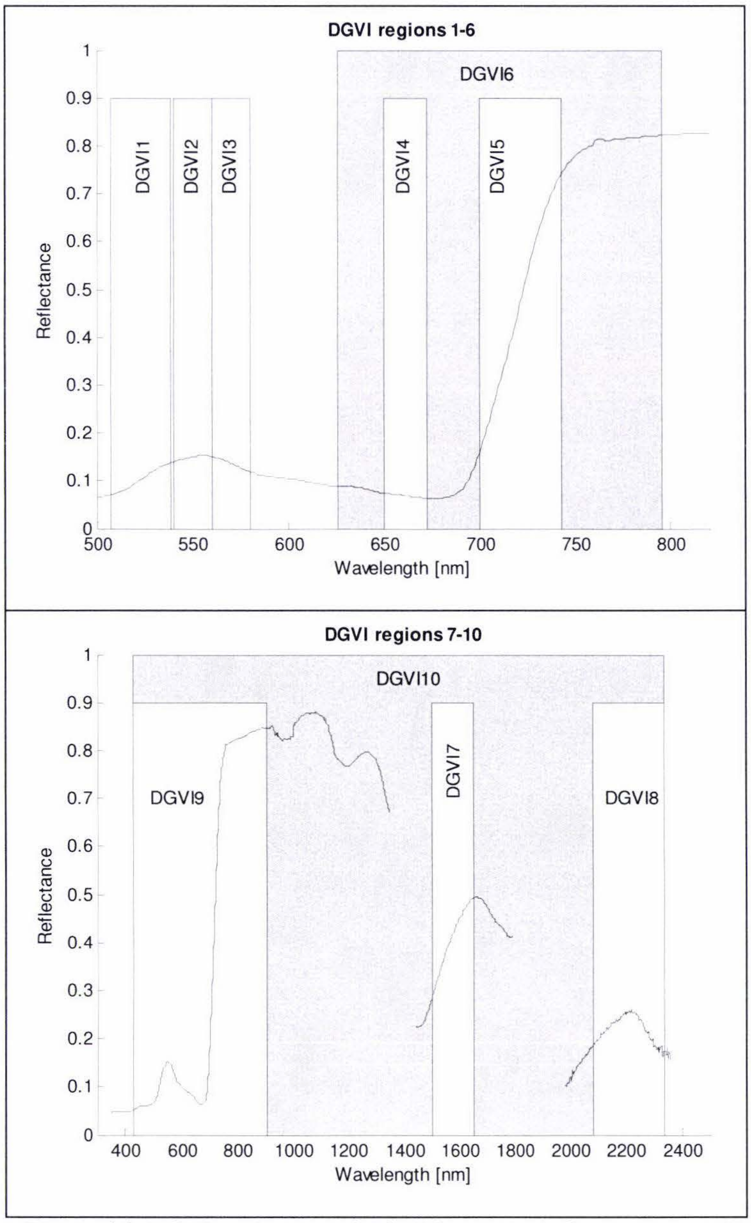


Figure 46: DGVI regions overlaid with a typical plant spectrum (Pittosporum eugenioides)

The DGVI transformation resulted in a new 10 dimensional space.

The discriminative power of the DGVIs was measured by means of the Wilcoxon test at a significance level of 0.01. The process was similar to the one described in section 3.5.4. The result is the count of species pairs with a statistically significant difference per DGVI (see Figure 47 and Table 9). This significance test was carried out on:

□ Unsmoothed data

□ Savitzky-Golay smoothed data with a polynomial order of 4 and filter size of 31

□ Savitzky-Golay smoothed data with a polynomial order of 4 and filter size of 51

□ Decimation by 10 (Savitzky-Golay smoothed (size 31, order 4) followed by downsampling by factor 10)

□ Decimation by 5 (Savitzky-Golay smoothed (size 31, order 4) followed by downsampling by factor 5)

□ Decimation by 5 (Savitzky-Golay smoothed (size 51, order 4) followed by downsampling by factor 5)

□ Hyperion synthesized

□ Hyperion synthesized preceded by a smoothing with a Savitzky-Golay filter (size 51, order 4)

Table 9: Mean frequencies of statistically significant differences in species pairs for DGVIs calculated for differing pre-processing parameters

	Raw data	Smoothed (31 4)	Smoothed (51 4)	Decimation by 10 (31 4)	Decimation by 5 (31 4)	Decimation by 5 (51 4)	Hyperion	Hyperion pre-smoothed (51 4)
DGVI_1	57	260	264	277	269	278	275	301
DGVI_2	65	252	248	256	249	254	268	272
DGVI_3	94	270	269	315	288	309	283	282
DGVI_4	62	274	280	277	278	274	281	289
DGVI_5	287	327	327	324	332	338	323	333
DGVI_6	50	250	243	247	242	247	258	276
DGVI_7	84	374	384	392	396	377	396	385
DGVI_8	15	47	103	99	47	150	101	282
DGVI_9	50	246	239	248	242	239	261	282
DGVI_10	12	92	251	215	138	256	206	300
Mean	77.6	239.2	260.8	265	248.1	272.2	265.2	300.2

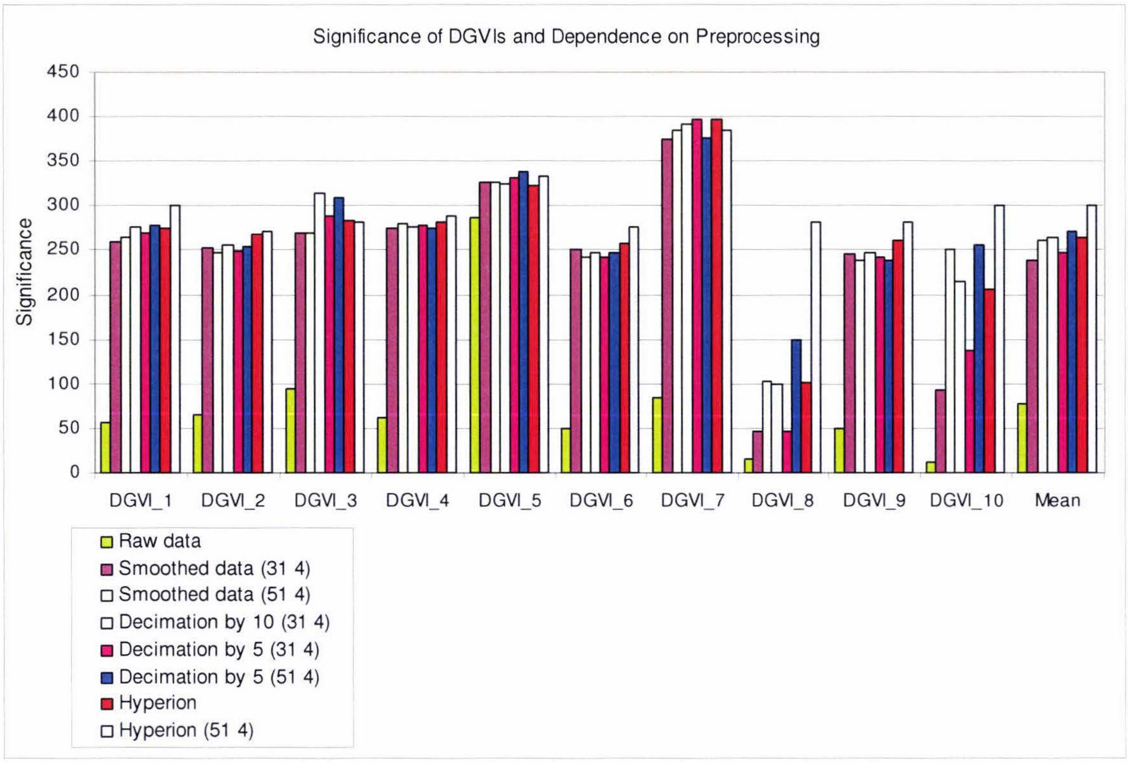


Figure 47: Frequency of statistically significant differences of DGVIs and their dependence on pre-processing

Unsmoothed raw data produced the lowest frequencies of all data types tested with a mean of 77.6. Only DGVI5 had a high frequency (287) which can be explained by less noise occurring in the red edge of the spectrum where the curve rises sharply enough to reduce the impact of noise. All other DGVIs had low frequencies, especially DGVI8 which covers the very noisy SWIR2 segment and DGVI10 which includes almost the full spectrum.

The impact of noise on the DGVIs was demonstrated by the fact that smoothed raw data produced much higher frequencies with a mean of 239.2 for filter size 31 and 260.8 for filter size 51. The bigger filter size produced smoother curves and resulted in better frequencies for the noisy DGVI segments 8 and 10.

Decimation by 10 preformed similarly to the smoothed raw data with a mean frequency of 265.

Decimation by 5 was again dependent on the filtering preceding the downsampling. A filter size of 31 produced a mean of 248.1 while a filter of size 51 resulted in a mean of 272.2. The most improvement by larger filter sizes was again found for DGVIs 8 and 10.

Hyperion synthesized data produced similar results as decimation by 10 and 5 respectively with a mean of 265.2.

The best overall result with a mean of 300.2 was achieved by Hyperion synthesized data preceded by a smoothing (filter size 51 order 4).

Regardless the pre-processing the highest frequencies occurred in the SWIR1 segment which is partly covered by DGVI7.

The 10 DGVIs define a feature space in which the species form distributions. This concept can be visualized in two dimensions by scatterplots of two DGVIs (see Figure 48). In this example, the combination of DGVI2 and DGVI6 showed a considerable overlap of the distribution of *Halocarpus* *biformis* with *Nothofagus menziesii* and *Pittosporum eugenioides*. The combination of DGVI2 and DGVI7 was more successful in separating the three species. The oval shape of the scatter for DGVI2 versus DGVI6 also indicated that these two dimensions were correlated.

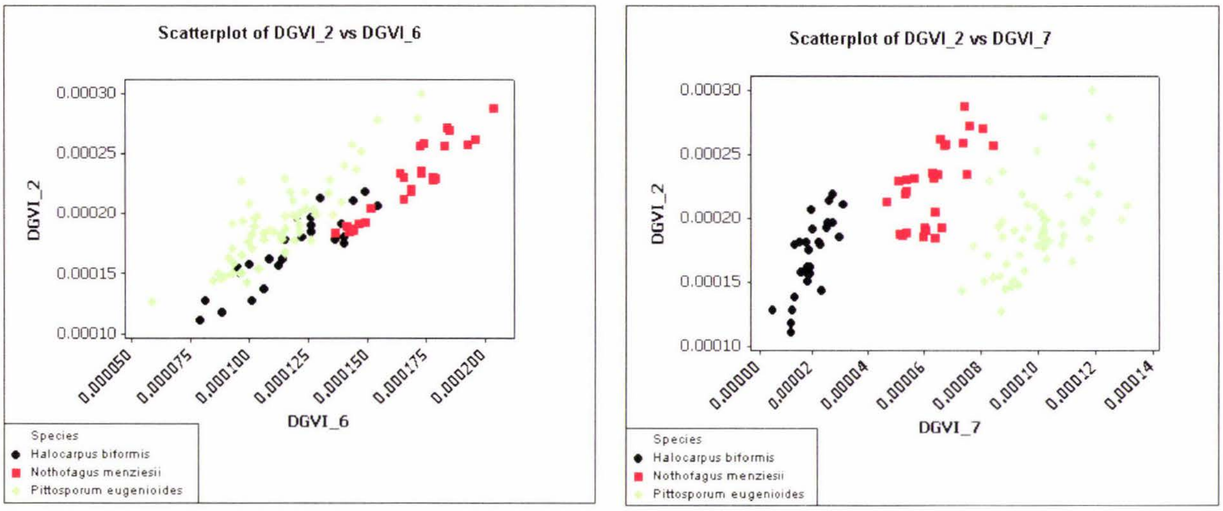


Figure 48: Example of the discrimination of species by DGVIs (calculation based on Hyperion synthesized data)

4.1.4.2 NTBIs

Based on the work on crops by Thenkabail et al. (2000) three narrow band NDVI type indices were used as examples for NTBIs: NTBI1 (550nm and 468nm), NTBI2 (550nm and 682nm) and NTBI3 (920nm and 696nm).

For direct comparison with the discriminating power of the DGVIs, the same pre-processing sets were used and a Wilcoxon test with a significance level of 0.01 was applied to all NTBIs (see Table 10 and Figure 49).

Not surprisingly, the pre-processing had little influence on the discriminating power of the three indices. The effect of smoothing operations was absolutely minimal. Generally, slightly broader bandwidths (~10nm) preformed a bit better than the very narrow (1nm) bandwidths.

Interestingly, with a mean frequency of 283.5 the NTBIs were more discriminating than the DGVIs which had an overall mean frequency of 241.

Table 10: NTBI and mean frequencies of statistically significant differences in species pairs

	Raw data	Smoothed (31 4)	Smoothed (51 4)	Decimation by 10 (31 4)	Decimation by 5 (31 4)	Decimation by 5 (51 4)	Hyperion	Hyperion (51 4)
NTB1	303	304	303.0	301	301	302.0	306	305
NTBI2	268	268	270.0	265	265	269.0	265	265
NTBI3	279	278	279.0	274	281	278.0	288	287
Mean	283.333	283.333	284	280	282.333	283	286.333	285.667

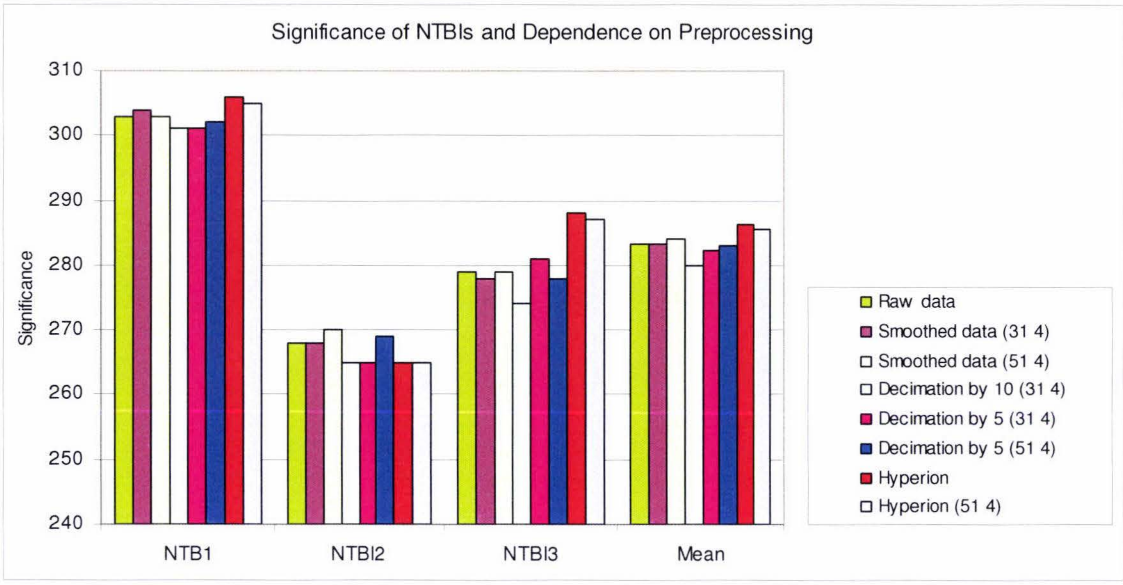


Figure 49: Frequency of statistically significant differences of NDVIs and their dependence on pre-processing

4.1.4.3 PCT

As a first step a principal component analysis was carried out on the Hyperion synthesized and Decimation 5 (pre-filtered with filter size 51 and order 4) data. The first 18 components including the eigenvalue, proportion and cumulative proportion of each dataset are shown in Table 11. The eigenvalues were in both cases rapidly falling (see Figure 50) which indicated that the principal component transformed data would require only around 10 components to explain most of the variation found in the data.

