ATMOSPHERIC EFFECTS ON LAND CLASSIFICATION USING SATELLITES AND THEIR CORRECTION

A thesis submitted in partial fulfilment
of the requirements of the degree of
Doctor of Philosophy

by

Asmala Ahmad

School of Mathematics and Statistics
University of Sheffield

June 2012

ATMOSPHERIC EFFECTS ON LAND CLASSIFICATION USING SATELLITES AND THEIR CORRECTION

Summary

Haze occurs almost every year in Malaysia and is caused by smoke which originates from forest fire in Indonesia. It causes visibility to drop, therefore affecting the data acquired for this area using optical sensor such as that onboard Landsat – the remote sensing satellite that have provided the longest continuous record of Earth's surface. The work presented in this thesis is meant to develop a better understanding of atmospheric effects on land classification using satellite data and method of removing them. To do so, the two main atmospheric effects dealt with here are cloud and haze. Detection of cloud and its shadow are carried out using MODIS algorithms due to allowing optimal use of its rich bands. The analysis is applied to Landsat data, in which shows a high agreement with other methods. The thesis then concerns on determining the most suitable classification scheme to be used. Maximum Likelihood (ML) is found to be a preferable classification scheme due to its simplicity, objectivity and ability to classify land covers with acceptable accuracy. The effects of haze are subsequently modelled and simulated as a summation of a weighted signal component and a weighted pure haze component. By doing so, the spectral and statistical properties of the land classes can be systematically investigated, in which showing that haze modifies the class spectral signatures, consequently causing the classification accuracy to decline. Based on the haze model, a method of removing haze from satellite data was developed and tested using both simulated and real datasets. The results show that the removal method is able clean up haze and improve classification accuracy, yet a highly non-uniform haze may hamper its performance.

Acknowledgement

I must thank my supervisor, Professor Shaun Quegan for his invaluable advice and guidance towards the completion of this thesis.

I am thankful to my colleagues in the Centre for Terrestrial Carbon Dynamics (CTCD) for the useful ideas and comments.

I would like to express my deepest appreciation to my sponsor, the Government of Malaysia for funding my PhD study. I am most grateful to the Malaysian Remote Sensing Agency, Malaysian Meteorological Department and Department of Environment, Malaysia for providing the data.

Finally, I am indebted to my parents, family and friends for their continuous supports and motivations throughout my study.

Table of Contents

Chapter 1

Introduction

1.1 Introduction

One of the most important types of information needed by regional and national governments concerns the condition and use of land within its territory, and how these are changing. This is particularly true for developing countries which are experiencing urban sprawl, deforestation, increase in impervious surfaces and other major modifications to the land surface. If land use is not monitored and managed properly, it may have effects on regional issues such as land degradation, loss of tropical rain forest, desertification or food security, as well as global issues such as climate change and loss of biodiversity. In countries such as Malaysia, land use change has caused undesirable impacts such as landslides, floods, loss of forest, loss of wetlands, loss of agricultural land and unplanned urbanisation.

We need to distinguish land use and land cover; *land use* refers to human exploitation of the land, e.g. for agriculture or forestry. This is related to, but is not the same as, *land cover*, which describes vegetation and artificial constructions covering the land surface (Anderson 1976). Satellite remote sensing instruments normally measure land cover, from which it may be possible to infer land use, and have been an excellent basis from which to observe large-scale landscapes systematically, consistently and synoptically (Wickland 1991). In this thesis we will be concerned with land cover mapping using remotely sensed data.

Land use and land cover information is vital for a wide variety of decision-making, planning and managing activities, as well as formulating measures to combat existing problems at global, regional and national levels. Such information must be gathered and stored systematically so that it can be retrieved without difficulty by users, ranging from students, technical workers, researchers, engineers and managers to policy makers.

Conventionally, information about land use and land cover was obtained from ground surveys, which require a huge amount of time and logistics, and are therefore very expensive; yet they have served many users for a long time. Later, aerial photographic surveys were implemented, which made land use and land cover mapping much easier but are logistically very expensive. Recent advances in remote sensing technology offer a much more practical way of mapping land use and land cover over large areas at an affordable cost.

For global needs, a number of land cover maps have been produced. One of the earliest was the University of Maryland Land Cover produced using the NOAA AVHRR satellite. Initially, in 1984, maps with 8 km resolution were produced, but later, in 1992, maps with 1 km resolution were produced (DeFries and Townshend 1994; DeFries et al, 1998; Loveland et al. 2000; Hansen et al. 2003). Also in 1992, researchers from the U.S. Geological Survey, University of Nebraska–Lincoln and the Joint Research Centre of the European Commission used NOAA AVHRR data to produce a 1 km resolution global land cover database known as DISCover (Loveland et al. 2000). Later, in 2000, the Joint Research Centre developed Global Land Cover 2000, popularly known as GLC2000, with 1 km resolution, using SPOT Vegetation data (Bartholome and Belward 2005). In the same year, the MODIS Vegetation Continuous Fields product, which contains information about vegetative cover types (i.e. woody vegetation, herbaceous vegetation, and bare ground), with 500 m resolution was produced by the University of Maryland using MODIS data (Hansen et al. 2003). In 2010, the European Space Agency and the Joint Research Centre produced GLOBCOVER with 300 m resolution using ENVISAT MERIS data (Bicheron et al. 2011). For GLOBCOVER and GLC2000, the land cover classification is based on the Food and Agricultural Organisation (FAO) Land Cover Classification System (LCCS), which assures its worldwide applicability and compatibility with other land cover mapping projects.

Although of benefit to many users, these global land cover maps are at a coarse resolution (i.e. 300 to 1000 m) and do not fulfil many of the needs at regional levels. Consequently, several regional land cover mapping projects were initiated in Europe, Africa and Asia. Developed regions, such as Europe, began such efforts much earlier. The Image and CORINE Land Cover 2000 (I & CLC 2000) project, initiated in 2000 by the European Environment Agency, was an extended version of the CLC project which started in the mid-1980s. With 1:200,000 scale and making use of Landsat and SPOT data as the primary input, the objectives of I & CLC 2000 were to (a) provide a satellite image snapshot of Europe in 2000, (b) update the CORINE land cover map and (c) produce land cover change maps for the period 1990-2000. For less developed regions, FAO has facilitated a number of land cover mapping projects, such as Africover, initiated in 1994 for Africa, and Asiacover, initiated in 1999 for Asia. Africover and Asiacover are based on FAO LCCS and used Landsat TM and ETM+ and ALOS-AVNIR data with mapping scales 1:100,000 to 1:200,000. Africover's East African module, covering ten countries (Burundi, Democratic Republic of Congo, Egypt, Eritrea, Kenya, Rwanda, Somalia, Sudan, Tanzania and Uganda), was completed in 2004, while the preparatory phase of Asiacover, which involved Cambodia, China (Province of Yunnan), Lao People's Democratic Republic, Malaysia, Myanmar, Thailand and Viet Nam, was completed in 2005.

Nevertheless, these regional maps were still at a quite coarse scale and therefore were less useful at national level. Consequently, national land cover mapping projects were initiated by countries such as the United States of America and the United Kingdom that possess up-to-date technologies, facilities and expertise. In the USA, the National Land Cover Data (NLCD) with 300 m resolution was started in the 1990s by the Multi-Resolution Land Characteristics Consortium, and its latest version, NLCD2001, with 30 m resolution, was completed in 2001. It used Landsat TM and ETM+ data. In the UK, the Land Cover Map 2000 (LCM2000) was produced by the Centre for Ecology and Hydrology in 2000 and was an upgraded version of the LCM Great Britain developed in 1990 (Fuller et al. 2000; Fuller 2005). The LCM2000 covers the whole Great Britain, i.e. England, Scotland, Wales and Northern Ireland with a spatial resolution of 25 m x 25 m and used a hierarchical classification scheme. It has been used for environmental impact assessments, checking agricultural censuses, metropolitan and landscape planning, catchment and groundwater management, flood risk assessment, telecommunications, health and hazard assessments, predicting climate change impacts, carbon accounting, conservation work, site assessments, and environmental and ecological research.

Malaysia was also determined to have her own national land cover maps. Since 1966, land cover maps were produced using aerial photographs by the Malaysian Department of Agriculture (DOA) (Mahmood et al. 1997). The use of remote sensing technology was initiated by the Malaysian goverment in 1988 with the establishment of the Agensi Remote Sensing Malaysia (ARSM), formerly known as the Malaysian Centre for Remote Sensing (MACRES), under the government's Ministry of Science, Technology and Innovation. The main objectives of ARSM are to develop remote sensing and related technologies and to operationalise their applications in user agencies for management of natural resources, environment and disasters, and strategic planning of the nation (ARSM 2011). Beginning the same year, as a joint effort between DOA and ARSM, land cover maps with 1:50,000 scale have been produced using Landsat and SPOT data. Initially, satellite data were purchased from neighbouring countries, such as Singapore and Thailand which have their own ground receiving stations. Since then, there has been a growing interest in the use of remote sensing and the amount of remote sensing projects and research has increased; this persuaded the Malaysian government to allocate more budget for the development of remote sensing and space related technologies (ARSM 2005). Eventually, in 2002, the Malaysian Ground Receiving Station (MGRS) was developed, which is capable of acquiring optical (i.e. Landsat, SPOT, MODIS and NOAA) and microwave (i.e. Radarsat-1) data (ARSM 2005). Major national projects coordinated by ARSM include National Resource and Environmental Management (NaREM) and Precision Farming.

Overview of Remote Sensing Activities in Malaysia

NaREM, the first national project of its kind, was initiated in 1999, with the aim of developing an operational natural resource and environmental management system using remote sensing and its related technologies to meet national development planning. NaREM encompasses three major components (ARSM 2011):

- NaSAT, which is an integrated database using remote sensing as the main source of data input, enhanced by baseline data on topography, agriculture, forestry, geology, coastal zone, environment, socio-economic and natural disaster.

- NaMOS, which provides models and algorithms for various applications, e.g. landslide hazard monitoring, coastal sensitivity index, soil erosion, ground water potential, agro-suitability zoning and total forest management,
- NaDES, which is an integrated development planning decision support system that incorporates resource, environmental, economic, socio-economic and policy information.

The concept of NaREM is illustrated in Figure 1.1.

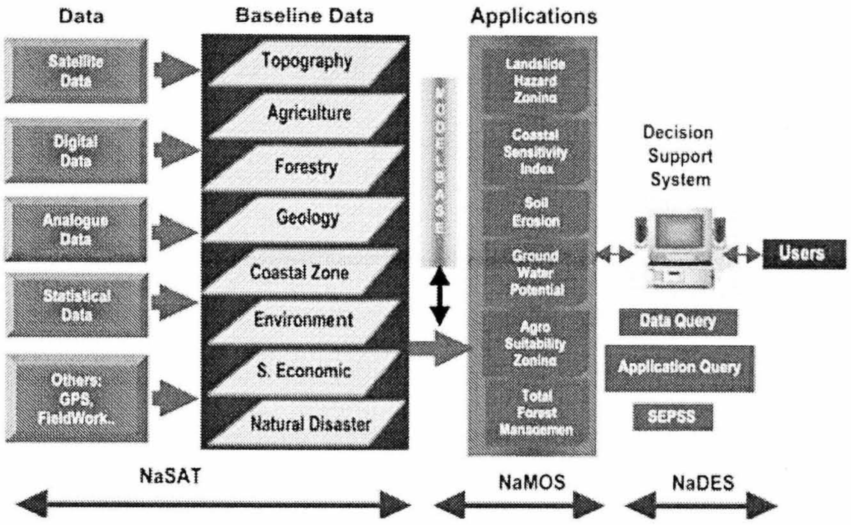


Figure 1.1: *NaREM major components (ARSM 2011).*

The main input to NaSAT is satellite data, provided by ARSM itself. Landsat data are the main source of satellite data for NaSAT and are obtained from MGRS. The operation of NaREM requires ancillary data from other government agencies, i.e. the Survey and Mapping Department, Department of Environment, Department of Statistics, Department of Fisheries, Department of Agriculture, Department of Forestry, Department of Mineral and Geoscience, Department of Irrigation and Drainage and Department of Meteorology, and experts in various fields, mainly from universities and industries (Figure 1.2). The outputs of NaREM are used by the Malaysian Economic Planning Unit, i.e. the agency responsible for economic policies for Malaysia (ARSM 2011).

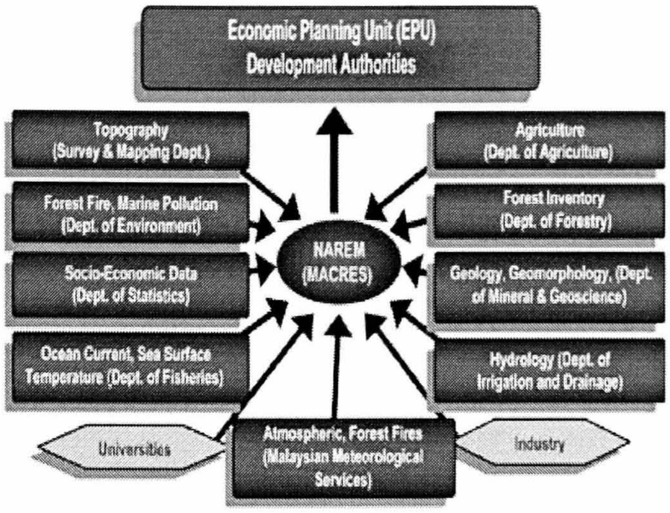

Economic Planning Unit (EPU)
Development Authorities
Topography
(Survey & Mapping Dept.)
Agriculture
(Dept. of Agriculture)
Forest Fire, Marine Pollution
(Dept. of Environment)
Forest Inventory
(Dept. of Forestry)
NAREM
(MACRES)
Socio-Economic Data
(Dept. of Statistics)
Geology, Geomorphology, (Dept. of Mineral & Geoscience)
Ocean Current, Sea Surface Temperature (Dept. of Fisheries)
Hydrology (Dept. of Irrigation and Drainage)
Universities
Atmospheric, Forest Fires
(Malaysian Meteorological Services)
Industry

Figure 1.2: *NaREM input and output components (ARSM 2011).*

Another important project coordinated by ARSM is in precision farming, which was also initiated in 1999. Its main objective is to enhance crop production through the integration of remote sensing, GIS and GPS into farming practices, whilst at the same time preserving the quality of the environment. The system emphasizes that agricultural input, such as fertilizers, pesticides and water, should be used at the right amount, time and place (ARSM 2011). The Precision Farming Concept is illustrated in Figure 1.3.

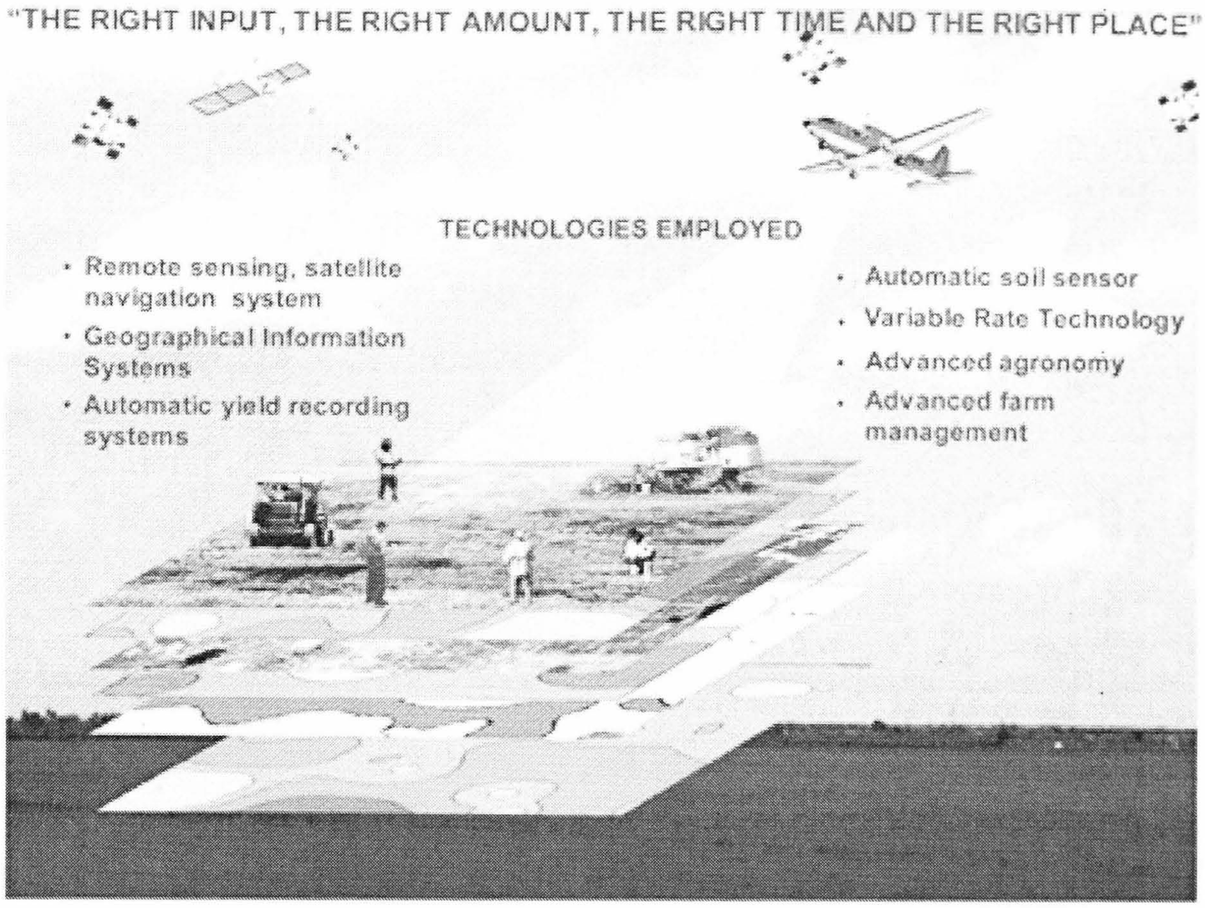

"THE RIGHT INPUT, THE RIGHT AMOUNT, THE RIGHT TIME AND THE RIGHT PLACE"
TECHNOLOGIES EMPLOYED
• Remote sensing, satellite navigation system
• Geographical Information Systems
• Automatic yield recording systems
• Automatic soil sensor
• Variable Rate Technology
• Advanced agronomy
• Advanced farm management