Table 11: First 18 components of the eigenanalysis of Hyperion-synthesized and Decimation by 5 data

Hyperion				Decimation by 5			
PC#	Eigenvalue	Proportion	Cumulative	PC#	Eigenvalue	Proportion	Cumulative
1	1.8699	0.890947	0.890947	1	3.5218	0.893990	0.89399
2	0.1858	0.088525	0.979472	2	0.3344	0.084881	0.97887
3	0.0215	0.010247	0.989719	3	0.0422	0.010716	0.989587
4	0.01	0.004774	0.994493	4	0.0203	0.005158	0.994744
5	0.0051	0.002446	0.996939	5	0.0089	0.002254	0.996998
6	0.0019	0.000899	0.997838	6	0.0032	0.000813	0.997811
7	0.0012	0.000573	0.99841	7	0.0025	0.000647	0.998459
8	0.0009	0.000422	0.998832	8	0.0016	0.000408	0.998866
9	0.0007	0.000318	0.99915	9	0.0012	0.000312	0.999178
10	0.0004	0.000178	0.999328	10	0.0008	0.000204	0.999382
11	0.0003	0.000136	0.999463	11	0.0005	0.000134	0.999515
12	0.0002	0.000096	0.999559	12	0.0004	0.000099	0.999614
13	0.0002	0.000080	0.999639	13	0.0003	0.000080	0.999695
14	0.0001	0.000050	0.999689	14	0.0002	0.000046	0.99974
15	0.0001	0.000045	0.999734	15	0.0001	0.000035	0.999776
16	0.0001	0.000027	0.999761	16	0.0001	0.000030	0.999806
17	0.0001	0.000026	0.999787	17	0.0001	0.000021	0.999827
18	0	0.000023	0.99981	18	0.0001	0.000019	0.999846

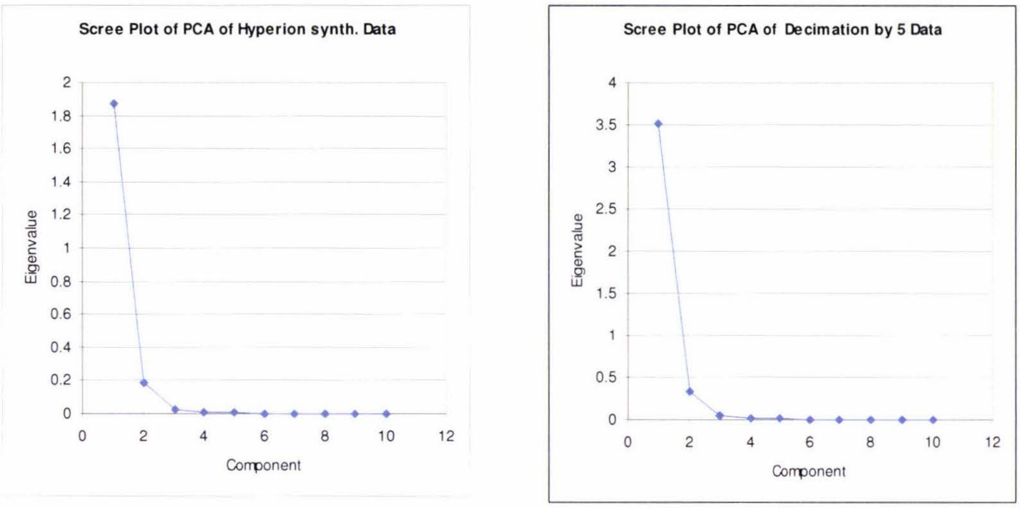


Figure 50: Scree plots of eigenvalues (Hyperion synthesized and Decimation 5 data)

The discriminating power of the components in the new feature space was assessed by applying the Wilcoxon test to all possible species pairs at a significance level of 0.01. Interestingly the frequency of significant differences was not strictly tied to the order of the components. Table 12 lists the ten components with the highest frequencies ordered by frequency. Component 11 had thus the highest discriminating power, followed by components 7, 8 and 10. These were followed by the first four components which all had frequencies between 321 and 325. The last two of these top ten components were of order 25 and 18, i.e. the highest frequencies are found the first sixth of all components. There was however a general drop in frequencies with increasing component order (see Figure 51).

Table 12: The 10 principal components with the highest significances (according to the Wilcoxon test) ordered by significance

Order	Significance
11	365
7	360
8	347
10	327
1	325
3	324
4	323
2	321
25	308
18	298

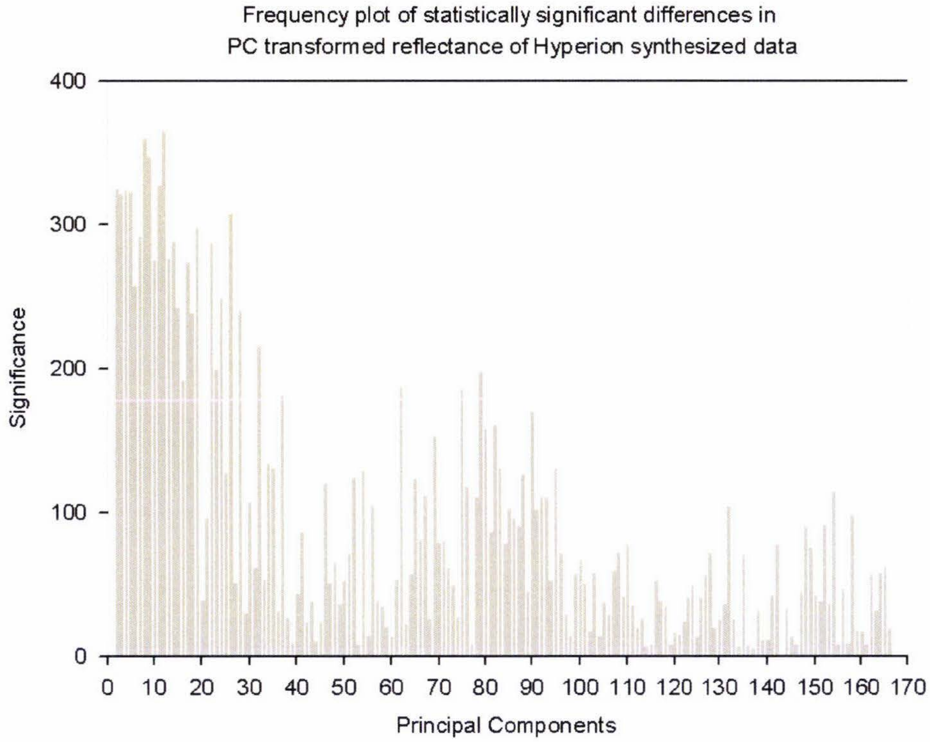


Figure 51: Histogram of statistically significant differences between species pairs for PC transformed Hyperion synthesized data

An analysis of the factor loadings of the components gave indications about the importance of the wavelengths. The average factor loadings were calculated for the visible, NIR, SWIR1 and SWIR2 segments (see Figure 52). Component 1 had the highest factor loadings in the SWIR1 and the lowest loadings in the visible. PC2 was dominated by NIR and SWIR2, PC3 by the visible, PC4 by SWIR1 and PC5 by visible. Interestingly, the coefficient plots formed shapes that were the negative (for PC1) and the positive (for PC2) of the typical spectral vegetation features (see Figure 53)

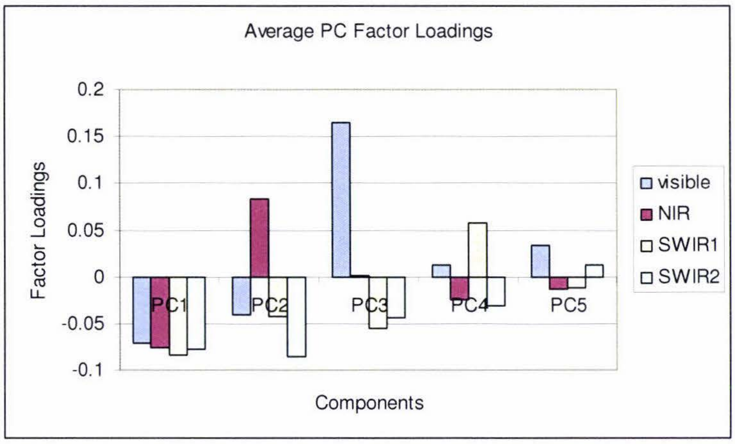


Figure 52: Average PC Factor Loadings for the first five components

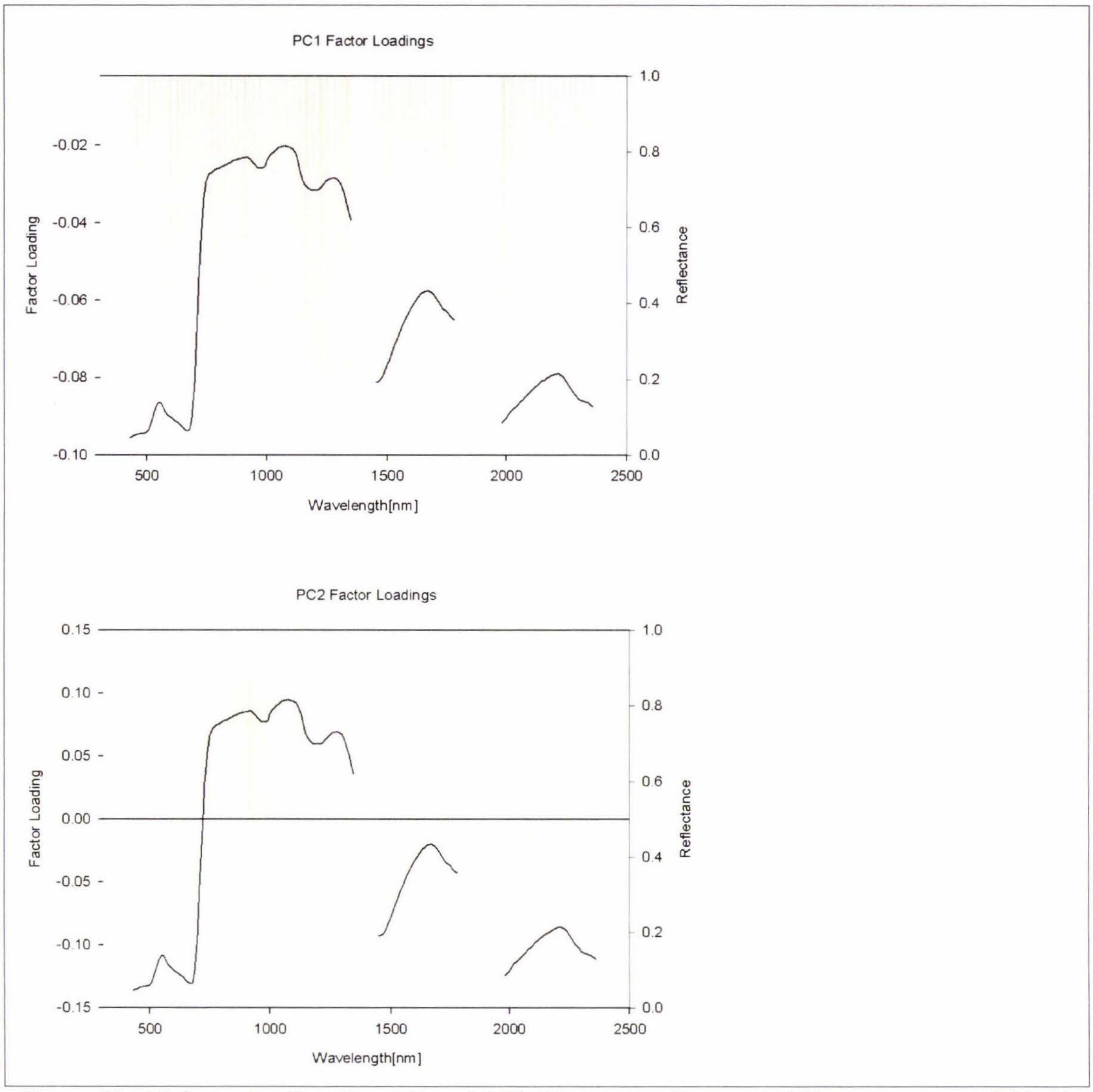


Figure 53: PC factor loadings for PC1 and PC2. The mean reflectance of Pittosporum eugenioides is displayed to relate the factors to typical vegetation reflectance features

4.1.5 Statistical Analysis

4.1.5.1 Discriminant Analysis

Classifications were carried out using three different discriminant functions: quadratic distance, general squared distance and SAM.

Two different datasets were classified: (a) the calibration data, i.e. the same data that were used to collect the statistical information used in the classification and (b) an independent dataset that contained the spectra of 15 species.

Smoothed Hyperion synthesized data was used to build three different feature spaces: DGVIs, NTBIs and PCT. The PCT was carried out using the first 25 components, based on the result of the Wilcoxon test on the PCT data that indicated that the highest frequencies of statistically significant differences occurred in the first 25 components.

The overall classification accuracy was found to be dependent on the feature space and the discriminant function (see Table 13). The highest overall accuracy was achieved by PC transformed data with 96.94% for the calibration dataset and 87.87% for the independent dataset respectively. The NTBI feature space was the least discriminating, which is probably directly related to its low dimensionality.

Error matrices were compiled for all classifications listing the correctly classified spectra per species, the errors of omission and commission, the row and column totals, the total number of classified spectra, the overall accuracy and the producer and user accuracies. Table 14 shows an example of an error matrix for the classification of the training data set in DGVI feature space using the quadratic distance discrimination function. *Metrosideros excelsa* had the lowest producer accuracy (20.00%) with only 8 out of 40 spectra being classified correctly. The omission errors in this case were: *Phormium tenax* (14), *Myoporum laetum* (12), *Macropiper excelsum* (3), *Corynocarpus laevigatus* (2) and *Hebe stricta* (1). A total of 1046 spectra were classified. The trace of the error matrix divided by total number of spectra gave an overall accuracy of 83.46. The minimum, maximum and mean of the producer and user accuracies were also calculated. For the above example, the minimum accuracy occurred in the producer accuracy (20%). The average user accuracy (90.93%) was higher than the mean producer accuracy (82.73%) (see Table 15).