Figure 1.3: *Concept of precision farming (ARSM 2011)*

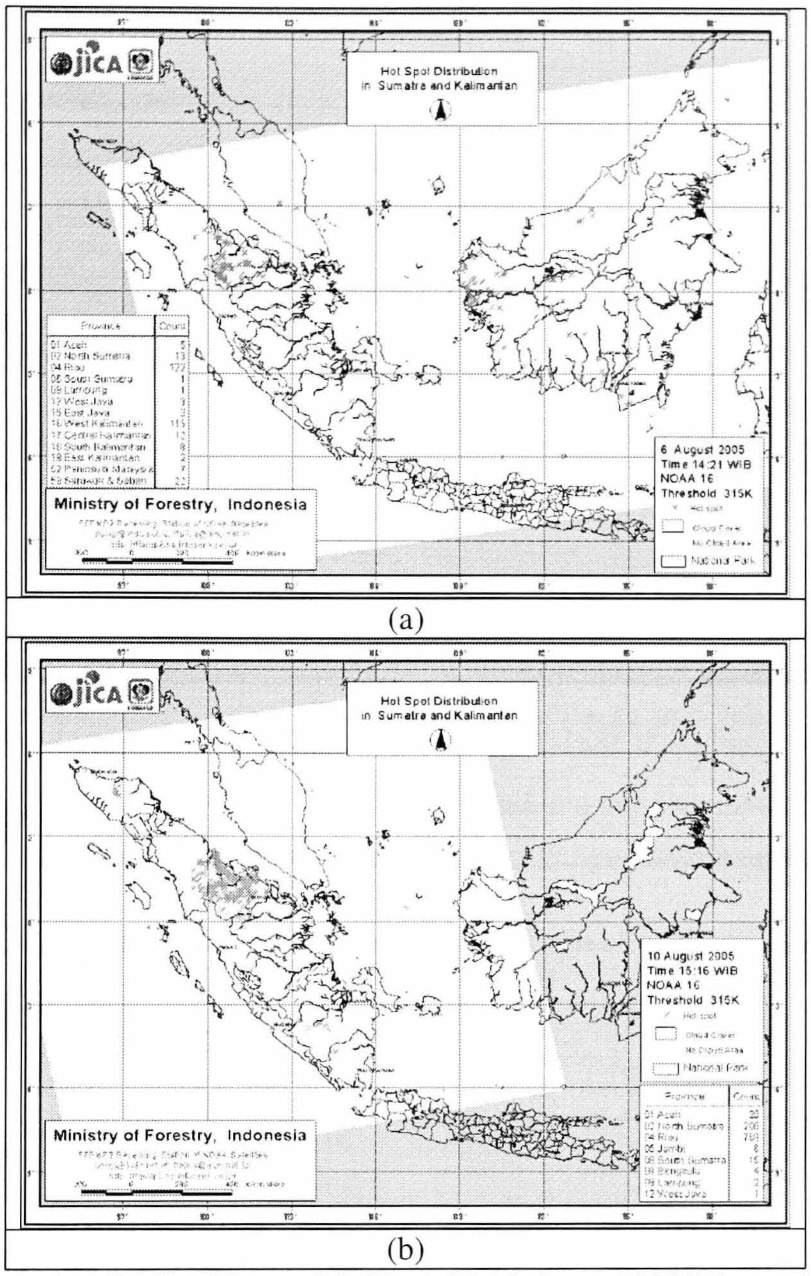
At present, precision farming is being implemented for two major Malaysian crops, rice and oil palm. Rice is a Malaysian staple food; approximately 70% of Malaysian rice consumption comes from national production and 30% is imported from Thailand. In precision farming of rice, due to its short life cycle, the incorporation of remote sensing sensors is very useful for providing data in a timely and cost-effective manner. Other national projects coordinated by ARSM include Integrated Geospatial Database and Planning System, Disaster Management, Fishing Zone Identification, Rice Monitoring and Yield Prediction System, Monitoring of Environmentally Sensitive Areas, Microwave Remote Sensing Research and Development, Integrated Remote Sensing and GIS Software Development and Satellite Image Map (ARSM 2011).

There are also remote sensing projects carried out by other government agencies, which use remote sensing as a tool to facilitate their routine tasks (ARSM 2005). These include the Department of Agriculture, Department of Mapping and Surveying, Department of Geology and Geoscience, Department of Fishery and Department of Meteorology. Research institutes that incorporate remote sensing technology in their work include the Malaysian Agriculture Research and Development Institute, the Rubber Research Institute and the Malaysian Palm Oil Board. In addition, due to job market demands, a number of universities have initiated remote sensing courses at postgraduate and undergraduate levels; e.g. Universiti Teknologi Malaysia, Universiti Sains Malaysia, Malaysian Multimedia University and Universiti Kebangsaan Malaysia (Hashim et al. 2004).

With the establishment of MGRS, Malaysia is now able to continuously acquire her own remote sensing data, without relying on other countries. As a huge amount of budget has been spent to establish the remote sensing facilities and more needs to be spent for maintaining them, the Malaysian government is looking forward to boosting remote sensing activities in Malaysia, so that as much benefit as possible will be gained by the country in return.

Haze Effects on Remote Sensing and Land Cover

Unfortunately, remote sensing Malaysian remote sensing data are affected by haze, which is a partially opaque condition of the atmosphere caused by tiny suspended solid or liquid particles in the lower atmosphere (Morris 1975). The thick haze that occurs in Malaysia is caused mainly by smoke originating from large forest fires in Indonesia, due to agricultural clean-up activities as farmers and large companies convert forests into plantations using fire to clear land (Hashim et al. 2004; Mahmud 2009). Major forest fires occurred in 1982-83, 1987, 1991, 1994, 1997-98, 2002, 2004 and 2005. For 2005, forest fire distributions in Indonesia from 6 and 10 August are shown in Figure 1.4.

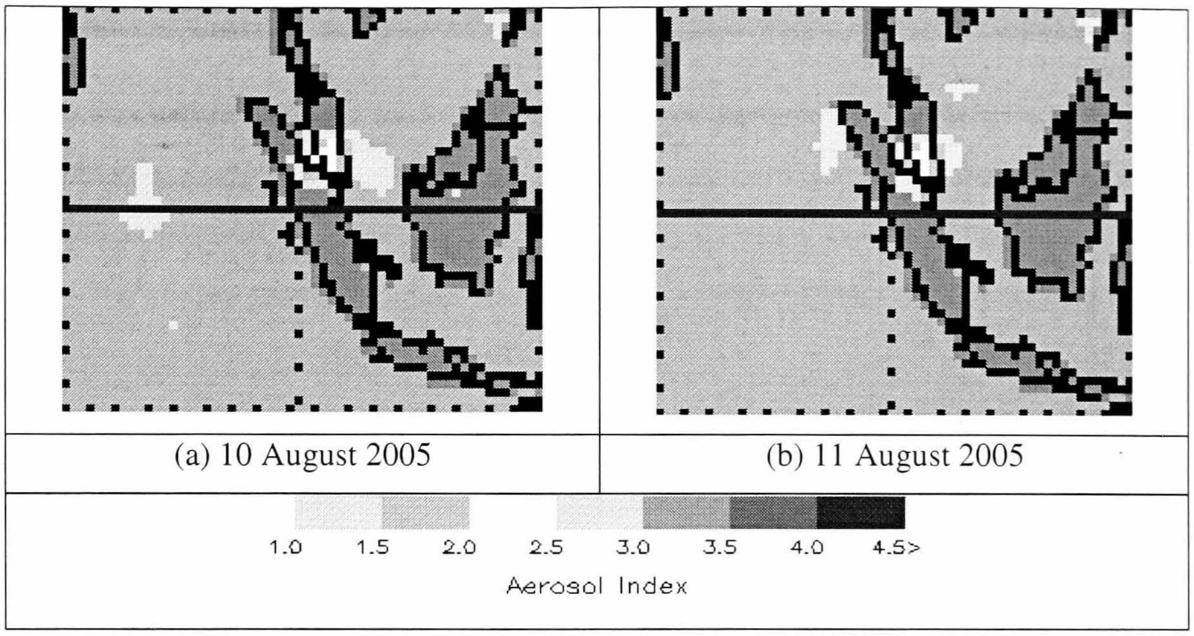


(a)

(b)

Figure 1.4: Fire distributions on (a) 6 and (b) 10 August 2005 determined from NOAA 16 satellite (Ministry of Forestry Indonesia 2010).

These forest fires release a huge amount of smoke that contains particulates and gases into the atmosphere (Mahmud 2009). The smoke is carried by the South-west Monsoon wind to neighbouring countries, such as Malaysia, Singapore, Thailand and Brunei (Mahmud 2009), and can travel hundreds of kilometres across the Southeast Asian region, reaching the Philippines. Haze conditions over Malaysia and Indonesia, based on the aerosol index measured using the Total Ozone Mapping Spectrometer (TOMS) from 10 and 11 August 2005 are shown in Figure 1.5. These are extreme examples, but lower level haze is a common occurrence, as seen in Figure 1.6.


(a) 10 August 2005

(b) 11 August 2005

Figure 1.5: *The aerosol index measured by the Total Ozone Mapping Spectrometer on (a) 10 and (b) 11 August 2005. The horizontal solid line and the vertical dashed line in the middle of the image represent latitude 0^o and longitude 100^o east respectively.*

Haze occurrences have also been reported in Africa and South America. In South America, plumes and haze layers originate from biomass burning that occurs every year over the central Amazon Basin due mainly to deforestation and land conversion (Guild et al. 2004). The haze layers occur at altitudes between 1000 and 4000 m and are 100 to 300 m thick but extend horizontally over several hundreds kilometres. The emissions from the burning significantly affect the chemical and optical characteristics of the atmosphere over the Amazon Basin (Andreae et al. 1988).

In West Africa, during the dry season, biomass burning occurs particularly in the Sahelian regions, due to the burning of agricultural waste (Haywood et al. 2008). Emissions from these fires were reported to reach as far as South Africa. The Southern African Regional Science Initiative (Justice et al. 1996) studied the generation, transport and deposition of the associated aerosols to develop better understanding of related environmental processes, such as the effect of aerosols on the global radiation balance.