Table 13: Classification results for calibration and independent datasets (accuracy in percentage)

	DGVI		NTBI		PCT	
	Calib. Set	Indep. Set	Calib. Set	Indep. Set	Calib. Set	Indep. Set
Quadratic distance	83.46	72.39	16.63	20.87	82.98	64.98
Gen. squared distance	75.04	72.05	28.2	23.56	96.94	87.87
SAM	33.07	31.98	19.4	16.16	35.85	22.22

Table 14: Error matrix for DGVIs of smoothed Hyperion synthesized data classified using the quadratic distance discriminant function

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Cyathea medullaris	Griselinia littoralis	Cordyline australis	Phormium tenax	Nothofagus truncata	Libocedrus bidwillii	Corynocarpus laevigatus	Coprosma robusta	Agathis australis	Macropiper excelsum	Hebe stricta	Pittosporum eugenioides	Leptospermum scoparium	Myrsine australis	Nothofagus solandri	Phyllocladus alpinus
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	31	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	43	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	32	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	23	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	23	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	14	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	41	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	28	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	55	[illegible]	[illegible]	[illegible]	[illegible]
13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	49	[illegible]	[illegible]	[illegible]
14	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	45	[illegible]	[illegible]
15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	26	[illegible]
16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	15
17	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
21	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
22	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
23	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
24	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
25	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
27	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
28	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
29	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
30	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
31	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
32	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Column Total	42	18	31	45	37	21	26	33	18	43	32	55	73	45	26	18
Producer Accuracy	92.86	88.89	100.00	95.56	86.49	42.86	88.46	69.70	77.78	95.35	87.50	100.00	67.12	100.00	100.00	83.33

Table 14 continued: Error matrix for DGVIs of smoothed Hyperion synthesized data classified using the quadratic distance discriminant function. Shaded cells are mentioned in the text

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32		
	Myoporum laetum	Hedycarya arborea	Pimelea buxifolia	Halocarpus biformis	Metrosideros excelsa	Brachyglottis repanda	Knightia excelsa	Dacrydium cupressinum	Cyathea dealbata	Nothofagus menziesii	Gleichenia dicarpa var. alpina	Cortaderia richardii	Podocarpus totara	Dicksonia squarrosa	Melicytus ramiflorus	Dacrycarpus dacrydioides	Row Total	User Accuracy
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	92.86
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	100.00
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	93.94
4	0	7	0	0	14	0	0	2	0	0	0	0	0	0	1	0	95	45.26
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	100.00
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	100.00
7	0	0	0	1	2	0	0	1	0	0	0	0	0	0	0	0	30	76.67
8	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	27	85.19
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	100.00
10	0	4	0	0	3	0	0	1	0	0	0	0	0	0	4	0	63	65.08
11	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	37	94.59
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	59	98.31
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	49	100.00
14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	49	91.84
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	100.00
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	100.00
17	47	0	0	0	12	0	0	1	0	0	0	0	0	0	0	0	66	71.21
18	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	100.00
19	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	17	100.00
20	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	22	100.00
21	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	8	100.00
22	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	15	100.00
23	0	0	1	0	0	0	21	3	0	0	0	0	0	0	0	2	28	75.00
24	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	7	100.00
25	0	1	0	0	0	0	0	2	35	0	0	0	0	0	0	5	66	53.03
26	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	23	100.00
27	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	18	100.00
28	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	27	100.00
29	0	0	0	4	0	0	0	3	0	0	0	0	42	0	0	0	56	75.00
30	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	19	100.00
31	0	5	0	0	0	0	0	0	0	0	0	0	0	0	55	0	60	91.67
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	12	100.00
Col. Total	48	23	18	27	40	15	21	20	35	27	18	27	42	19	60	20	1046	
Prod. Acc.	97.92	26.09	94.44	81.48	20.00	100.00	100.00	35.00	100.00	85.19	100.00	100.00	100.00	100.00	91.67	60.00		83.46

Table 15: Producer and user accuracy statistics for DGVIs of smoothed Hyperion synthesized data classified by the quadratic distance discriminant function (accuracy in percentage)

Prod. Acc. Min	20.00	User Acc. Min	45.26
Prod. Acc. Max	100.00	User Acc. Max	100.00
Prod. Acc. Mean	82.73	User Acc. Mean	90.93

4.1.5.2 Separability Analysis

The distances between the species in feature space were measured by calculating the JM and the Bhattacharya distances. The analysis was carried out on the following datasets:

- □ DGVIs of Hyperion synthesized data
- □ DGVIs of Hyperion synthesized data pre-smoothed with a Savitzky-Golay filter of size 51 and order 4
- □ PCT of Hyperion synthesized data using the first 25 components (data pre-smoothed with a Savitzky-Golay filter of size 51 and order 4)

As an example, the matrix showing JM and B distances in the upper and lower triangles respectively is presented for the DGVIs of Hyperion synthesized data in Table 17. The best separability was achieved by the PCT data with a mean JM distance of 2.00 (see Table 16). Out of a total of 496 species pairs 495 (99.79%) had a JM distance > 1.99. A JM value of 2.0 indicates full separability and a value of 1.9 a good separability. PCT data therefore achieved a very good separability while DGVI data with a mean of 1.82 still contained some overlaps of species distributions. Interestingly the DGVIs calculated from pre-smoothed Hyperion data did not perform better than the DGVIs based on non pre-smoothed Hyperion data as could be expected with regard to the results of the Wilcoxon test of the DGVIs.

The B distance measure produced similar results to the JM distance with PC transformed data having the best separability. In fact, due to limits in numerical precision, some B distances were infinite.

Table 16: Statistics of separability analysis

	DGVIs (Hyperion)	DGVI (smoothed Hyperion)	PCT (smoothed Hyperion)
JM Min	1.23	1.17	1.99
JM Max	1.98	1.98	2.00
JM Mean	1.82	1.82	2.00
B Min	0.96	0.88	5.20
B Max	4.55	4.50	∞
B Mean	2.59	2.65	∞

Table 17: JM distances (upper triangle) and B distances (lower triangle) between species in DGVI feature space

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Cyathea medullaris	Griselinia littoralis	Cordyline australis	Phormium tenax	Nothofagus truncata	Libocedrus bidwillii	Corynocarpus laevigatus	Coprosma robusta	Agathis australis	Macropiper excelsum	Hebe stricta	Pittosporum eugenioides	Leptospermum scoparium	Myrsine australis	Nothofagus solandri	Phylocladus alpinus
1	-	1.90	1.85	1.72	1.78	1.89	1.79	1.86	1.82	1.65	1.53	1.68	1.79	1.63	1.90	1.96
2	2.99	-	1.92	1.85	1.76	1.78	1.82	1.79	1.94	1.88	1.76	1.84	1.87	1.89	1.82	1.93
3	2.61	3.17	-	1.64	1.80	1.89	1.83	1.89	1.93	1.84	1.74	1.80	1.87	1.83	1.91	1.97
4	1.97	2.59	1.72	-	1.64	1.90	1.76	1.74	1.86	1.72	1.30	1.59	1.59	1.45	1.84	1.97
5	2.23	2.10	2.32	1.70	-	1.67	1.77	1.69	1.82	1.79	1.51	1.71	1.78	1.82	1.76	1.83
6	2.91	2.19	2.91	2.96	1.79	-	1.86	1.84	1.88	1.92	1.84	1.89	1.87	1.93	1.72	1.76
7	2.24	2.42	2.48	2.12	2.17	2.69	-	1.82	1.90	1.76	1.76	1.71	1.88	1.90	1.86	1.95
8	2.67	2.27	2.95	2.04	1.86	2.51	2.40	-	1.93	1.81	1.66	1.78	1.75	1.82	1.78	1.95
9	2.40	3.54	3.42	2.67	2.43	2.84	2.96	3.32	-	1.90	1.81	1.85	1.83	1.90	1.92	1.96
10	1.74	2.85	2.54	1.97	2.24	3.27	2.10	2.33	2.97	-	1.64	1.51	1.86	1.66	1.86	1.97
11	1.45	2.11	2.04	1.05	1.40	2.52	2.13	1.78	2.37	1.71	-	1.45	1.70	1.60	1.85	1.95
12	1.82	2.55	2.32	1.58	1.93	2.90	1.93	2.22	2.60	1.41	1.28	-	1.80	1.67	1.83	1.96
13	2.27	2.76	2.77	1.59	2.20	2.73	2.79	2.07	2.45	2.64	1.91	2.32	-	1.69	1.80	1.95
14	1.68	2.87	2.49	1.30	2.40	3.41	2.95	2.43	3.01	1.76	1.60	1.80	1.88	-	1.84	1.97
15	3.01	2.43	3.14	2.53	2.13	1.95	2.68	2.22	3.19	2.68	2.60	2.47	2.28	2.51	-	1.84
16	3.86	3.32	4.17	4.10	2.46	2.12	3.72	3.72	3.87	4.22	3.64	3.99	3.73	4.35	2.50	-
17	1.97	2.73	1.82	1.09	2.25	3.26	2.41	2.44	2.87	2.06	1.26	1.76	2.19	1.41	2.83	4.55
18	2.68	2.31	3.03	2.19	1.37	2.23	2.36	2.21	3.09	2.45	1.74	2.15	2.18	2.85	2.15	2.61
19	3.52	3.21	2.60	3.47	2.62	2.35	2.90	3.38	3.44	3.48	3.26	3.21	3.81	4.04	2.72	3.57
20	3.49	2.83	3.24	3.62	1.96	1.83	3.34	3.07	3.50	3.84	3.15	3.39	3.25	3.75	2.00	1.81
21	2.08	1.71	2.10	1.46	1.20	2.05	1.92	1.28	2.50	1.78	0.96	1.49	2.02	1.94	1.76	3.27
22	2.97	3.34	3.57	2.42	2.29	3.24	2.93	2.89	3.04	2.41	2.03	1.75	3.11	2.61	3.14	4.03
23	2.75	3.00	2.81	2.62	2.60	2.84	2.03	2.69	3.66	2.74	2.46	2.38	3.26	3.00	2.66	4.23
24	2.94	2.43	2.78	2.86	2.07	1.55	2.71	2.81	3.12	2.99	2.35	2.33	2.89	2.80	1.99	2.48
25	1.34	3.04	2.83	1.87	2.34	3.20	2.34	2.73	1.98	1.52	1.64	1.51	2.35	1.73	3.01	4.36
26	3.33	2.40	3.40	2.84	1.78	2.01	2.62	2.18	2.89	3.38	2.53	2.90	2.65	3.47	2.08	2.48
27	2.77	2.41	2.66	2.44	1.62	2.06	2.03	2.19	2.96	2.41	2.32	2.20	2.56	2.93	1.47	2.52
28	2.52	2.92	2.09	2.23	2.64	2.90	1.95	2.89	3.32	1.96	2.21	1.85	2.84	2.63	2.80	3.94
29	2.19	3.14	2.75	1.92	2.29	2.95	2.51	2.54	2.71	1.97	1.83	1.82	2.53	1.76	2.56	4.05
30	2.75	3.10	3.06	2.93	2.58	2.69	2.26	2.78	2.73	2.80	2.65	2.66	3.34	3.31	2.84	4.17
31	1.74	2.11	2.75	1.58	1.68	2.69	1.98	1.91	2.54	1.56	1.41	1.39	2.27	1.75	2.34	3.84
32	2.76	3.10	3.57	3.09	2.05	2.28	3.08	3.15	2.96	3.15	2.67	2.96	2.52	2.77	2.61	2.40

Table 17 continued: JM distances (upper triangle) and B distances (lower triangle) between species in DGVI feature space

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
	Myoporum laetum	Hedycarya arborea	Pimelea buxifolia	Halocarpus biformis	Metrosideros excelsa	Brachyglottis repanda	Knightia excelsa	Dacrydium cupressinum	Cyathea dealbata	Nothofagus menziesii	Gleichenia dicarpa var. alpina	Cortaderia richardii	Podocarpus totara	Dicksonia squarrosa	Melicytus ramiflorus	Dacrycarpus dacrydioides
1	1.72	1.86	1.94	1.94	1.75	1.90	1.87	1.89	1.48	1.93	1.87	1.84	1.78	1.87	1.65	1.87
2	1.87	1.80	1.92	1.88	1.64	1.93	1.90	1.82	1.90	1.82	1.87	1.89	1.91	1.91	1.76	1.91
3	1.67	1.90	1.85	1.92	1.76	1.94	1.88	1.88	1.88	1.93	1.86	1.75	1.87	1.91	1.87	1.94
4	1.33	1.78	1.94	1.95	1.53	1.82	1.85	1.89	1.69	1.88	1.83	1.78	1.71	1.89	1.59	1.91
5	1.79	1.49	1.85	1.72	1.40	1.80	1.85	1.75	1.81	1.66	1.60	1.86	1.80	1.85	1.63	1.74
6	1.92	1.79	1.81	1.68	1.74	1.92	1.88	1.58	1.92	1.73	1.74	1.89	1.90	1.86	1.86	1.80
7	1.82	1.81	1.89	1.93	1.71	1.89	1.74	1.87	1.81	1.85	1.74	1.71	1.84	1.79	1.72	1.91
8	1.83	1.78	1.93	1.91	1.45	1.89	1.86	1.88	1.87	1.77	1.78	1.89	1.84	1.88	1.70	1.91
9	1.89	1.93	1.94	1.94	1.84	1.90	1.95	1.91	1.72	1.89	1.90	1.93	1.87	1.87	1.84	1.91
10	1.74	1.83	1.94	1.96	1.66	1.82	1.87	1.90	1.56	1.93	1.82	1.72	1.72	1.88	1.58	1.91
11	1.43	1.65	1.92	1.91	1.23	1.74	1.83	1.81	1.61	1.84	1.80	1.78	1.68	1.86	1.51	1.86
12	1.66	1.77	1.92	1.93	1.55	1.65	1.82	1.81	1.56	1.89	1.78	1.69	1.68	1.86	1.50	1.90
13	1.78	1.77	1.96	1.92	1.73	1.91	1.92	1.89	1.81	1.86	1.84	1.88	1.84	1.93	1.79	1.84
14	1.51	1.88	1.96	1.95	1.71	1.85	1.90	1.88	1.65	1.94	1.89	1.86	1.65	1.93	1.65	1.87
15	1.88	1.77	1.87	1.73	1.66	1.91	1.86	1.73	1.90	1.75	1.54	1.88	1.84	1.88	1.81	1.85
16	1.98	1.85	1.94	1.67	1.92	1.96	1.97	1.83	1.97	1.83	1.84	1.96	1.97	1.97	1.96	1.82
17		1.84	1.93	1.96	1.53	1.82	1.85	1.88	1.67	1.93	1.86	1.72	1.79	1.90	1.63	1.92
18	2.50	-	1.91	1.85	1.54	1.72	1.89	1.81	1.86	1.83	1.67	1.90	1.89	1.89	1.64	1.77
19	3.31	3.16	-	1.88	1.86	1.97	1.92	1.86	1.95	1.93	1.79	1.88	1.94	1.93	1.89	1.96
20	3.91	2.57	2.78	-	1.86	1.95	1.94	1.71	1.96	1.76	1.66	1.94	1.93	1.94	1.93	1.82
21	1.44	1.47	2.65	2.68	-	1.70	1.79	1.74	1.74	1.79	1.59	1.72	1.73	1.81	1.49	1.86
22	2.43	1.96	4.23	3.79	1.90	-	1.91	1.91	1.80	1.93	1.88	1.92	1.90	1.92	1.75	1.94
23	2.61	2.95	3.22	3.55	2.24	3.16		1.82	1.89	1.90	1.88	1.83	1.85	1.88	1.84	1.95
24	2.84	2.35	2.64	1.92	2.04	3.09	2.39	-	1.91	1.81	1.86	1.87	1.85	1.91	1.86	1.82
25	1.81	2.67	3.75	3.98	2.02	2.31	2.88	3.12	-	1.94	1.88	1.81	1.72	1.88	1.64	1.88
26	3.31	2.46	3.38	2.10	2.24	3.33	3.02	2.38	3.43	-	1.76	1.93	1.89	1.87	1.88	1.86
27	2.65	1.80	2.25	1.77	1.58	2.84	2.83	2.29	2.81	2.11	-	1.84	1.85	1.85	1.70	1.86
28	1.98	3.00	2.85	3.45	1.98	3.18	2.46	2.75	2.33	3.37	2.51	-	1.84	1.87	1.80	1.94
29	2.24	2.91	3.49	3.39	2.00	3.00	2.62	2.60	1.97	2.90	2.60	2.55	-	1.90	1.71	1.89
30	2.95	2.94	3.36	3.49	2.37	3.23	2.82	3.13	2.80	2.71	2.59	2.76	2.96	-	1.89	1.93
31	1.69	1.71	2.93	3.33	1.36	2.10	2.51	2.69	1.71	2.80	1.91	2.30	1.93	2.88	-	1.86
32	3.28	2.18	3.82	2.42	2.68	3.44	3.60	2.33	2.81	2.64	2.63	3.53	2.89	3.30	2.69	-

4.1.5.3 Most Discriminating Bands

The results of the Wilcoxon test with a significance level of 0.01 for raw spectra, Hyperion synthesized spectra and 1st derivative of Hyperion synthesized spectra are graphically depicted as histograms in Figures 55-57. Table 18 lists the overall maximum, minimum, mean and standard deviation frequency and the same measurements for the visible (350-670nm), NIR (671-1349), SWIR1 (1441-1789nm) and SWIR2 (1981-2359). These segments divide the spectrum by the position of the red edge and the filtered water bands.