Visibility will be used as the key parameter to describe haze severity. It is defined as the greatest distance at which a black object located on the ground can be seen and recognized when observed against the horizon sky during daylight or could be seen and recognized during the night if the illumination were raised to the normal daylight level (WMO 2003) (see also Section 4.1). Air quality and visibility is measured at a number of stations by the Malaysian Meteorological Department; a more detailed discussion on air quality monitoring and measurements in Malaysia will be presented in Chapter 4. Here we display data from one of these stations, Petaling Jaya, in Selangor, Malaysia, to demonstrate haze occurrence and characteristics in Malaysia. Figure 1.6 shows a plot of daily visibility against day from 1999 to 2008. White, yellow, green, violet and red colours indicate clear (above 10 km visibility), moderate (5 – 10 km visibility), hazy (2 – 5 km visibility), very hazy (0.5 – 2 km visibility) and extremely hazy (less than 0.5 km visibility) conditions respectively. For most years, a drop in visibility can be observed at the end of the year, indicating the occurrence of increased haze. The extreme values seen for 2005 correspond to Figure 1.4 and Figure 1.5.

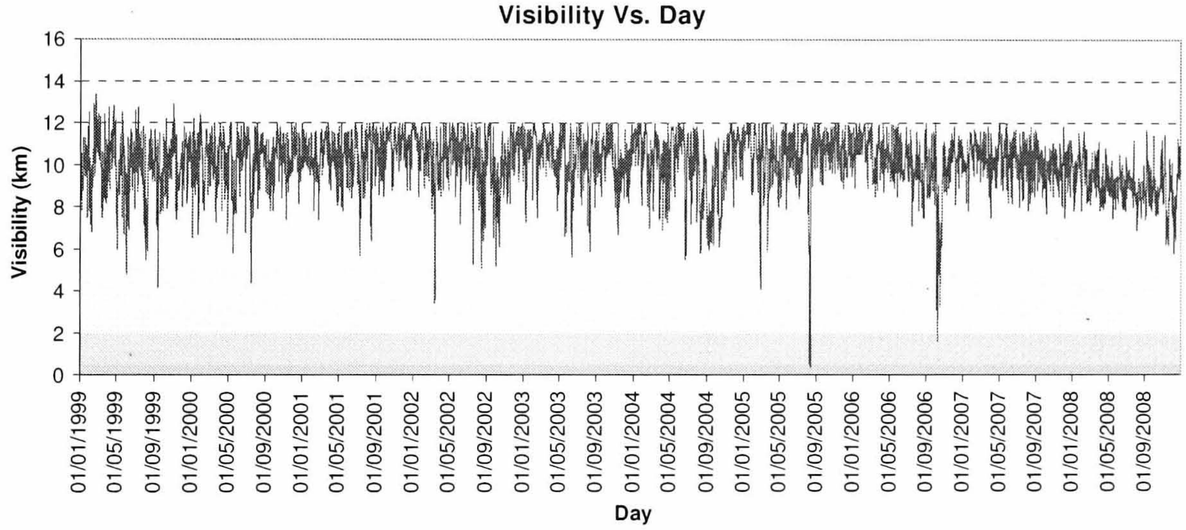


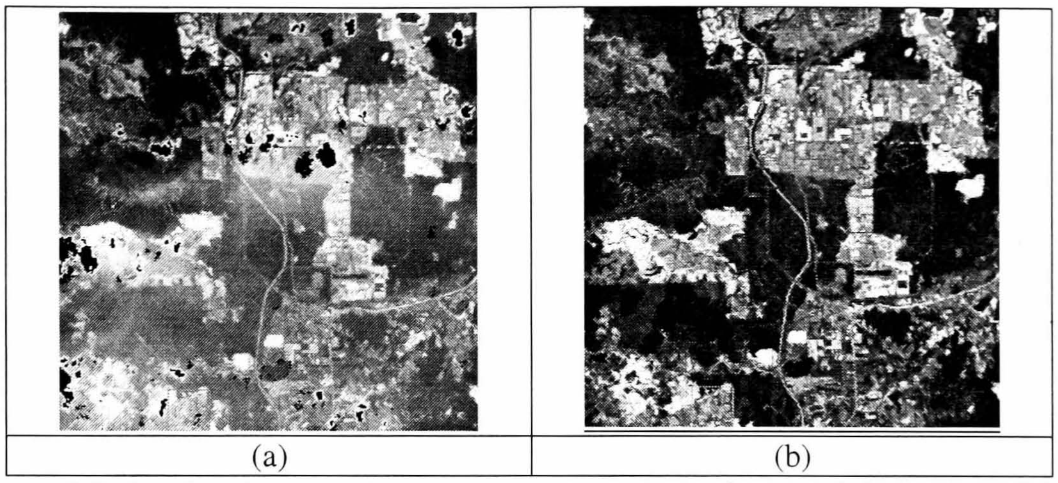
Figure 1.6: *Visibility against day for Petaling Jaya from 1999 to 2008.*

Table 1.1 summarises the number of days for clear, moderate, hazy, very hazy and extremely hazy conditions in Petaling Jaya from 1999 to 2008. The years which have the most days when the visibility is 10 km and less are 2008 (309), followed by 1999 (196), 2007 (159) and 2006 (156). *We will show later in this thesis (Chapter 4 and 5), that when visibility drops to less than 10 km, haze causes classification accuracy to drop below an acceptable level. Since classification accuracy is the key element that* *determines the quality of satellite derived-maps, such situations could severely degrade the quality of land cover maps for the area.*

Table 1.1: *Number of days for clear, moderate, hazy, very hazy and extremely hazy conditions in Petaling Jaya from 1999 to 2008.*

Year	> 10 km (Clear)	5 to 10 (Moderate)	2 to 5 (Hazy)	500 to 2 (Very hazy)	< 0.5 km (Extremely hazy)
1999	179	184	2	0	0
2000	229	135	1	0	0
2001	265	100	0	0	0
2002	237	126	2	0	0
2003	234	131	0	0	0
2004	222	143	0	0	0
2005	259	101	3	1	1
2006	209	149	6	1	0
2007	206	159	0	0	0
2008	56	309	0	0	0

To visualise the effects of haze, Figure 1.7 shows Landsat images of Bukit Beruntung in Selangor (approximately 30 km from Petaling Jaya) for (a) 6 August (5.8 km visibility) and (b) 22 August (11.7 km visibility) 2005; Landsat bands 3, 2 and 1 are assigned to red, green and blue respectively. For 6 August (Figure 1.7(a)), small patches of cloud and its shadow, masked in black, can be seen mainly on the top and left of the image, while haze covers mainly the middle and bottom parts of the image. Due to the haze, the distinction between different types of land cover is blurred, and their spectral signatures are altered. For 22 August (Figure 1.7(b)) the land cover can be recognised easily due to the clear conditions; bright areas represent urban, while dark areas, agricultural sites. *We will show in Chapter 5 that the haze seen in Figure 1.7(a) will cause a drop of 25% in classification accuracy.*



(a) (b)

Figure 1.7: *Landsat images of Bukit Beruntung, acquired on (a) 6 August and (b) 22 August 2005, with bands 3,2, 1 assigned to red, green and blue.*

Haze also greatly hinders projects that require continual near real-time data, such as precision farming and NaREM (particularly concerning natural hazard, e.g. landslides). The possible impact on precision farming of paddy is given here. Paddy requires approximately 120 days to grow before it can be harvested, and satellite data is one of the key inputs in monitoring its growth stages (e.g. through satellite-derived vegetation indexes). Figure 1.8 shows visibility against Landsat overpass date (i.e. 16 days interval) for 1999 to 2008. For convenience, data with visibility 10 km and less are indicated by vertical bars. The red bars are data that overlap with the main paddy planting season (August to December), while the black bars show dates outside the planting season (January to July).

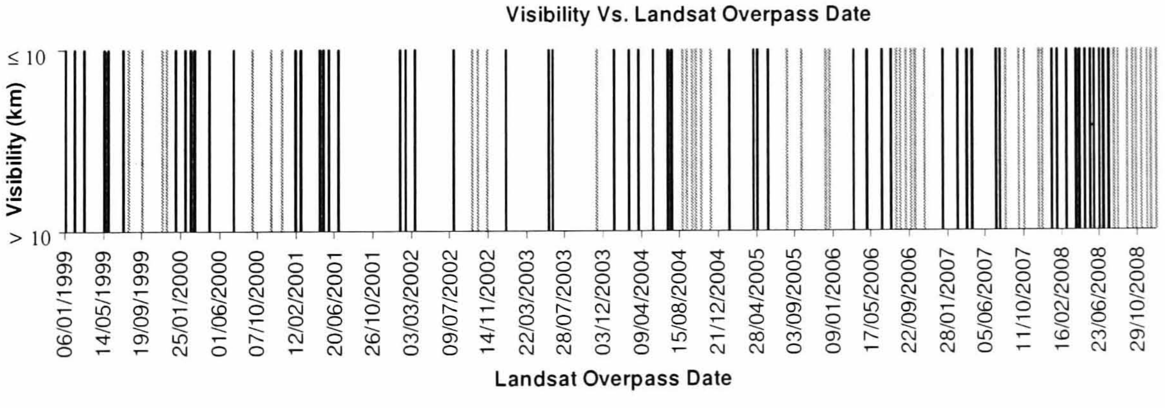


Figure 1.8: *Visibility against Landsat overpass date in 2005 for Petaling Jaya. Black (off season) and red (main season) bars are Landsat data having visibility 10 km and less, no bar indicates data with visibility more than 10 km. The red bars are the haze affected data for the main planting season, while the black bars, for those outside the planting season.*

Table 1.2 summarises the number of Landsat overpass days overlapping with the main planting season (10 days) and having visibility 10 km and less, for Petaling Jaya from 1999 to 2008. Landsat gives 23 overpasses of the area each year, and 10 of them occur during the paddy main planting season. Out of these data, some of them have visibility 10 km and less, indicating that they were significantly affected by haze. 2008 has the most of haze-affected data (i.e. 9), followed by 2006 and 2004 (6), 2007 (5) and 2005 and 1999 (4). For other years, the number of days was 3 days and less. *Consequently, for 2008, only one acquisition could be used during the main planting season, and only 4 for 2006 and 2004. 2007 (5) and 2005 and 1999 (6), have the most haze-free data.*

Table 1.2: *Number of Landsat overpass days occurring during the main planting season and the number having visibility 10 km and less, for Petaling Jaya from 1999 to 2008.*

Year	No. of overpasses overlapping with the main planting season and have visibility 10 km and less
1999	4
2000	3
2001	0
2002	3
2003	1
2004	6
2005	4
2006	6
2007	5
2008	9

1.2 Statement of the Problem, Aim and Objectives of the Thesis

Haze modifies spectral signatures and reduces the accuracy of land cover classification using satellite data (Kaufman and Sendra 1988). Also, haze can significantly hinder practices that require continual input from remote sensing data (e.g. precision farming). The current approach to handling hazy data is simply to remove the data from further analysis; however, this causes losses of valuable surface information (Lu et al. 2007). On the other hand, if these data are considered for further processing, they are likely to degrade subsequent satellite-derived information (e.g. land cover and vegetation index maps) unless they are first corrected for the haze.

Hence, the aim of this thesis is to develop and test methods for removing haze from satellite data. Achieving this aim requires a systematic development of several subsidiary objectives:

(a) Masking cloud from remote sensing data (Chapter 2).
(b) Classifying land covers in the study area (Chapter 3).
(c) Assessing the effects of haze on land covers (Chapter 4) and
(d) Developing and testing of haze removal procedures (Chapter 5).

1.3 Thesis plan

Land cover mapping from remote sensing data is an important asset in providing useful information for managing land activities at local and global scales. Unfortunately, at certain places and times, satellite data are affected by haze. To overcome this problem, this thesis develops and tests methods for removing haze from satellite data and is organised as follows:

Haze shares some characteristics with cloud, which also creates problems for land cover classification. Hence, Chapter 2 is concerned with cloud detection and masking for Malaysian satellite data. In this chapter, MODIS data, due to the richness of the spectral bands, will be analysed to develop understanding of the spectral properties of cloud (Ackerman et al. 1998; Ackerman et al. 2010). We then relate and apply the analysis to Landsat data, which will be used in later chapters.

In Chapter 3, we carry out land cover classification using Landsat data based on the ML (Maximum Likelihood) classification. The performance of ML classification is assessed by comparison with the ISODATA (Iterative Self-organizing Data Analysis

Technique) clustering in terms of visual analysis, classification accuracy, band correlations and decision boundaries (Thomson et al. 1998; Low and Choi 2004).

Chapter 4 is mainly concerned with investigation of haze effects on satellite data. For this purpose, hazy datasets are modelled and simulated by incorporating the haze path radiance and the effects of signal attenuation into the Landsat dataset. This makes use of the 6S (Second Simulation of a Satellite Signal in the Solar Spectrum) radiative transfer model (Vermote et al. 1997; Kotchenova et al. 2006). The simulated hazy datasets undergo ML classification and accuracy analysis (Song et al. 2001; Zhang et al. 2002) so that the effects of haze on the classification can be assessed.

Chapter 5 is devoted to the development and testing of a haze removal procedure for hazy satellite data (Chavez 1988; Schott et al. 1988; Liang et al. 2001). Physical and mathematical descriptions of the haze removal are discussed. We assess the haze removal performance based on the quality of simulated and real data and classification accuracy. For the real data, Landsat data from Bukit Beruntung from 6 August 2005 will be used due to the hazy conditions, while a clear satellite data of the same area from 22 August 2005 will be used as a reference data (see Figure 1.7).

Chapter 6 summarises the main conclusions of this thesis and gives recommendations for future work.

Chapter 2

Cloud Detection and Masking

2.1 Introduction

Our main concern in this thesis is to characterise the effects of haze on satellite data of land surfaces, and to use this understanding to develop methods to mitigate these effects, particularly in the context of land cover mapping, though the outcomes are also relevant for other remote sensing applications. However, atmospheric contamination of surface information is also caused by cloud, which, if thick, can completely obscure the surface within the satellite field of view or, if thin, attenuate solar radiation both on the incident path and after reflection and scattering at the surface. This is particularly important over tropical regions where cloud is persistent.

The later chapters of this thesis rely heavily on methods of land use classification, which need to take account of cloud (and cloud shadow). One approach would be to simply treat cloud as another land cover type and use the same methods as for any other land cover. However, this is unsatisfactory for at least two reasons: (1) unlike most land covers, cloud has known physical characteristics affecting its spectral response at different wavelengths, and it is advantageous to exploit these in its detection; (2) cloud occurs in different types, and hence characterising it in an overall classification scheme is not straightforward. Hence, in common with many other studies (Meng et al. 2009; Luo et al. 2008; Ackerman et al. 2006), we prefer to use an approach that detects and masks (thick) cloud and cloud shadows before undertaking land cover classification.

We are also interested in studying cloud because cloud and haze share some spectral properties; this is exploited in Chapter 4 where cloud data are used to learn some of the statistical properties of haze. Furthermore, cloud detection schemes have difficulty in removing thin cloud, and in many cases thin cloud needs to be treated similarly to haze (Ji 2008; Moro and Halounova 2007; Lu 2007; Zhang et al. 2002).

The primary data used in later chapters is from Landsat, and the scheme we use to deal with data in these images is MODIS cloud mask. However, this scheme is intended for global use, and may not be optimised for tropical regions, such as Malaysia. We therefore perform a critical analysis of this scheme in order to assess its likely weaknesses when used over Malaysia (and hence ways in which it might be improved, although we do not develop such an improved scheme here). To do this we make use of the spectrally rich satellite data provided by MODIS, which is equipped with 36 bands ranging from visible to thermal wavelengths, several of which overlap with those of Landsat. Although MODIS has much coarser spatial resolution than Landsat (250 to 1000 m vs. 30 to 120 m), this analysis is valid, since our principal concern is spectral behaviour.

The principal aims of this chapter are therefore:

1. to analyse the relationship between the spectral properties of cloud and haze.
2. to determine a suitable cloud detection method for Malaysia
3. to analyse the method most relevant for this thesis
4. to apply the cloud analysis onto Landsat data

We begin in Section 2.2 with a brief survey of cloud properties, including their morphology, physical properties and associated spectral signatures, placing particular emphasis on the types of cloud and their occurrence throughout the year in a Malaysian context. The relations between haze and cloud are also discussed in this section. Section 2.3 explains how the physical and spectral properties discussed in Section 2.2 can be translated into detection approaches for cloud and cloud shadow, and follow this in Section 2.4 with a survey of the main approaches relevant to this thesis. In Section 2.5 we provide a critical analysis of the scheme most important for this thesis (MODIS), in particular examining how well the global thresholding approach lying at the centre of this scheme is adapted to Malaysian conditions and likely errors arising from use of the global schemes. This section also describes the datasets and methods for cloud masking over Malaysia.

:arrying out this analysis, we have to confront the issue of how to validate the ıd detections. Clearly we have no independent data which we can use as a :rence, so have adopted a pragmatic approach which is visual analysis. As an ended analysis we will compare the results with Landsat ACCA (Automatic Cloud /er Assessment) scheme. We will show that the analysis of MODIS scheme can be lied to Landsat data with reasonably high accuracy. To further validate this, we ry out the scheme on Landsat data with different cloud conditions.

:tion 2.6 summarises the chapter and explains which aspects of it will be exploited :r in the thesis.

2.2 Cloud Morphology and Physical Properties

Over 60 per cent of the Earth's surface is covered by cloud at any time (Rossow et al. 1993, Choi and Ho 2009), where a cloud is a visible mass of condensed water droplets or ice crystals suspended in the atmosphere above the Earth's surface. Cloud is made of either water droplets or ice particles or both with diameters ranging from 10 to several hundreds μm. It scatters electromagnetic energy in UV through mid-infrared wavelengths due to the much larger particle diameter than the wavelengths and therefore causing Mie scattering. This leads to the two most obvious features of clouds seen from space; they are white and bright. The primary cloud types are cumulus, stratus and cirrus (Figure 2.1). Those further classified from the main types include cumulonimbus, nimbostratus, stratocumulus, altocumulus and cirrocumulus; depending on their height and appearance from ground, i.e. *cirro-* (curl), *alto-* (mid), *strato-* (layer), *nimbo-* (precipitation) and *cumulo-* (heap).