Table 18: Significance statistics

	Raw		**Hyperion**		**1st Derivative of Hyperion**	
	Significance	%	Significance	%	Significance	%
Min	227	45.8	261	52.6	10	2.0
Max	366	73.8	364	73.4	424	85.5
Mean	328.6	66.2	325.2	65.6	287.7	58.0
Stddev	22.1	4.5	21.8	4.4	89.5	18.0
Mean + Stddev	350.7	70.7	347.0	70.0	377.2	76.1
Min Visible	282	56.9	283	57.1	285	57.5
Max Visible	361	72.8	359	72.4	371	74.8
Mean Visible	327.7	66.1	325.6	65.6	335.8	67.7
Stddev Visible	18.5	3.7	22.2	4.5	23.9	4.8
Min NIR	282	56.9	286	57.7	38	7.7
Max NIR	343	69.2	343	69.2	403	81.3
Mean NIR	317.6	64.0	317.5	64.0	302.7	61.0
Stddev NIR	11.4	2.3	11.0	2.2	53.1	10.7
Min SWIR1	324	65.3	329	66.3	209	42.1
Max SWIR1	366	73.8	364	73.4	424	85.5
Mean SWIR1	350.7	70.7	350.8	70.7	357.2	72.0
Stddev SWIR1	9.4	1.9	9.0	1.8	40.8	8.2
Min SWIR2	227	45.8	261	52.6	10	2.0
Max SWIR2	355	71.6	354	71.4	319	64.3
Mean SWIR2	328.6	66.2	332.5	67.0	165.9	33.4
Stddev SWIR2	31.1	6.3	28.3	5.7	84.9	17.1

Table 19: Number of bands with frequencies higher than mean plus one standard deviation

	Raw		**Hyperion**		**1st Derivative of Hyperion**	
	# of bands	%	# of bands	%	# of bands	%
Total	336.0	19.4	51.0	30.7	10.0	6.1
Visible	22.0	1.3	3.0	1.8	0.0	0.0
NIR	0.0	0.0	0.0	0.0	2.0	1.2
SWIR1	209.0	12.1	28.0	16.9	8.0	4.9
SWIR2	105.0	6.1	20.0	12.0	0.0	0.0

The maximum frequency for the raw data was 366 in the SWIR1 region at 1727nm, i.e. this wavelength was statistically significant different for 73.8% of all species pairings. The SWIR1 region also had the highest mean significance of 350.7. Separability was generally better in the visible portion of the spectrum than in the NIR. The lowest significance of the visible and NIR was found around 673nm which is the start of the red edge.

The significance for the Hyperion synthesized data was very similar to the raw data. The maximum was slightly lower by 0.4% while the minimum was higher by 6.8%. The average significance of Hyperion was 0.6% lower than that of the raw data.

The significance frequencies for both raw and Hyperion synthesized data varied with a standard deviation of 22.1 and 21.8 respectively.

For the first derivative of Hyperion synthesized data, the significance frequencies had a standard deviation of 89.5, thus the variations in frequency were much higher than for zero derivative data. The derivative data did not show any drastic decrease of frequencies in the red edge but dropped to a value of 38 around 1000nm where the spectra often show a little step due to the switch over of internal sensor elements.

A threshold was calculated by adding the standard deviation to the mean (see Table 18). The number of bands that had frequencies equal to or higher than this threshold was reported for the full spectrum and the visible, NIR, SWIR1 and SWIR2 segments (see Table 19). For the raw data 19.4% of all bands had a frequency of 350.7 or higher, i.e. were significantly different for at least 70.7% of all species pairings. For Hyperion synthesized data 30.7% of all bands had a frequency of 347 or higher which meant that at least 70% of all species pairs were significantly different at these bands.

Only 6.1% of all bands were equal to or higher than the threshold for the 1st derivative. These bands had significant differences for at least 76% of all species pairs.

Figure 54 compares the percentage of bands with frequencies higher than the threshold in the spectrum segments. The highest percentage for zero order derivatives was in the SWIR1 segment, followed by SWIR2 and visible. The NIR segment had no bands with frequencies above the threshold. SWIR1 recorded the highest percentage for the 1st derivative, followed by NIR with no bands above the threshold for the visible and SWIR2 segments.

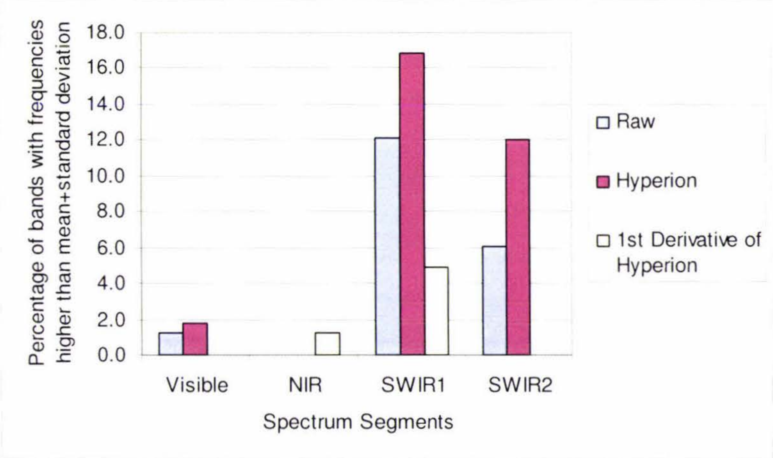


Figure 54: Graphical comparison of the number of bands with frequencies higher than the threshold (mean + standard deviation) per spectrum segment

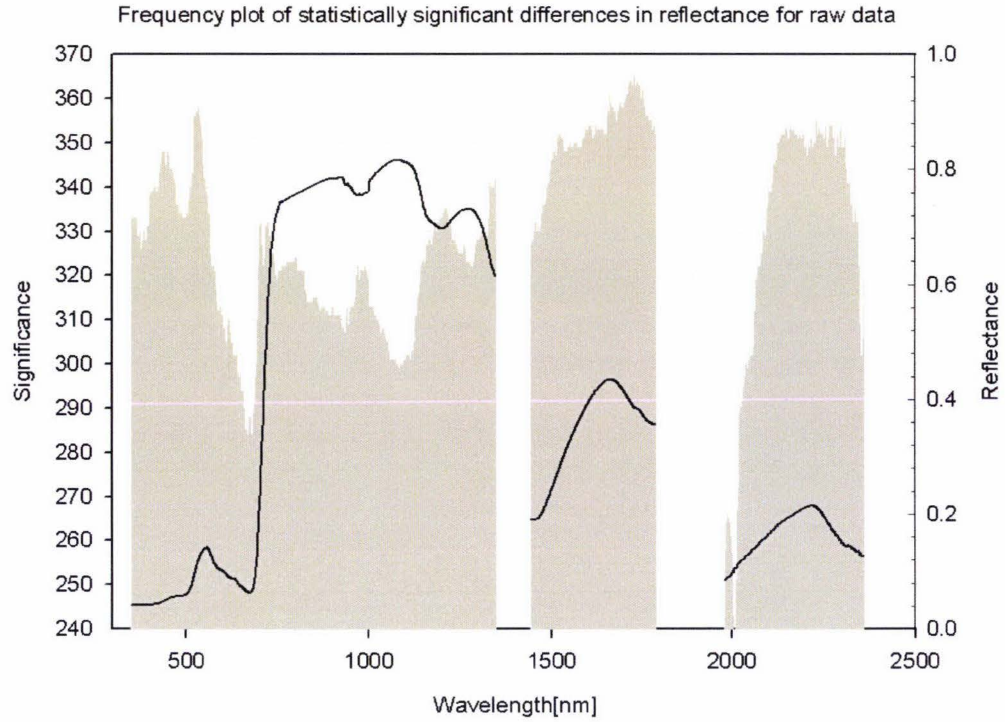


Figure 55: Histogram of the statistically significant differences in reflectance calculated using raw data of all library relevant species. The mean reflectance of Pittosporum eugenioides is displayed to relate the frequency to typical vegetation reflectance features.

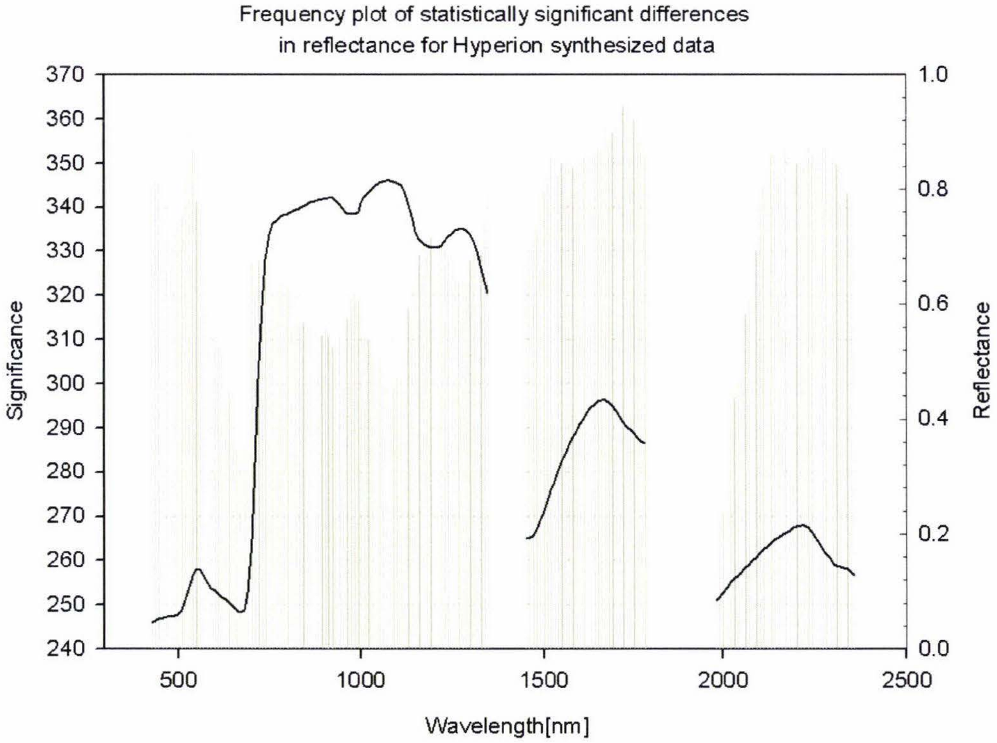


Figure 56: Histogram of the statistically significant differences in reflectance calculated using Hyperion synthesized data of all library relevant species. The mean reflectance of Pittosporum eugenioides is displayed to relate the frequency to typical vegetation reflectance features

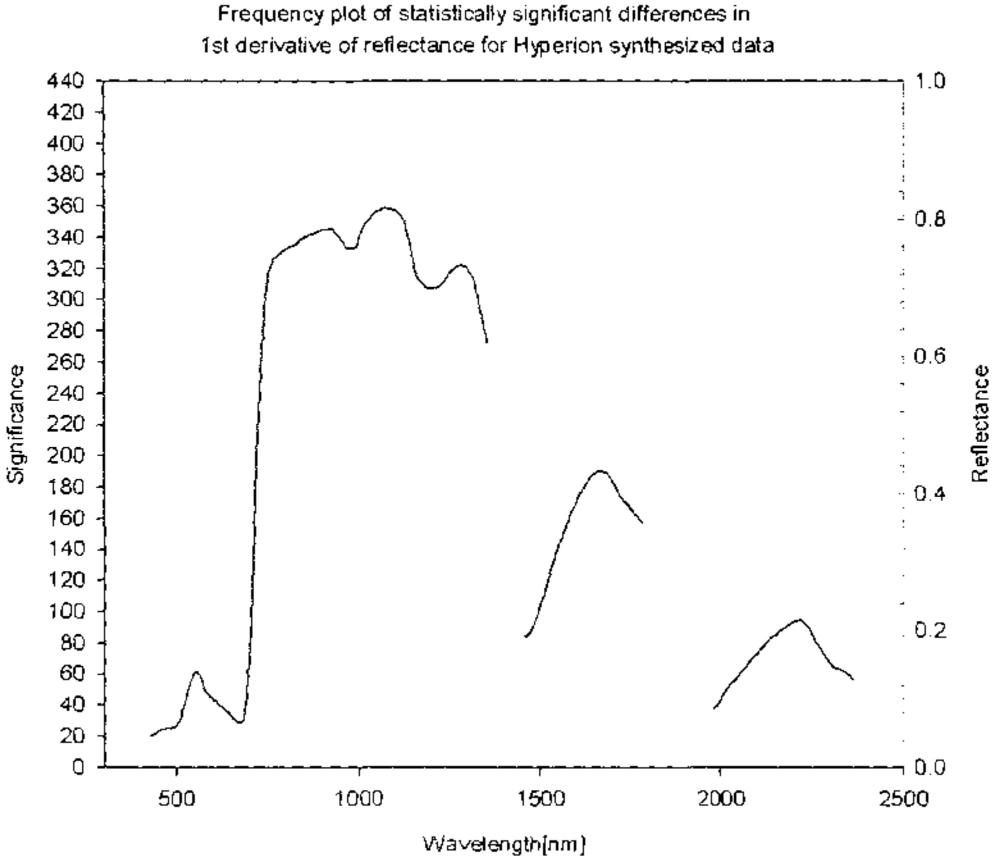


Figure 57: Histogram of the statistically significant differences in the first derivative of reflectance calculated using Hyperion synthesized data of all library relevant species. The mean reflectance of Pittosporum eugenioides is displayed to relate the frequency to typical vegetation reflectance features

4.2 Mixed Spectral Signatures

4.2.1 Paper/Plant Mixture

A spectral plot of the mixtures of kawakawa leaves and paper revealed that the endmembers were indeed encompassing their mixtures. The paper endmember defined the maximum and kawakawa the minimum of the spectral reflectances found (see Figure 58). The position of the mixtures was linearly dependent on the abundances, i.e. the higher the abundance of an endmember in the mixture, the closer the spectral curve was to the endmember curve.