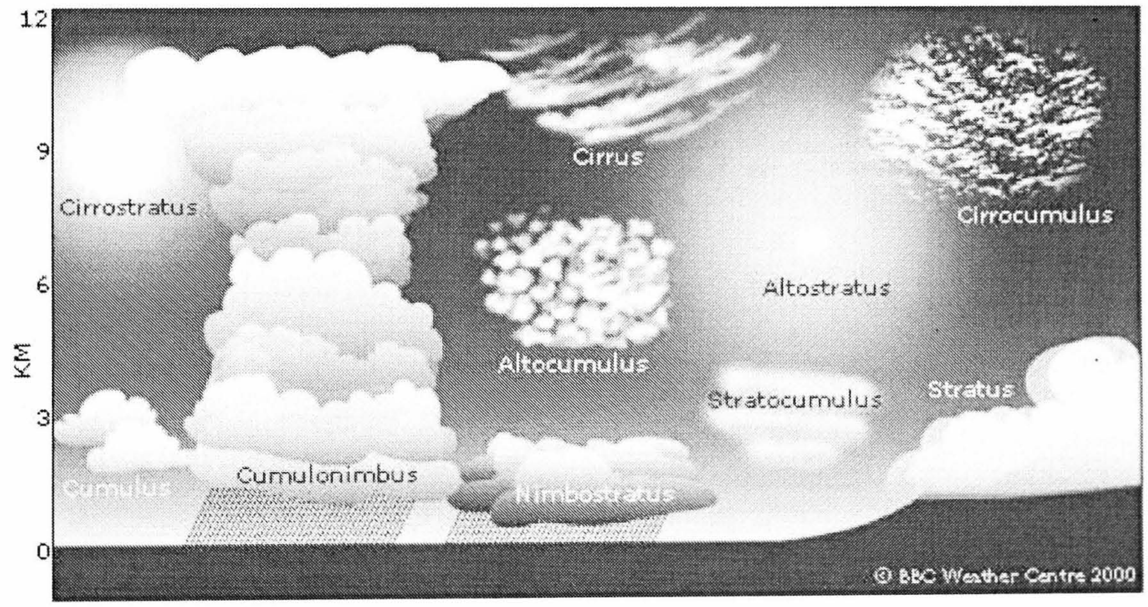


Figure 2.1: *Common types of cloud (BBC 2011).*

For simplificaiton, these cloud types often categorised based on their heights. Table 2.1 shows clouds that are typically divided into three main categories, i.e. high-level clouds (i.e. cirrus, cirrostratus and cirrocumulus), mid-level clouds (i.e. altostratus and altocumulus) and low-level clouds (i.e. stratus, stratocumulus, nimbostratus, cumulus and cumulonimbus). From these clouds, only cumulus and cumulonimbus fall into convective clouds, i.e. those have a larger vertical extent (thickness), but smaller horizontal extent; while the remaining are stratiform clouds, i.e. those have a far larger horizontal extent than the thickness.

Table 2.1: *Category, type and description of clouds (Weather Forecast Office 2011).*

Category	Type		Description
High level clouds (5000 – 13000 m) They are given the prefix *cirro-*. Due to cold temperatures at these levels; the clouds primarily are composed of ice crystals and often appear thin, streaky, and white	Cirrus (Ci)	Stratiform clouds	They are the highest of all clouds, and are thin and wispy. They composed entirely of ice crystals, which evaporate high above the earth surface.
	Cirrostratus (Cs)		Sheet-like thin clouds that usually cover the entire sky. Sometimes, the sun or moon will appear to have a halo around in the presence of cirrostratus clouds. They consist of ice crystals.
	Cirrocumulus (Cc)		Appear across the sky as patches or thin layers of cloud consisting of tiny individual smaller clouds. They are usually a transitional phase between cirrus and cirrostratus clouds and composed of ice crystals.
Mid level clouds (2000 – 5000 m) They are given the prefix *alto-*. Depending on the altitude, time of year, and vertical temperature structure of the troposphere, these clouds may be composed of liquid water droplets, ice crystals, or a combination of the two, including super-cooled droplets (i.e., liquid droplets whose temperatures are below freezing).	Altostratus (As)		Known as *strato* type clouds that possess a flat and uniform type texture in the middle latitudes. They can appear as thin or thick layers of clouds.They composed of both water droplets and ice crystals, and produce occasionally light showers or snow.
	Altocumulus (Ac)		Known as *cumulo* type clouds that usually occur as a layer or patch of more or less separate cloudlets in the form of heaps, rolls, billows or pancakes. They mainly consist water droplets of, but ice crystals are often present. Usually they produce no or very occasional light rain.
Low-level clouds (below 2000 m) They normally consist of liquid water droplets or even super-cooled droplets, except during cold winter storms when ice crystals (and snow) comprise much of the clouds.	Stratus (St)		Appear uniform and flat, producing a grey layer of cloud cover which may be precipitation-free or may cause periods of light precipitation or drizzle. They consist of water droplets and commonly form near coasts and mountains..
	Stratocumulus (Sc)		Usually appear as low and puffy clouds but sometimes they line up in rows or spread out. They consist of water droplets and may produce light rain or snow.
	Nimbostratus (Ns)		They formed from thick, dense stratus or stratocumulus clouds that produce steady rain or snow. They common occur in middle latitudes and composed of water droplets, snow flakes and ice crystals.

	Cumulus (Cu)	Convective clouds	Thick fluffy clouds, with a flat base. A cumulus cloud starts forming at a very low altitude, but it has the ability to cover a significant vertical distance, which gives it a gigantic appearance. They commonly occur over land and located worldwide, except polar regions. They composed of water droplets and can produce brief showers.
	Cumulonimbus (Cb)		Much larger and more vertically developed than cumulus clouds which form in a more stable atmosphere. Larger cumulonimbus clouds can produce heavy downpours and even thunderstorms. They commonly occur in tropics and temperate regions but rare at poles and may composed of water droplets and ice crystals.

In cloud observation, cloud amount can be measured as the average "amount" for a given period is the product of frequency-of-occurrence (f) and amount-when-present (awp). For example, if in a particular season altocumulus is reported in 30% of the usable observations and if it covers 40% of the sky when it is present, then f=0.3, awp=0.4, and the seasonal average altocumulus amount is 12% (=0.3*0.4) (Warren and Hahn 2002).

Figure 2.2 shows Malaysian and global monthly average of daily cloud amount for 26 years, i.e. from 1971 to 1996; Malaysia has about 30% more cloud than the global. For Malaysia, the highest cloud amount occurs in November (87%), followed by October, September and December (86%), observed from land stations located approximately 2.5° North and 102.5° East (Hahn and Warren 1999). The higher cloud amount at the end of the year is due to the occurrence of Northeast Monsoon (November to February) which brings much rain to Malaysia, while the lower cloud cover in the middle of the year is associated with Southwest Monsoon (May to September) which brings less rain.

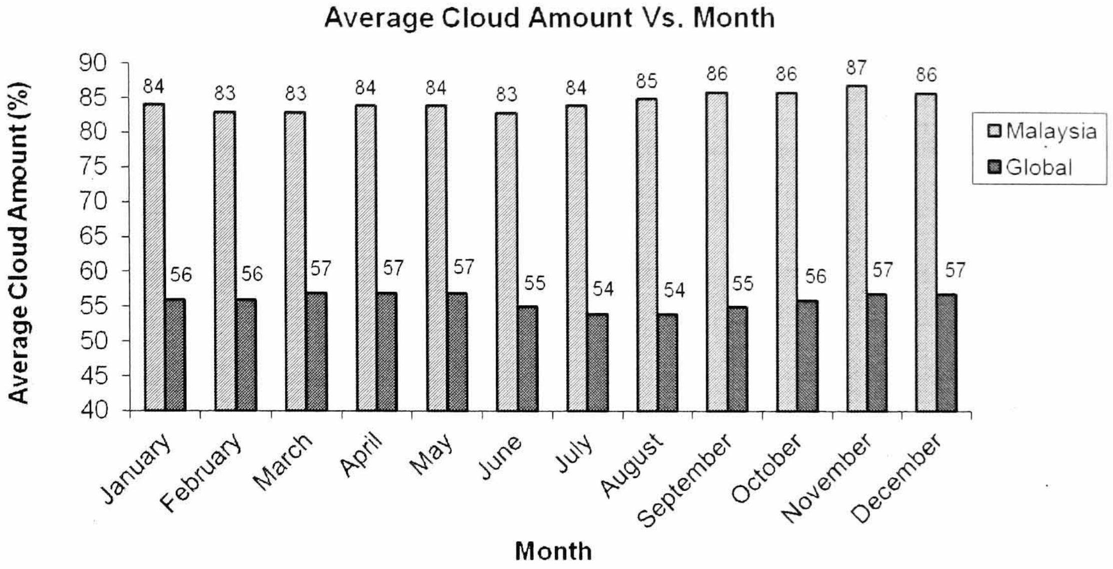


Figure 2.2: *Overall average cloud amount versus month for Malaysia from 1971 to 1996 (Hahn and Warren 1999).*

Figure 2.3 shows plots of the average of daily cloud amount for each cloud type against month for Malaysia from 1971 to 1996 observed from land stations. It can be seen that cloud amount for cirrus, cirrostratus and cirrocumulus (high clouds) and altocumulus are much higher in term of percentage of cloud amount than other cloud types and invariant throughout the year, while stratus is quite low, but also invariant throughout the year. During the Northeast Monsoon (November to February) and Southwest Monsoon (May to September), there is a noticeable increase in cumulus, cumulonimbus and nimbostratus. It also can be seen that stratocumulus is much higher during the Northeast Monsoon than during the Southwest Monsoon. In this period occurrence of completely clear sky has not been recorded.

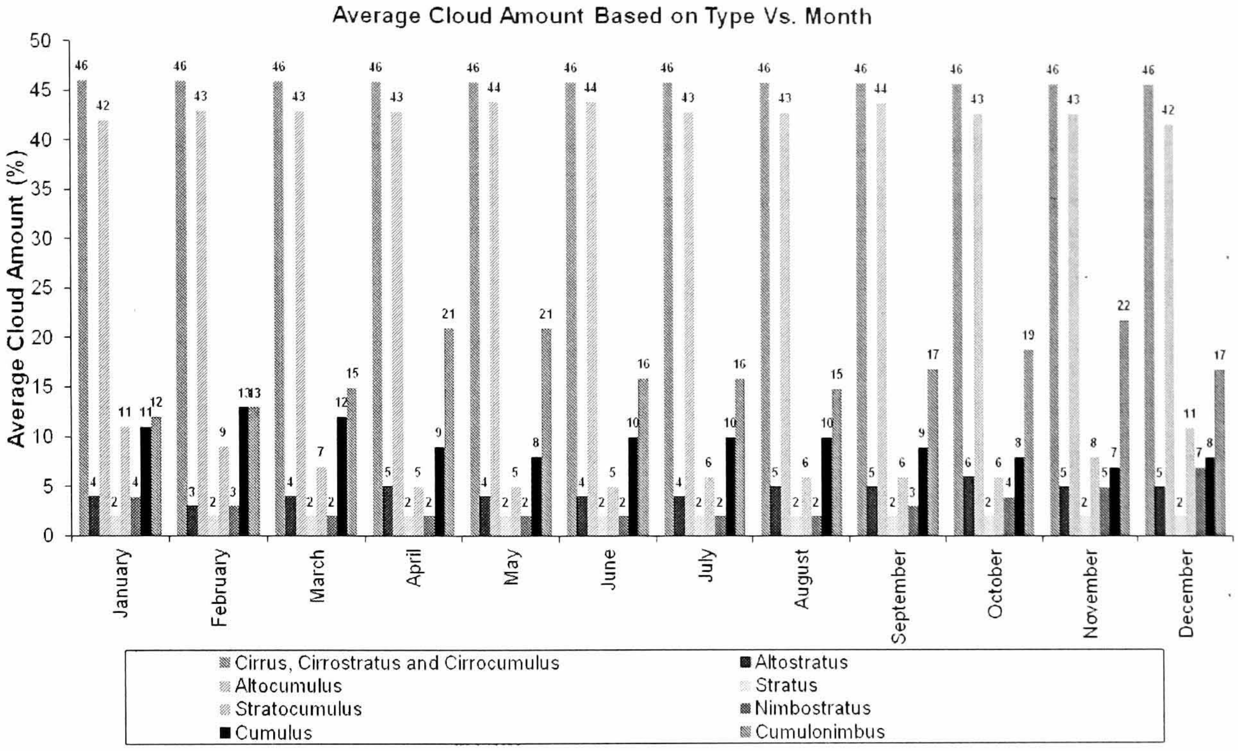


Figure 2.3: *Average cloud amount based on cloud type versus month for Malaysia from 1971 to 1996 (Hahn and Warren 1999).*

Spectral Properties of Cloud and Their Relationship with Haze

Both cloud and haze scatter solar radiation but the former has higher scattering intensity, therefore is more reflective than the later. Haze often occurs at a wider horizontal scale than cloud, so tends to distribute more homogeneously and therefore has a lower standard deviation than cloud (Martin et al. 2002). Figure 2.4 shows histogram of 3 x 3-window standard deviation of haze and clouds when sampled from 0.55 μm MODIS band 12 and 4 (1000 and 500 m spatial resolutions); haze has lower standard deviations than cloud in both bands. As haze gets more severe, it scatters more solar radiation and eventually becomes as reflective as cloud. Hence, it is sensible to assume that if haze is very thick, it possesses the standard deviation of cloud. In our study, we will make use of the cloud properties (i.e. covariance) to simulate haze for use in studying its effects on satellite data (Chapter 4).

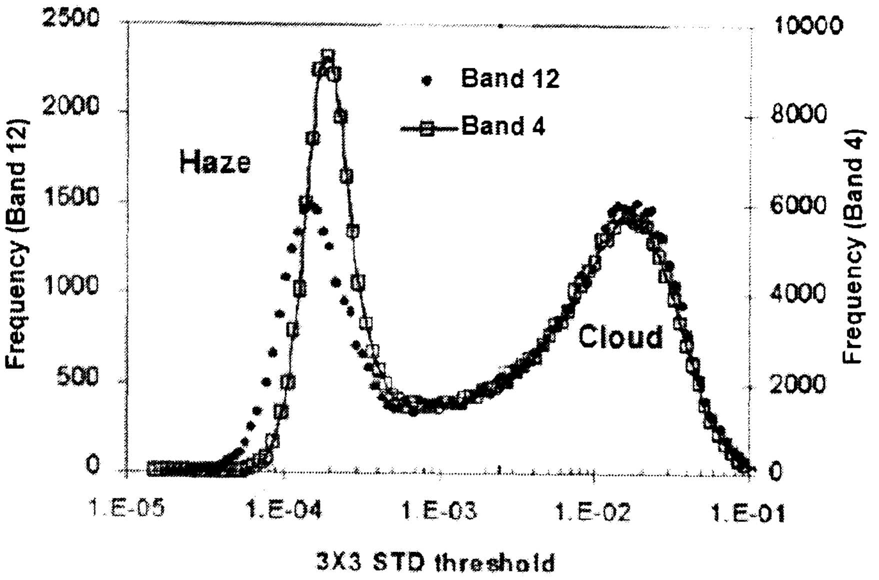


Figure 2.4: *Histogram of 3 x 3-window standard deviation at 0.55 μm from band 4 (500 m spatial resolution) and band 12 (1000 m spatial resolution) showing the basic separation between haze and clouds at 1000 and 500 m spatial resolutions (modified after Martin et al. 2002).*

2.3 Cloud Detection from Satellites

Cloud detection from satellites data is based on radiative properties in visible and thermal infrared spectral range. Cloud appears very brighter in the visible wavelengths due to the shorter path of the photon come from the sun and reflected by cloud particles towards the satellite sensor, than in a cloud-free atmosphere, while darker in the thermal wavelengths due to the lower temperature than the surroundings (Couvert and Seze 1997; Chen et al. 2002; Jose et al. 2003).

In visible wavelengths, the larger the water content and the thicker the cloud, the higher the reflectance measured from the satellite sensor, therefore it appears brighter (Li et al. 2003). The convective clouds look brighter than the stratiform clouds because they contain more water droplets and are thicker. Among the convective clouds, cumulonimbus is brighter than cumulus. Hence, in most cases, cloud formed in the lower levels is brighter that the higher levels. In near infrared wavelengths, a cloud with high cloud top height looks bright and a cloud with low cloud top height look dark. Among the stratiform clouds, high level clouds are the brightest, followed by middle level clouds and low level clouds. In terms of forms, a stratiform cloud often appears with fairly large extent of cloud area, while the convective cloud exists as a rather small cloud cluster. In terms of texture, a stratiform cloud has a smooth and even cloud surface, while a convective cloud has an uneven and ragged cloud surface.

Spherical albedo represents a mean value of the reflection function over all solar and observational zenith and azimuth angles and the reflection function is subject to particle size (King et al. 1992). Figure 2.5 shows plots of spherical albedo for various cloud effective radius r_e as a function of wavelength. It is obvious that cloud droplets with smaller r_e gives higher spherical albedo than those with bigger r_e; therefore, the higher the reflectance, is the smaller the cloud effective radius is. Also shown is the wavelength locations of selected MODIS bands which signify the relevance of using MODIS bands in cloud detection.

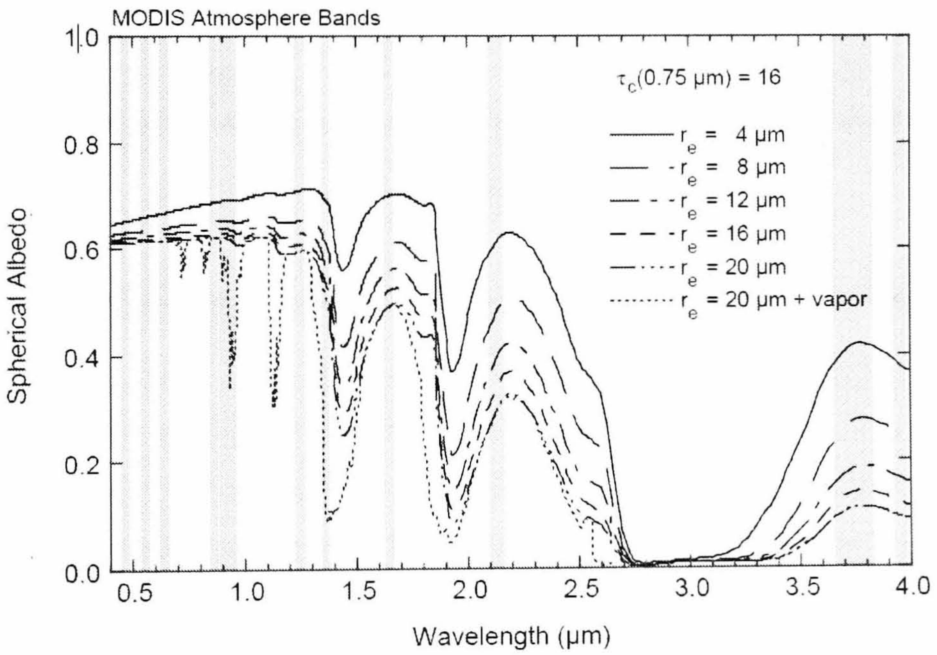


Figure 2.5: *Cloud spherical albedo for selected effective radius of cloud droplets as a function of wavelength (King et al. 1992). Also shown is the location of selected MODIS bands.*

In thermal infrared wavelengths, thicker cloud has lower brightness temperature than thinner cloud, therefore appears darker; hence convective clouds often look darker than stratiform clouds. Figure 2.6 shows the brightness temperature spectrum between

9.1 and 16.7 μm over clear scene and optically thin, moderate and thick cirrus clouds, and location of MODIS bands 30 to 36 (King et al. 1992). From 10 to 13 μm, it is clear that thick cloud has the lowest brightness temperature compared with those of moderate and thin cloud and clear scene.