This linearity was also well illustrated by synthesizing Landsat7 data and plotting band 1 against band 7 (see Figure 59).

Unmixing was carried out in Matlab using Hyperion synthesized data for endmembers and mixtures. The RMSE for the abundances was 2.68% (see Table 20).

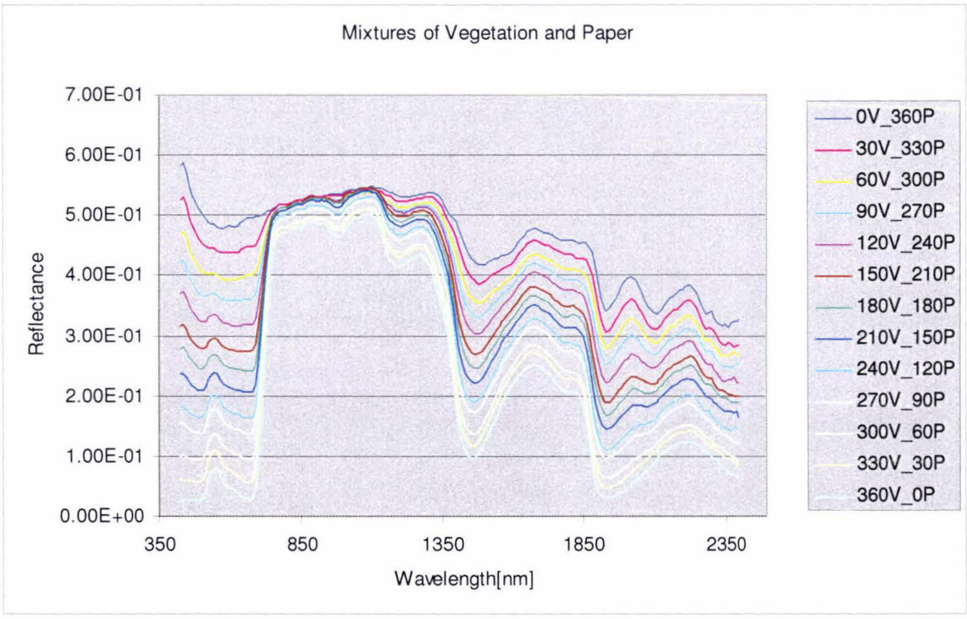


Figure 58: Spectral curves of mixtures of vegetation and paper

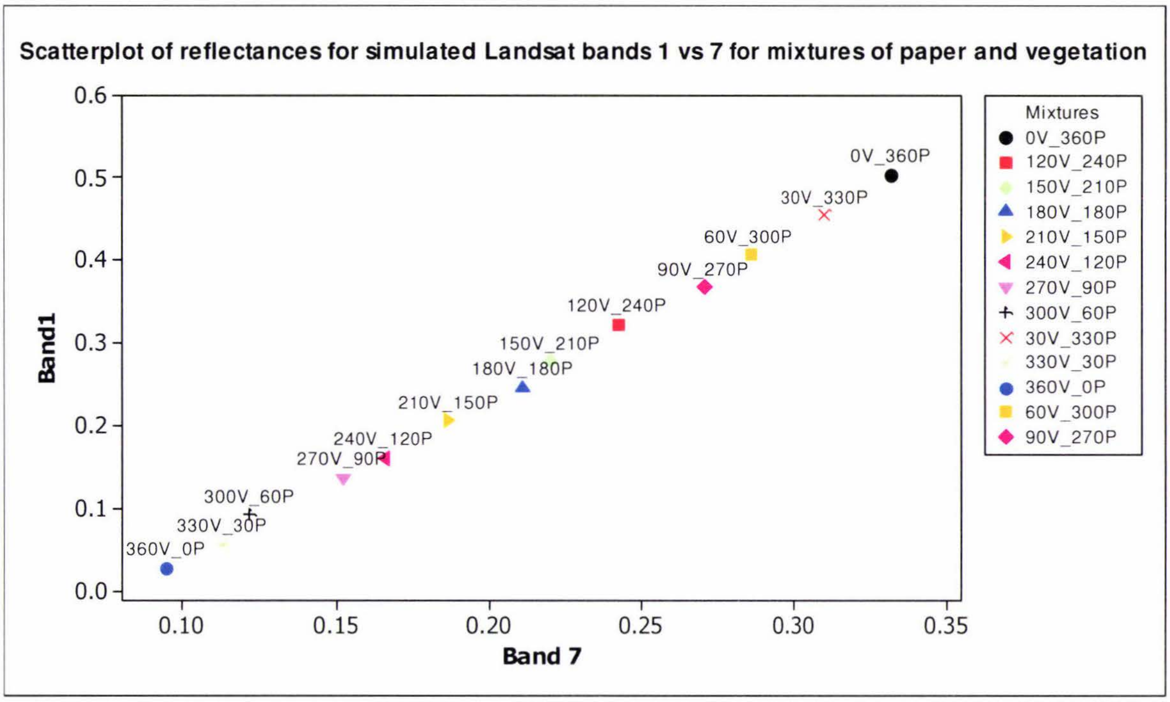


Figure 59: Scatterplot of reflectance values for simulated Landsat bands 1 vs 7 for mixtures of paper and vegetation

Table 20: Unmixing results for vegetation/paper mixtures

Mixture	Vegetation abundance [%]	Paper abundance [%]	Estimated vegetation abundance [%]	Estimated paper abundance [%]	Error [%]
30V_330P	8	92	10.03	89.97	2.03
60V_300P	17	83	19.69	80.31	2.69
90V_270P	25	75	27.40	72.60	2.40
120V_240P	33	67	36.00	64.00	3.00
150V_210P	42	58	45.94	54.06	3.94
180V_180P	50	50	52.46	47.54	2.46
210V_150P	58	42	60.24	39.76	2.24
240V_120P	67	33	70.12	29.88	3.12
270V_90P	75	25	75.75	24.25	0.75
300V_60P	83	17	86.95	13.06	3.95
330V_30P	92	8	92.94	7.06	0.94

4.2.2 Paper/Plastic/Plant Mixture

The mixture experiment involving three different endmembers was harder to interpret using the full spectral plots as the spectra for the endmembers vegetation and plastic overlapped (see Figure 60). However, by plotting reflectances in band 7 versus band 1 of Landsat7 synthesized data, it was evident that the endmembers define the extremes of the space that holds the mixtures (see Figure 61). The three endmembers define a triangle in this two dimensional space with the mixtures lying inside the boundaries of this triangle.

The unmixing was carried out using Hyperion synthesized data. The estimated abundances were generally in reasonable ranges. It is however worth noting that some abundances were negative, e.g. an estimated vegetation abundance of -0.59% for the V0_P180_PL180 mixture (see Table 21). This was due to the fact that the negativity constraint was not added to the unmixing procedure. The RMSE for vegetation, paper and plastic were 7.37%, 4.48% and 4.77% respectively.

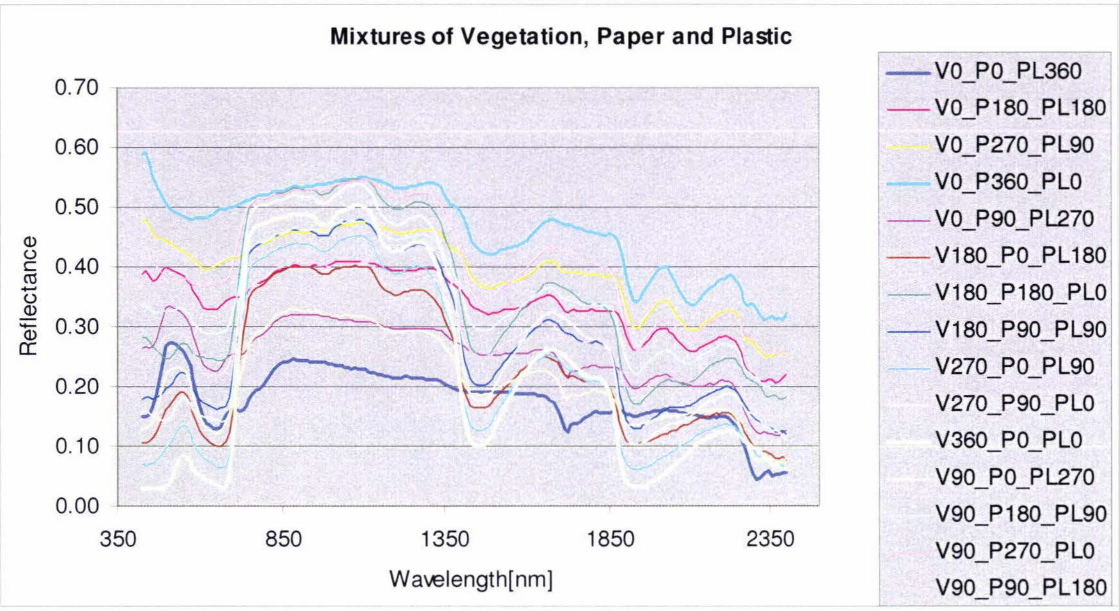


Figure 60: Spectral curves of mixtures of vegetation, paper and plastic. Endmembers are plotted in thick lines

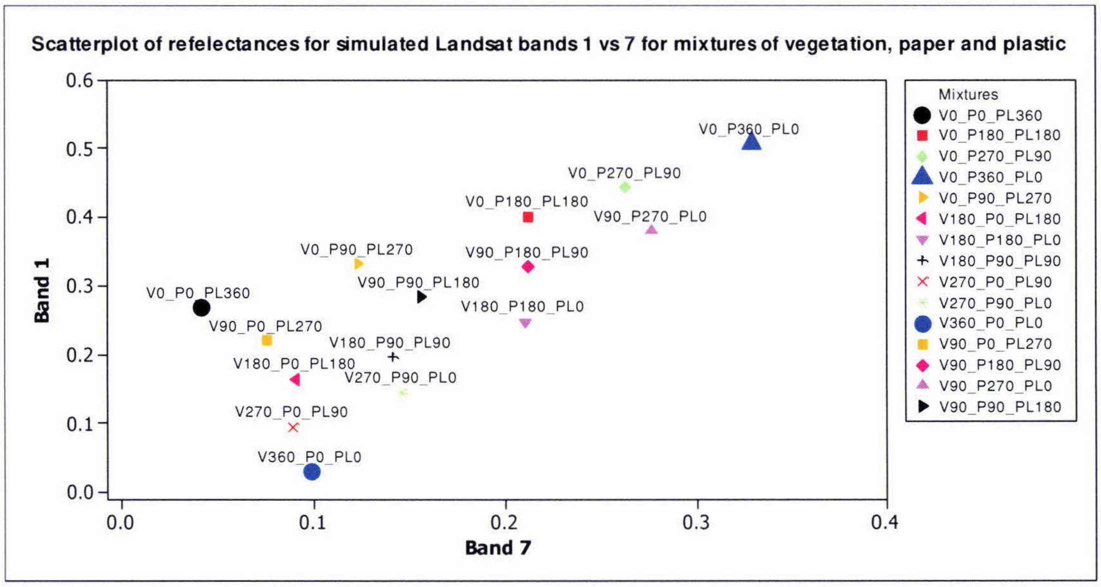


Figure 61: Scatterplot of reflectances for simulated Landsat bands 1 vs 7 for mixtures of vegetation, paper and plastic. Endmembers are plotted in bigger symbols

Table 21: Unmixing results for vegetation/paper/plastic mixtures

	Vegetation abundance [%]	Paper abundance [%]	Plastic abundance [%]	Estimated vegetation abundance [%]	Estimated paper abundance [%]	Estimated plastic abundance [%]	Error Vegetation [%]	Error Paper [%]	Error Plastic [%]
V0_P180_PL180	0	50	50	-0.59	56.27	44.32	-0.59	6.27	-5.68
V0_P270_PL90	0	75	25	0.09	75.51	24.40	0.09	0.51	-0.60
V0_P90_PL270	0	25	75	-0.39	26.38	74.00	-0.39	1.38	-1.00
V180_P0_PL180	50	0	50	53.89	7.92	38.19	3.89	7.92	-11.81
V180_P180_PL0	50	50	0	59.88	48.26	-8.14	9.88	-1.74	-8.14
V180_P90_PL90	50	25	25	59.01	26.53	14.47	9.00	1.53	-10.53
V270_P0_PL90	75	0	25	76.93	2.97	20.11	1.93	2.97	-4.89
V270_P90_PL0	75	25	0	82.12	25.36	-7.47	7.11	0.36	-7.47
V90_P0_PL270	25	0	75	26.51	6.34	67.15	1.51	6.34	-7.85
V90_P180_PL90	25	50	25	28.17	53.06	18.76	3.17	3.06	-6.24
V90_P270_PL0	25	75	0	27.91	75.53	-3.44	2.91	0.53	-3.44
V90_P90_PL180	25	25	50	27.01	33.52	39.48	2.01	8.52	-10.53

4.2.3 Three plant mixture

The endmembers were not discernible in the full spectral plots (see Figure 62). Furthermore, the endmembers no longer defined the boundaries in which the mixtures fell as is illustrated by plotting band 1 versus band 7 of Landsat7 synthesized data (see Figure 63). The results of the unmixing as shown in Table 22 had large errors for most of the estimated abundances with many percentages being negative. The RMSE for Kawakawa, Lemonwood and Karaka were 68.79%, 51.45% and 26.88% respectively.

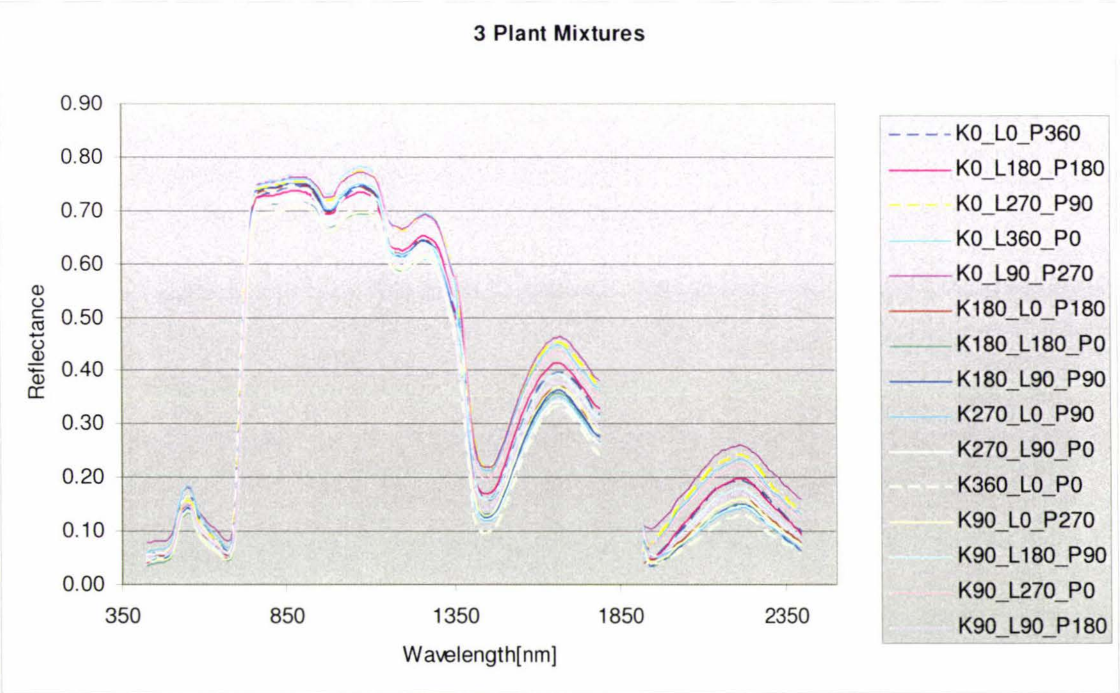


Figure 62: Spectral curves for mixtures of three plants. Endmembers are plotted as dashed lines.

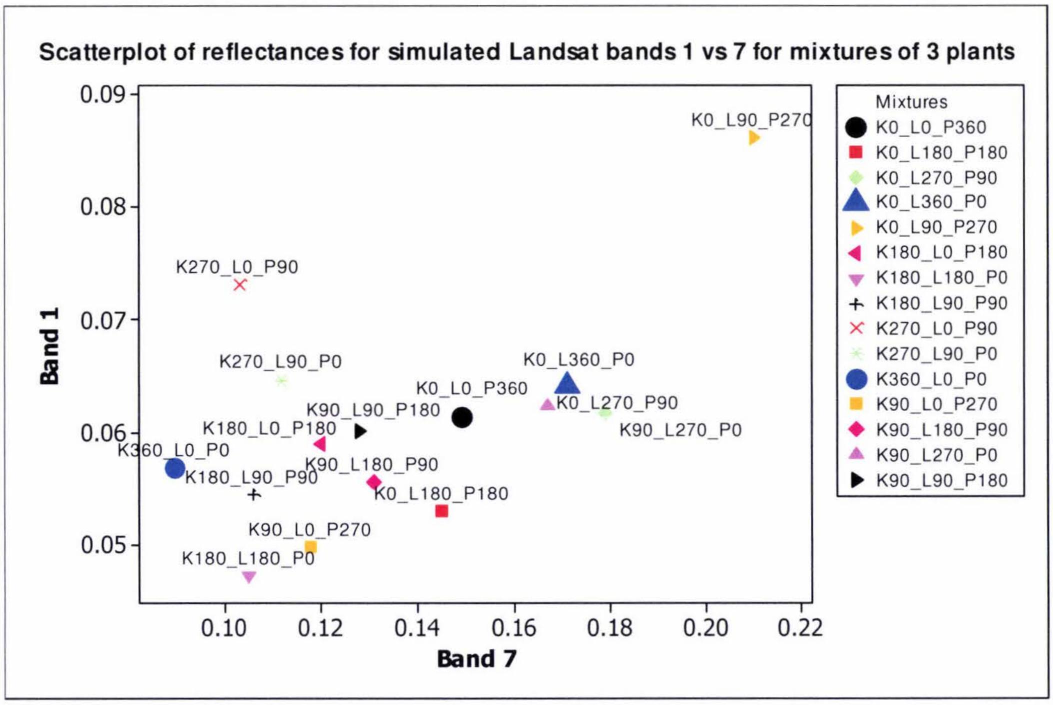


Figure 63: Scatterplot of reflectances for simulated Landsat bands 1 vs 7 for mixtures of 3 plants

Table 22: Unmixing results for 3 plant mixtures

	Kawakawa abundance [%]	Lemonwood abundance [%]	Karaka abundance [%]	Kawakawa abundance [%]	Lemonwood abundance [%]	Karaka abundance [%]	Error Kawakawa [%]	Error Lemonwood [%]	Error Karaka[%]
Kar0_L180_Kaw180	50	50	0	99.94	8.37	-8.31	49.94	-41.64	-8.31
Kar0_L270_Kaw90	25	75	0	53.83	72.89	-26.72	28.83	-2.11	-26.72
Kar0_L90_Kaw270	75	25	0	79.84	72.41	-52.24	4.84	47.41	-52.24
Kar180_L0_Kaw180	50	0	50	22.01	20.51	57.48	-27.99	20.51	7.48
Kar180_L180_Kaw0	0	50	50	126.96	-49.09	22.14	126.96	-99.09	-27.87
Kar180_L90_Kaw90	25	25	50	0.85	26.37	72.77	-24.15	1.37	22.77
Kar270_L0_Kaw90	25	0	75	-22.27	33.80	88.47	-47.27	33.80	13.47
Kar270_L90_Kaw0	0	25	75	107.89	-43.21	35.32	107.89	-68.21	-39.68
Kar90_L0_Kaw270	75	0	25	161.51	-62.95	1.44	86.51	-62.95	-23.56
Kar90_L180_Kaw90	25	50	25	95.76	-2.58	6.82	70.76	-52.58	-18.18
Kar90_L270_Kaw0	0	75	25	6.78	87.80	5.42	6.78	12.80	-19.58
Kar90_L90_Kaw180	50	25	25	147.75	-44.68	-3.07	97.75	-69.68	-28.07

4.2.4 Positional Dependence of Paper/Plastic Mixtures

The positional experiment confirmed that the position of both the tungsten lamp and the sun for illumination does influence the resulting spectra. The three different mixtures (plastic abundances of 0.25, 0.5 and 0.75) form three groups when plotted (see Figure 64). Ideally, with no positional dependence, the spectra of these groups should be identical. Two intra group differences could be observed: offsets of the spectral curves and shape differences. The shape differences are easily discernible between wavelengths 950-1180nm, 1300-1450nm and 1750-2000nm where a considerable difference in slope gradient between positions 3 and 4 and positions 1 and 2 can be seen. This is best illustrated by the 1st derivative, shown for the 50% mixtures in Figure 65.

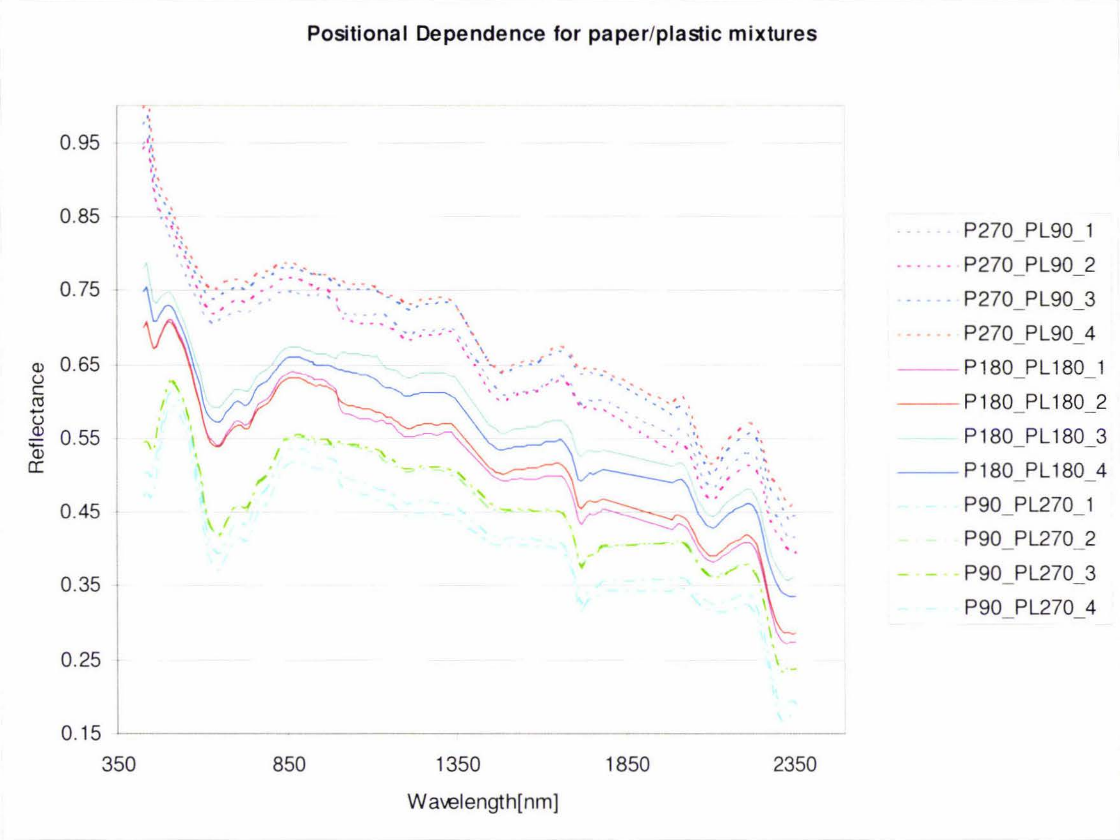


Figure 64: Positional dependence for paper/plastic mixtures

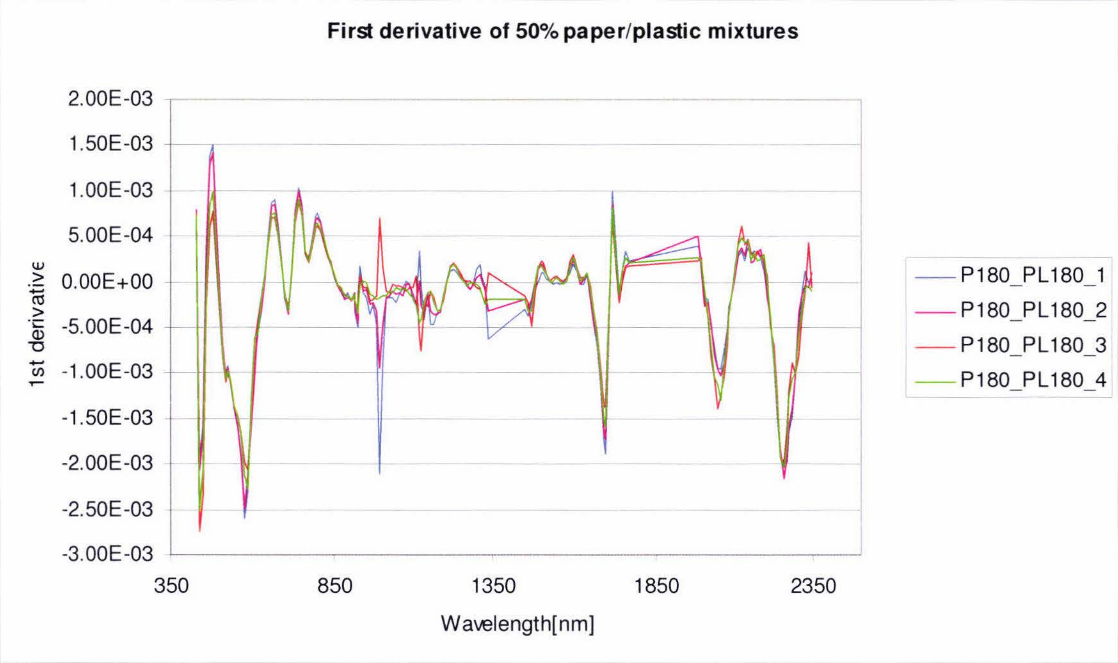


Figure 65: First derivative of 50% paper/plastic mixtures

4.2.5 Probe Rotation

Visually the four positions of the probe resulted in very similar spectra (see Figure 66). RMSE's were calculated between the mean and the four positional spectral curves for every wavelength. The mean of all the RMSE's was 0.00589.

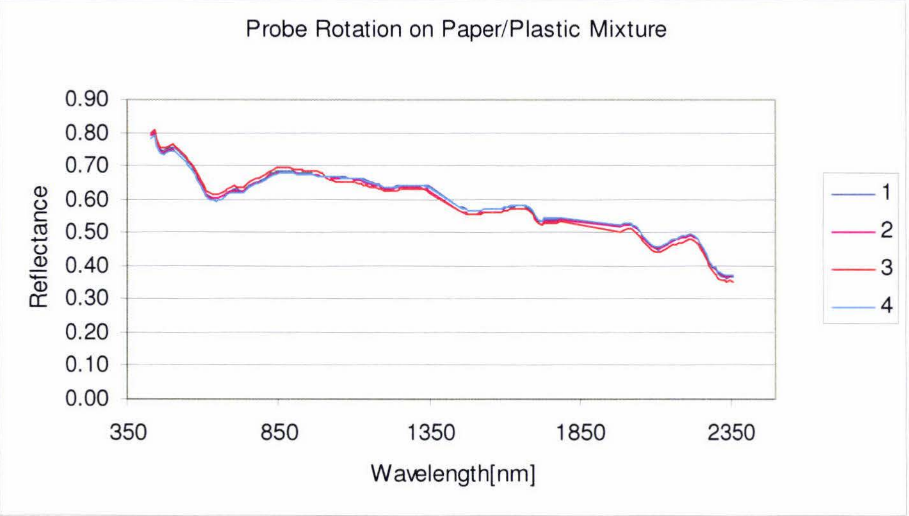


Figure 66: Probe Rotation on paper/plastic mixture

5 Discussion

5.1 Collection of Spectral Data of New Zealand Native Plants

The collection of spectral data of New Zealand native plants that was assembled during this research provides a valuable data source for future research. It must however be noted that its use will be restricted by the fact that critical metadata like atmospheric condition was not collected. A further restriction is the temporal change of spectra that is not described by the current collection. The main reason for these shortcomings lies in the setup of this project as data were collected well before all implications were known. Future sampling campaigns should therefore be more thoroughly planned with the appropriate database structure ,sampling protocol and recording of metadata.

5.2 Spectral Databases

The database developed for this project proved to be ideal for the data analysis that was carried out. It was, however, not designed to act as a repository for spectra that could be accessed by persons having no prior knowledge of the stored spectra. Therefore information such as the instrument used, illumination conditions, collector details and extensive target description was not included. Furthermore, the hierarchical structuring that features species, sites and spectra could be regarded as too restrictive. The experiences gained so far indicate that the chosen structure applies to most experiments. In some cases the site level might not be needed, but this inconvenience could be solved by a simple software modification leaving the database structure intact.

The database approach also enabled the data to be stored in a central place and the simultaneous data access by several users posed no problem because the database system ensured the data integrity. The implemented system however does not offer multi-user capability, i.e. users cannot store their own personalized settings.

Future spectral databases should provide multi-user access to studies and more information on the instrumentation and environmental conditions of the sampling. The direct linkage with a geographic information system (GIS) should also be considered when designing the database.

5.3 Spectral Processing Chain

The spectral processing chain consisted of waveband filtering, smoothing, data reduction (sensor synthesizing / downsampling), derivative calculation and feature space transformation. These are the most commonly used operations in hyperspectral studies and all pre-processing applied to the data in this study was achieved by these operations. It is clear though that the implemented steps are not conclusive. Other data processing such as continuum removal and special indices like band depth indices are in use in the research community. Such operations do not fit into the current chain. Furthermore one could argue about the logical order of the processing steps. E.g. the derivative calculation could be before or after the data reduction. For such a modification, a more flexible approach would be needed where the processing methods would be modularised allowing the interactive building of processing chains.

5.4 Processing Speed of Smoothing Operations

Of the two implemented algorithms for the application of Savitzky-Golay filters for smoothing of data, the moving window (MW) outperformed the FFT. Theoretically the FFT provides faster processing than conventional convolution above a certain number of N data points. Conventional convolution requires in the order of N^2 operations while FFT needs N*lg(N) operations where lg(N) is the logarithm-base-2 of N. FFT outperforms conventional convolution for around 64-128 data points (Smith, 2003). The reason why MW preformed faster than FFT lies in the processing overhead required for setting up the vectors, transforming them into frequency space and back into time space and storing the result again in the internal data structures. To avoid this, the spectral data should at all stages be stored in the Matrix objects supplied by the NewMat library.

5.5 Data Reduction

The issue of high data redundancy of hyperspectral data was addressed by data reduction techniques, either by the synthesizing of other sensor responses/downsampling or PCT. Analysis of original data and reduced data showed that the loss of vital information is minimal. E.g. the histograms highlighting the most discriminating bands were virtually identical for raw and Hyperion synthesized data. The sharp drop of the eigenvalues also confirmed that the data had a high redundancy. The first few components explained almost all variations found in the data.

Data reduction was also successful in the reduction of noise which greatly influenced the calculation of derivatives as was shown on the example of DGVIs.

5.6 Discriminative Power of Feature Spaces

In this study three different feature spaces were compared: DGVI, NTBI and PCT. PCT had the best discrimination of species, followed by DGVI and NTBI.

The DGVI was originally designed for correlation with plant properties. It is as such not optimized for the discrimination between species. A closer study of the DGVI regions (Figure 46) reveals that no narrow (~20nm) regions exist in the NIR (700-1300nm) and the SWIR segment 1 is also only partly covered by DGVI7. According to the result for the most discriminating bands carried out for the 1st derivative, SWIR1 had the highest frequency of statistically significant differences between bands, followed by NIR. One could expect a better discrimination if the DGVI regions were redefined, possibly featuring narrower regions for SWIR1 and SWIR 2 and new regions in the NIR.

The NTBI feature space had the lowest discriminative power. This however is very likely a direct result of the low dimensionality. Adding more dimensions should increase the discrimination. The selection of the best NTBI's could be achieved by a data-mining process where the Wilcoxon test would be applied to all possible two band combinations for all species pairs.

Both DGVI and NTBI feature spaces were found prone to have high correlations between dimensions. E.g. for DGVIs a discriminant analysis could not be carried out in Minitab because the correlations of certain variables were too high. Even PCT data had correlations between bands despite the fact that in theory PCT should be a zero correlation transform. A likely solution to this could be the building of a set of most discriminating but least correlated variables. This could be achieved by subjecting the variable set to a stepwise discriminant analysis.

5.7 Discrimination and Classification

Thirty two species were collected as training data set in this study and the highest classification accuracy (96.94%) was achieved using a generalized squared distance discriminant function on PCT data. This percentage of correctly classified spectra was astonishing, given that all plant spectra look very similar. However, one could expect that the classification accuracy would drop if more species were added to the data set. At some point an over crowding of feature space would take place resulting in overlaps of the species clusters. It may be necessary to segment the species by spatial properties or temporal information in order to limit the possible species that are used for classifications.

While the classification of the training data demonstrated the capability of discrimination of species by spectral data, the application of this technology relies on the result gained from the independent dataset. The independent set used in this study contained 15 species, i.e. less than half of the training set species. The maximum classification accuracy (87.87%) should therefore be regarded with caution. Ideally, all training set species should be included in the independent test set.

5.8 Principal Component Analysis

The variation explained by the first two components of Hyperion synthesized data was 97.9%. This was higher than the percentage of 85% reported by Thenkabail et al. (2004a). The high factor loadings in the SWIR mentioned by Thenkabail et al. were also found for New Zealand native plants. But the visible part of the spectrum had also high loadings, especially for PC3, which was different to the result of Thenkabail et al. who found the SWIR2 segment to have the highest loadings for PC3.

These findings indicate that the loading factors differ considerably with the training dataset and information about importance of bands for vegetation discrimination based on the analysis of PC factor loadings can not be readily generalized for all vegetation types.

5.9 Linear Transformations

PCT was used as a linear transformation in this study. Excellent results have found for both reduction of dimensionality and discrimination in the resulting feature space. The application of the Wilcoxon test however showed that the frequency of significant differences was not strictly tied to the components. One possible explanation could be that the variance explained by the components is partially to be attributed to the inherent noise. This noise would then decrease the frequency of statistically significant differences between species.

In the context of linear transformations like PCT the application of the MNF transformation to spectral data collections would be of interest. As MNF was designed to order the components by their signal to noise ratio, one could expect to find the frequency of significant differences tied to the component order when subjected to the Wilcoxon test. Traditionally, the MNF has been applied to imagery and the estimation of the noise covariance matrix has used the differences between neighbouring pixels (Green et al., 1988; Lee et al., 1990). The application to time series has been demonstrated by Hundley et al. (2001).

Whether the MNF transformation can be applied to hyperspectral signatures that have no spatial component and thus no spatial neighbours remains to be seen. The critical factor will remain the estimation of the noise covariance matrix.

5.10 Most Discriminating Bands

Interestingly the histogram of the most discriminating bands exhibited some differences to the data shown by Schmidt and Skidmore (2003). They reported that the most discriminating wavebands for saltmarsh vegetation occurred in the NIR and SWIR regions (740-1820 nm) of the spectra. The same analysis applied to New Zealand native plants indicated that NIR had the lowest overall frequencies of statistically significant differences between species pairs while SWIR segments 1 and 2 had the highest overall frequencies. This suggests that analyses of the most discriminating bands can again not be generalised but must be carried out for each differing set of spectral vegetation data.

The histogram calculated from 1st derivative data suggested that the NIR part of the spectrum contained important information. Analyses using 1st derivatives should therefore make use of the NIR region.

5.11 Separability Analysis and Discriminant Analysis

The separability analysis gave indications about the separability of species in certain feature spaces. PCT data had a mean JM distance of 2.0 which would indicate full separability. However, these results could not be used directly as a prediction for the accuracy that was achieved in the discriminant analysis. PCT data was not reaching 100% accuracy as could be expected. The reason for this lies in the different metrics. The discriminant functions used for the classification are not identical to the distance measurement of the JM or B distance.

5.12 Spectral Unmixing

While the unmixing of paper/plant and paper/plastic/plant mixtures worked really well, the abundance estimation for the mixtures of three different plants was unsatisfactory with root mean square errors between 26.88% and 68.79%. The reason for this is a phenomenon described by Price (1994): if an endmember can be described by the combination of two other endmembers the result of the unmixing is unlikely to yield useful results. Exactly this situation applied to the three endmembers of this experiment. The Kawakawa curve lay between the Lemonwood and Karaka curves and thus could have been a mixture of the latter two (see Figure 67).

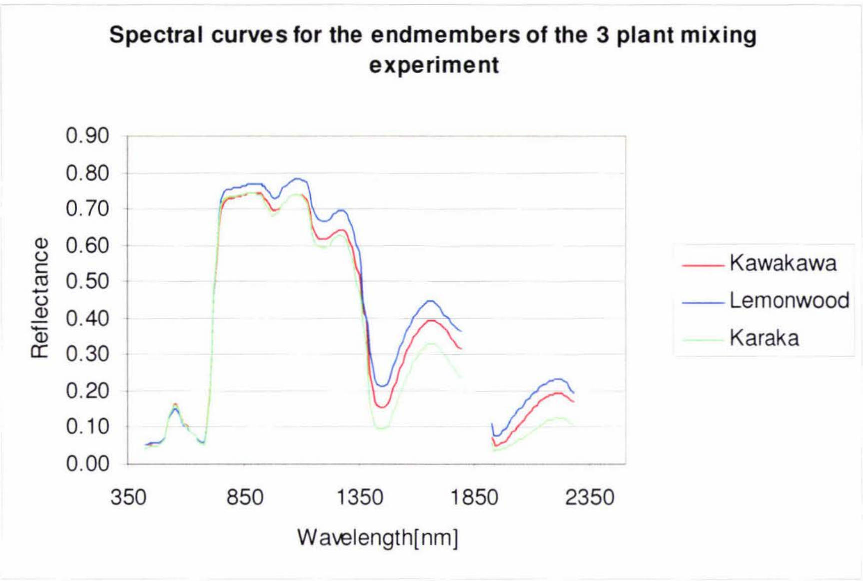


Figure 67: Spectral curves for the endmembers of the 3 plant mixing experiment

Another problem found was the positional dependence. The spectroradiometer sampled the field of view homogenously as was demonstrated by the probe rotation experiment. The positional dependence was therefore the result of the BRDF (Bidirectional Reflectance Distribution Function) with changing viewing geometry and fixed illumination geometry. The BRDF therefore biased the abundance estimations. A quantification of the BRDF influence would require further mixing experiments. One can however expect that the errors reported for the mixtures are at least partly the result of the BRDF influence.

5.13 Atmospheric Correction of Hyperion Imagery

In order to be able to compare the signature of Hyperion pixels with the collected ground data an atmospheric correction was applied using the FLAASH module in ENVI (Research Systems Inc., 2005).

A satisfying output could not be achieved despite using various settings for atmospheric and aerosol models and other parameters. A more thorough investigation into the matter would be needed. Such an effort was unfortunately beyond the timeframe of this research.

Recent findings point to the fact that FLAASH can produce good results if the scale file is edited in a certain way. One could also expect to improve the results if FLAASH were coupled with an empirical line correction.

6 Conclusion

This study has shown that the results of the analysis of hyperspectral data are heavily influenced by the preceding pre-processing. The main contributing factors were (a) the smoothing which depended on filter sizes and polynomial orders, (b) the data reduction achieved by synthesizing other sensor responses or downsampling and (c) the derivative calculation where a Savitzky-Golay filter was effecting a double smoothing. The best set of parameters for these operations was identified by testing different settings followed by statistical analysis.

It became clear that fast and repeatable data processing is a key factor to the efficient study of hyperspectral data. By storing all spectral data in a database, all subsequent operations could be carried out on the same dataset which remained unchanged. The implementation of software with a database interface that handled data input, processing and output proved to be a very effective way of hyperspectral data processing. The processing chain developed in this study contained methods that are most commonly used in hyperspectral studies. It is recommended that future processing chains should be of a modular nature to accommodate more varieties of data processing steps. Statistical research should be carried out in other software packages and only if a certain method has proven to be useful and often needed should it be implemented in the database interface software.

The atmospheric correction of Hyperion imagery was found to be difficult and no good match between ground data and pixel signatures could be achieved. The effort to improve these results was beyond the timeframe of this research.

The species of New Zealand native plants that were studied showed a very good potential for discrimination. More research is needed to gain knowledge of temporal and spatial variations. A possible outcome of such a study might be the collection of spectral reference data for certain seasons or regions.

7 Bibliography

Analytical Spectral Devices Inc. "Technical Guide." **2006** http://www.asdi.com/TG_Ref4_web.pdf.

Apan, A., Held, A., Phinn, S. & Markley, J. (2003). Formulation and assessment of narrow-band vegetation indices from EO-1 Hyperion imagery for discriminating sugarcane disease. Proceedings of the Spatial Sciences Conference, Canberra.

Bell, I. E. & Baranoski, G. V. G. (2004). "Reducing the Dimensionality of Plant Spectral Databases." IEEE Transactions on Geoscience and Remote Sensing **42**(3).

Ben-Dor, E. & Levin, N. (2000). "Determination of surface reflectance from raw hyperspectral data without simultaneous ground data measurements: a case study of the GER 63-channel sensor data acquired over Naan, Israel." International Journal of Remote Sensing **21**(10): 2053-2074.

Bojinski, S., Schaepman, M., Schlaepfer, D. & Itten, K. (2003). "SPECCHIO: a spectrum database for remote sensing applications." Computers & Geosciences **29**: 27-38.

Clark, M. L., Roberts, D. A. & Clark, D., B. (2005). "Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales." Remote Sensing of Environment **96**: 375-398.

Clark, R. N., Swayze, G. A., Gallagher, A. J., King, T. V. V. & Calvin, W. M. (1993). The U. S. Geological Survey, Digital Spectral Library: Version 1: 0.2 to 3.0 microns. U.S. Geological Survey Open File Report. **93:** 1340.

Cochrane, M. A. (2000). "Using vegetation reflectance variability for species level classification of hyperspectral data." International Journal of Remote Sensing **21**(10): 2075-2087.

Cocks, T., Jenssen, R., Stewart, A., Wilson, I. & Shields, T. (1998). The HYMAP airborne hyperspectral sensor: the system, calibration and performance. 1st EARSEL Workshop on Imaging Spectroscopy, Zurich.

Coops, N. C., Smith, M.-L., Martin, M. E. & Ollinger, S. V. (2003). "Prediction of Eucalypt Foliage Nitrogen Content From Satellite-Derived Hyperspectral Data." IEEE Transactions on Geoscience and Remote Sensing **14**(6): 1338-1346.

Davies, R. (2002). NewMat.http://www.robertnz.net.

Dymond, J. R. & Shepherd, J. D. (2004). "The spatial distribution of indigenous forest and its composition in the Wellington region, New Zealand, from ETM+ satellite imagery." Remote Sensing of Environment **90**: 116-125.

Elvidge, C. D. & Chen, Z. (1995). "Comparison of broadband and narrowband red and near-infrared vegetation indices." Remote Sensing of Environment **54**: 38-48.

Fliege, N. J. (1994). Multirate Digital Signal Processing. Chichester, John Wiley & Sons.

Fyfe, S. K. (2003). "Spatial and temporal variation in spectral reflectance: Are seagrass species spectrally distinct?" Limnol. Oceanography **48**(1): 464-479.

GER (2000). "GER EPS-H Series Airborne Imaging Spectrometer System." http://www.ger.com/epshman.html.

Green, A. A., Berman, M., Switzer, P. & Craig, M. D. (1988). "A Transformation for Ordering Multispectral Data in Terms of Image Quality with Implications for Noise Removal." IEEE Transactions on Geoscience and Remote Sensing **26**(1): 65-74.

Haboudane, D., Miller, J. R., Pattey, E., Zarco-Tejada, P. J. & Strachan, I. B. (2004). "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture." Remote Sensing of Environment **90**: 337-352.

Herold, M., Roberts, D. A., Gardner, M. E. & E., D. P. (2004). "Spectrometery for urban remote sensing - Developement and analysis of a spectral library from 350 to 2400 nm." Remote Sensing of Environment **91**: 304-319.

Huang, Z., Turner, B., Dury, S., Wallis, I. & Foley, W. (2004). "Estimating foliage nitrogen concentration from HYMAP data using continuum removal analysis." Remote Sensing of Environment **93**(1-2): 18-29.
The

Hundley, D., Anderle, M. & Kirby, M. (2001). A Solution Procedure for Blind Signal Separation Using the Maximum Noise Fraction Approach: Algorithms and Examples. Proceedings of the Conference on Independent Components Analysis, San Diego, CA.

Hutsinpiller, A. (1988). "Discrimination of Hydrothermal Alteration Mineral Assemblages at Virginia City, Nevada, Using the Airborne Imaging Spectrometer." Remote Sensing of Environment **24**: 53-66.

Integrated Spectronics Pty Ltd "HyMap Airborne Scanners." **2006** http://www.intspec.com/Products/HyMapProd.htm.

Keshava, N. & Mustard, J. F. (2002). "Spectral Unmixing." IEEE Signal Processing Magazine **19**(1): 44-57.

Kokaly, R. F. & 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**: 267–287.

Kokaly, R. F., Despain, D. G., Clark, R. N. & Livo, E. K. (2003). "Mapping vegetation in Yellowstone National Park using spectral feature analysis of AVIRIS data." Remote Sensing of Environment **84**: 437-456.

Kruse, F. A. (2004). "Comparison of ATREM, ACORN and FLAASH Atmosperic corrections using low-altitude AVIRIS data of Boulder, CO." www.hgimaging.com/PDF/Kruse-JPL2004_ATM_Compare.pdf.

Labsphere Inc. North Sutton, NH, USA.

Landgrebe, D. (1997). On Information Extraction Principles for Hyperspectral Data, Purdue University.

Landgrebe, D. (2003). Signal Theory Methods in Multispectral Remote Sensing. Hoboken, New Jersey, John Wiley & Sons.

Lee, J. B., Woodyatt, A. S. & Berman, M. (1990). "Enhancement of High Spectral Resolution Remote-Sensing Data by a Noise-Adjusted Principal Components Transform." IEEE Transactions on Geoscience and Remote Sensing **28**(3): 295-304.

Liao, L. & Jarecke, P. (undated). "Performance Characterization of the Hyperion Imaging Spectrometer Instrument." www.eoc.csiro.au/hswww/oz_pi/docs/hyperion_performance.pdf.

Lillesand, T. M., Kiefer, R. W. & Chipman, J. W. (2004). Remote Sensing and Image Interpretation, John Wiley & Sons.

Lusch, D. P. (1989). Fundamental Considerations for Teaching the Spectral Reflectance Characteristics of Vegetation, Soil and Water. Current Trends in Remote Sensing Education. D. M. Nelliset al. Hong Kong, Geocarta International Centre.

Martin, M. E. & Aber, J. D. (1997). "High spectral resolution remote sensing of forest canopy lignin, nitrogen, and ecosystem processes." Ecological Applications **7**(2): 431-443.

Mathur, A., Mann Bruce, L. & Byrd, J. (2002). "Discrimination of Subtly Different Vegetative Species via Hyperspectral Data." IEEE International Geoscience and Remote Sensing Symposium **2**: 805-807.

Mazer, A. S., Martin, M., Lee, M. & Solomon, J. E. (1988). "Image Processing Software for Imaging Spectrometry Data Analysis." Remote Sensing of Environment **24**: 201-210.

Miller, J. N. & Miller, J. C. (2005). Statistics and Chemometrics for Analytical Chemistry. London, Pearson Education Limited.

Milton, E. J. (2001). "Methods in Field Spectroscopy." www.soton.ac.uk/~epfs/methods/spectroscopy.shtml.

Minitab Inc. (2003). MINITAB Statistical Software. State College, Pennsylvania.

Mundt, J. T., Glenn, N. F., Weber, K. T., Prather, T. S., Lass, L. W. & Pettingill, J. (2005). "Discrimination of hoary cress and determination of its detection limits via hyperspectral image processing and accuracy assessment techniques." Remote Sensing of Environment **96**: 509-517.

MySQL AB (2005). MySQL.http://www.mysql.com.

Olsen, R. C., Bergman, S. & Resmini, R. G. (1997). Target detection in a forest environment using spectral imagery. SPIE 1997 Annual International Symposium on Optical Science, Engineering and Instrumentation, San Diego.

Papula, L. (1994). Mathematik fuer Ingenieure und Naturwissenschaftler, Viewegs.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2002). Numerical Recipies in C++. Cambridge, Cambridge University Press.

Price, J. C. (1994). "How Unique Are Spectral Signatures?" Remote Sensing of Environment **49**: 181-186.

Ramsey, E. & Nelson, G. (2005). "A whole image approach using field measurements for transforming EO1 Hyperion hyperspectral data into canopy reflectance spectra." International Journal of Remote Sensing **26**(8): 1589-1610.

Ramsey, E., Rangoonwala, A., Nelson, G., Ehrlich, R. & Martella, K. (2005). "Generation and validation of characteristic spectra from EO1 Hyperion image data for detecting the occurrence of the invasive species, Chinese tallow." International Journal of Remote Sensing **26**(8): 1611-1636.

Research Systems Inc. (2004). ENVI Tutorials. Boulder, CO.

Research Systems Inc. (2005). ENVI. Boulder, CO.

Richards, J. A. (1993). Remote Sensing Digital Image Analysis. Berlin, Springer Verlag.

Richardson, A. D., Reeves, J. B. & Gregoire, T. G. (2003). "Multivariate analyses of visible/near infrared (VIS/NIR) absorbance spectra reveal underlying spectral differences among dried, ground conifer needle samples from different growth environments." New Phytologist **161**: 291-301.

Riedmann, M. (2003). "Laboratory Calibration of the Compact Airborne Spectrographic Imager (CASI-2)." http://www.ncaveo.ac.uk/site-resources/pdf/MRCasiCal.pdf.

Roberts, D. A., Gardner, M. E., Church, R., Ustin, S., Scheer, G. & Green, R. O. (1998). "Mapping Chaparral in the Santa Monica Mountains Using Multiple Endmember Spectral Mixture Models." Remote Sensing of Environment **65**: 267-279.

Savitzky, A. & Golay, M. J. E. (1964). "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." Analytical Chemistry **36**(8): 1627-1639.

Schmidt, K. S. & Skidmore, A. K. (2003). "Spectral discrimination of vegetation types in a coastal wetland." Remote Sensing of Environment **85**: 92-108.

Schmidt, K. S. & Skidmore, A. K. (2004). "Smoothing vegetation spectra with wavelets." International Journal of Remote Sensing **25**(6): 1167-1184.

Shannon, C. E. (1949). "Communication in the presence of noise." Proc. IRE **37**: 10-21.

Shaw, G. & Manolakis, D. (2002). "Signal Processing for Hyperspectral Image Exploitation." IEEE Signal Processing Magazine **19**(1): 12-16.

Shepherd, K. D. & Walsh, M. G. (2002). "Development of Reflectance Spectral Libraries for Characterization of Soil Properties." Soil Science Society Am. J. **66**: 988-998.

Smith, G. M. & Milton, E. J. (1999). "The use of the empirical line method to calibrate remotely sensed data to reflectance." Int. J. Remote Sensing **20**(13): 2653-1662.

Smith, J. O. (2003). Mathematics of the Discrete Fourier Transform (DFT), with Music and Audio Applications, W3K Publishing.

The MathWorks Inc. (2004). Matlab. Natick, MA.

Thenkabail, P. S., Enclona, E. A. & Ashton, M. S. (2004a). "Accuracy assessment of hyperspectral waveband performance for vegetation analysis applications." Remote Sensing of Environment **91**: 354-376.

Thenkabail, P. S., Enclona, E. A., Ashton, M. S., Legg, C. & Minko, J. D. D. (2004b). "Hyperion, IKONOS, ALI and ETM+ Sensors in the study of African rainforests." Remote Sensing of Environment **90**: 23-43.

Thenkabail, P. S., Smith, R. B. & De Pauw, E. (2000). "Hyperspectral Vegetation Indices and Their Relationship with Agricultural Crop Characteristics." Remote Sensing of Environment **71**: 158-182.

Thenkabail, P. S., Smith, R. B. & De Pauw, E. (2002). "Evaluation of Narrowband and Broadband Vegetation Indices for Determining Optimal Hyperspectral Wavebands for Agricultural Crop Characterization." Photogrammetric Engineering & Remote Sensing **68**(6): 607-621.

Tsai, F. & Philpot, W. (1998). "Derivative Analysis of Hyperspectral Data." Remote Sensing of Environment **66**: 41-51.

University of Waikato (2005). WEKA.http://www.cs.waikato.ac.nz/~ml/weka/.

Unser, M. (2000). "Sampling—50 Years After Shannon." Proceedings of the IEEE **88**(4): 569-587.

USGS (2005). "EO-1 User's Guide: Data Properties: Hyperion." http://eo1.usgs.gov/userGuide/hyp_prop.html.

van Till, M., Bijlmer, A. & de Lange, R. (2004). "Seasonal Variability in Spectral Reflectance of Coastal Dune Vegetation." EARSel eProceedings **3**(2): 154-165.

Vane, G. & Goetz, A. F. H. (1988). "Terrestrial Imaging Spectroscopy." Remote Sensing of Environment **24**: 1-29.

Venables, W. N., Smith, D. M. & and the R Development Core Team (2005). R: A Programming Environment for Data Analysis and Graphics.

Williams, D. & Summers, R. (2004). "Database Design." http://gisweb.massey.ac.nz/Topic/DatabaseDesign/lectures/introduction.html.

Williams, D. J., Rybicki, N. B., Lombana, A. V., O'Brien, T. M. & Gomez, R. B. (2002). "Preliminary Investigation of Submerged Aquatic Vegetation Mapping Using Hyperspectral Remote Sensing." Environmental Monitoring and Assessment **81**(1-3): 383 - 392.

Younan, N. H., King, R. L. & Bennett, H. H. J. (2004). "Classification of Hyperspectral Data: A Comparative Study." Precision Agriculture **5**: 41-53.

Zanoni, V., Davis, B., Ryan, R., Gasser, G. & Blonski, S. (2002). "Remote Sensing Requirements Development: A Simulation-Based Approach." ISPRS.www.isprs.org/commission1/proceedings/paper/VZanoni_ISPRS2002.pdf.

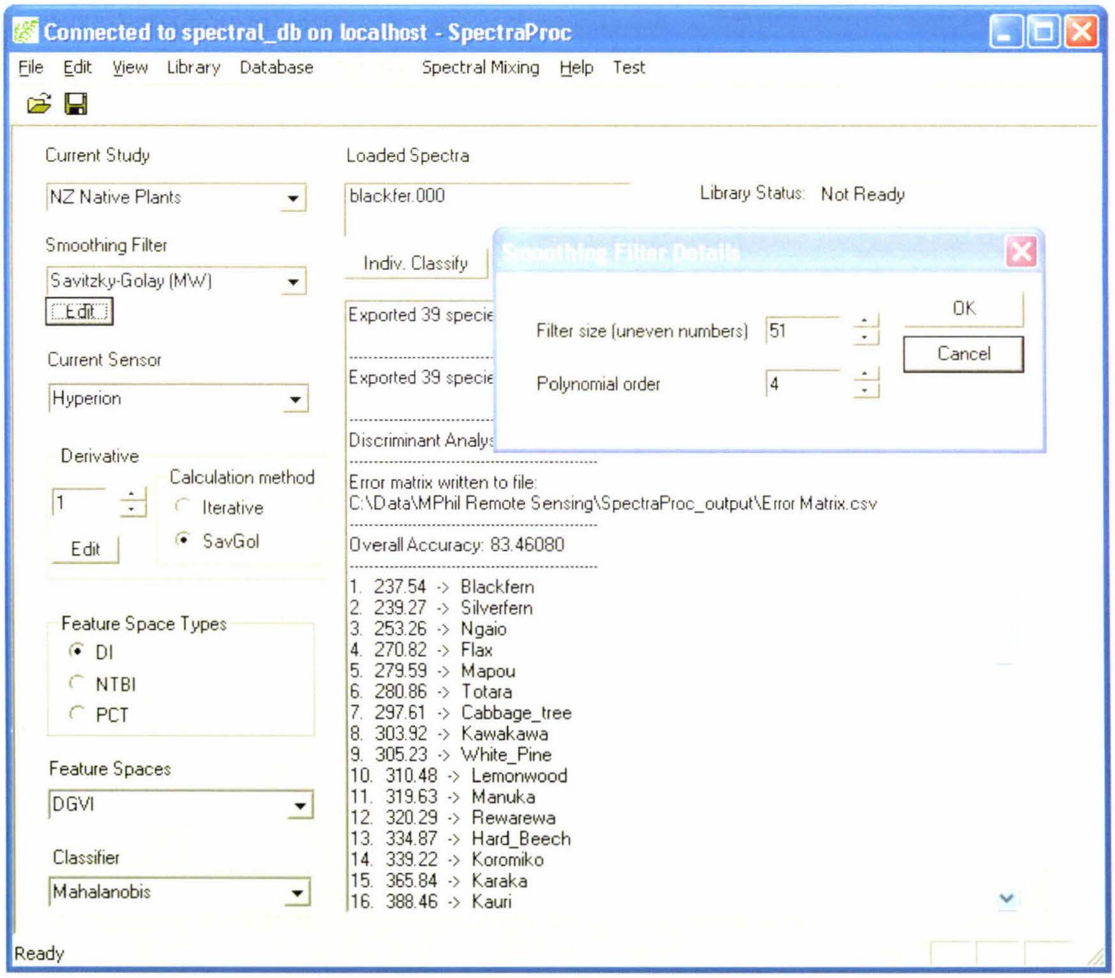
8 Appendix

8.1 SpectraProc Graphical User Interface

The graphical user interface (GUI) (for a screenshot please see Figure 68) was based on the structure of the processing chain (see Figure 19). The left side of the main window consists of controls for the selection of the study and the main settings for smoothing filter, synthesizing, derivative calculation, feature space transformation and classifier discriminant function. Processing details are entered in pop up windows, shown here with the example of the smoothing function. The text output panel in the middle of the main window is used to display processing and error information.

The listbox on top of the text output panel is used to display spectra files that are loaded directly into memory. The 'Indiv. Classify' button under it classifies the selected, individually loaded spectra against the current library.

The library status box on the top right of the screen indicates whether statistical information has been compiled for the current pre-processing settings.



Current Study: NZ Native Plants

Smoothing Filter: Savitzky-Golay (MW)

Current Sensor: Hyperion

Derivative: 1

Calculation method: ○ Iterative ● SavGol

Feature Space Types: ● DI ○ NTBI ○ PCT

Feature Spaces: DGVI

Classifier: Mahalanobis

Loaded Spectra: blackfer.000

Library Status: Not Ready

Filter size (uneven numbers): 51

Polynomial order: 4

```
Exported 39 specie
-----------------------------------
Exported 39 specie
-----------------------------------
Discriminant Analy:
-----------------------------------
Error matrix written to file:
C:\Data\MPhil Remote Sensing\SpectraProc_output\Error Matrix.csv
-----------------------------------
Overall Accuracy: 83.46080
-----------------------------------
1. 237.54 -> Blackfern
2. 239.27 -> Silverfern
3. 253.26 -> Ngaio
4. 270.82 -> Flax
5. 279.59 -> Mapou
6. 280.86 -> Totara
7. 297.61 -> Cabbage_tree
8. 303.92 -> Kawakawa
9. 305.23 -> White_Pine
10. 310.48 -> Lemonwood
11. 319.63 -> Manuka
12. 320.29 -> Rewarewa
13. 334.87 -> Hard_Beech
14. 339.22 -> Koromiko
15. 365.84 -> Karaka
16. 388.46 -> Kauri
```

Ready

Figure 68: Screen capture of SpectraProc