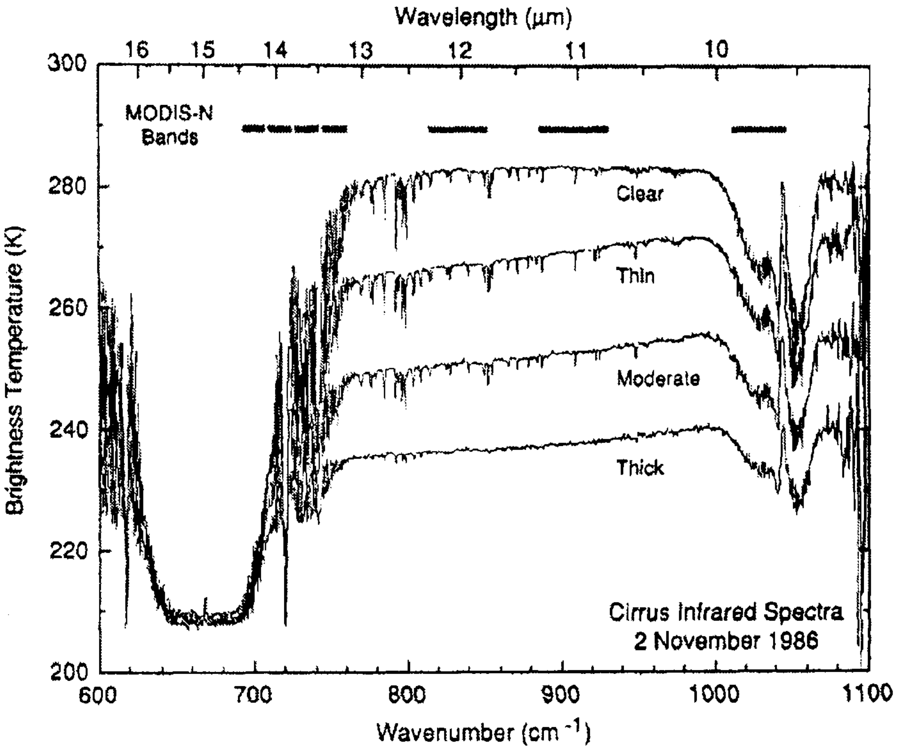


Figure 2.6: *Brightness temperature spectrum between 9.1 and 16.7 μm over clear scene and optically thin, moderate and thick cirrus clouds. Also shown are location and bandwidth of MODIS bands 30 to 36 (King et al. 1992).*

Due to the higher reflectance and lower temperature values than land, cloud can be identified by selecting threshold values that denote the lowest cloud reflectance and the highest cloud temperature in an image (Buriez et al. 1997; Baum and Trepte 1999; Bendix et al. 2004). The exceptions to this rule in the visible wavelengths are snow, ice, and white sand, which can have reflectance values that are greater than or equal to the cloud reflectance values (Di Vittorio and Emery 2002). Such exceptions can be ignored as most of the study areas are highly vegetated land areas.

Clouds have higher optical thicknesses in the visible spectral range compared to all other atmospheric constituents such as haze and fog, therefore often block the surface from the solar radiation. Cloud, can consist of water or ice droplets, often have different spectral properties at different wavelengths so requires different spectral bands with appropriate thresholds. The spectral properties of cloud over land differ significantly from ocean. Hence, using different thresholds for such conditions tends to give better results than using the same thresholds.

Most of cloud detection scheme employs cloud detection algorithm involving a number of tests which are based on differences between the spectral properties of cloud and non-cloud features. The tests are applied to each pixel within a satellite field-of-view, where the pixels that are flagged as cloud in some of the tests are judged as cloudy; in other words only those identified as cloud-free pixels in every test are judged to be cloud-free.

In day time, both visible and thermal bands of the satellite data can be used, so detection of cloud is more informative than night time. Generally high and thick clouds are easier to detect than others. The accuracy of cloud detection depends very much on the properties of the underlying surface. Higher accuracy can be gained for remote sensing data covering surfaces having fairly constant temperature and emissivity (Saunders 1986). This is due to the little variation of the spectral properties for these surfaces; this provides a quite constant difference between them and those of the cloud.

Cloud detection tests can be categorised into four categories, i.e. brightness temperature test, brightness temperature difference test, simple reflectance test and reflectance ratio test.

(a) Brightness Temperature Tests

The tests commonly performed using brightness temperature measurements are from 11 μm and 14 μm wavelengths. 11 μm measurement was initially used for partial coherence test (Coakley and Bretherton 1982; Saunders 1986; Franca and Cracknell 1995). The idea behind this technique is that the absolute value of BT should be lower than surface area that the variability of brightness temperature for cloudy pixels should be higher than clear-sky pixels. This can be carried out by using standard deviation value of an array of pixels. For high latitude regions, cloud pixels are indicated with standard deviation less than 0.2 K, while 0.4 K for equatorial regions. However, the main problem of this technique is its performance in detecting cloudy pixels over land and coastal areas (Saunders 1986; France and Cracknell 1995). In more recent years, Ackerman et al. (2010) used this test as a clear-sky restoral test over sea and land; if pixels are determined as cloudy from initial tests, it may be restored to clear given the brightness temperature exceeds certain thresholds.

CO_2 absorption bands (near 14 μm) can be used to distinguish transparent clouds from opaque clouds and clear-sky. Using this test, clouds at various levels of the atmosphere to be detected, though is particularly effective for detecting thin cirrus clouds that are often missed by simple infrared and visible tests (Wylie et al 1994).

(b) Brightness Temperature Difference Test

The frequently used brightness temperature difference tests are $BT_{(11)} - BT_{(12)}$, $BT_{(11)} - BT_{(3.9)}$ and $BT_{(8.6)} - BT_{(11)}$.

$BT_{(11)} - BT_{(12)}$ test can be used to detect thin cloud (i.e. cirrus) because they are larger than that of clear-sky and thick cloud conditions (Inoue 1987; Saunders and Kriebel 1988). This test has been widely used for cloud screening using satellite sensor such as MODIS, NOAA AVHRR and GOES.

$BT_{(8.6)} - BT_{(11)}$ test indicate certain cloud properties based on the difference of water vapour absorption between 8.6 and 11 μm wavelengths. This is because at 8.6 μm wavelength, ice/water particle absorption is low, while atmospheric water vapour absorption is quite high; the reverse is true at 11 μm wavelength. Large positive values of $BT_{(8.6)} - BT_{(11)}$ indicate the presence of cirrus clouds (ice clouds), due to the larger increase in the imaginary index of refraction of ice over that of water. On the other hand, negative values of $BT_{(8.6)} - BT_{(11)}$ indicate clear conditions, due to stronger atmospheric water vapour absorption at 8.6 μm than at 11 μm (Ackerman et al. 1998).

$BT_{(11)} - BT_{(3.9)}$ test can be used to differentiate between cloud over land and water; its value over land is different from over water. For cloudy pixels over land, the longwave minus shortwave brightness temperature (i.e. $BT_{(11)} - BT_{(3.9)}$) has a large negative value during the day for thick clouds. This is because much of the energy sensed by the satellite comes from the Earth's surface and atmosphere below the cloud, and the 3.9-μm channel's response to warm pixel temperatures is greater than it is at 11 μm, resulting in negative difference values during the day.

(c) Reflectance Test

The frequently used reflectance tests are such as $R_{(0.66)}$, $R_{(1.38)}$ and $R_{(0.76)}$; $R_{(0.66)}$ has been widely used in discriminating clouds from vegetated land due to the difference reflectance properties measured at 0.66 μm wavelength. $R_{(1.38)}$ in day time can be used to detect the presence of high-level clouds, particularly thin cirrus, due to the strong water vapour absorption at that region (Gao et al. 1993). Another useful band is $R_{(0.76)}$, which is based on oxygen absorption band at 0.76 μm and have been used in the past to estimate pressure in MERIS 0.76 μm band (band 11) (Fischer et al. 1997), so is also useful for cloud detection. Surface pressure can be calculated from the ratio of pixel observations made at 0.76 μm to observations made at 0.75 μm. The presence of thin cirrus cloud can produce errors of up to 150 hPa to the calculated surface pressure. This effect on the surface pressure can be used as an indirect means of detecting thin cirrus cloud.

(d) Reflectance Ratio Test

This test was proposed by Saunders and Kriebel (1988) and is based on the ratio of reflectance in the near-infrared and visible infrared bands. For cloudy pixels, due to similar reflectance properties resulted from quite similar scattering effects (Mie Scattering) in both spectral bands, $R_{(0.87)}/R_{(0.66)}$ values are close to 1, i.e. between 0.8 and 1.1 (Saunders and Kriebel 1988; Ackerman et al. 1998). For land pixels, $R_{(0.87)}/R_{(0.66)}$ values are higher than 1 due to the higher reflectance in the near-infrared than the visible band.

2.4 Literature Survey

A number of global cloud detection schemes have been developed over the years but the most popular ones are such as APOLLO (AVHRR Processing Scheme Over Cloud, Land and Ocean), CLAVR (Clouds from AVHRR), EUMETSAT (European Organisation for the Exploitation of Meteorological Satellites), SCANDIA (SMHI(Swedish Meteorological and Hydrological Institute) Cloud Analysis Model Using Digital AVHRR Data), ISCCP (International Satellite Cloud Climatology Project), CERES (Clouds and the Earth's Radiant Energy System), MODIS (Moderate-resolution Imaging Spectroradiometer) and Landsat ACCA (Automatic Cloud Cover Assessment). For convenience, we denote the satellite measured solar reflectance as R, and the infrared radiance as brightness temperature denoted as BT. Subscripts with bracket refer to the wavelength while subscript without bracket refer to the satellite band number, at which the measurement is made.

(a) APOLLO scheme

The APOLLO scheme was among the earliest scheme and used all five NOAA AVHRR (Advanced Very high Resolution Radiometer) (Saunders and Kriebel 1988). The five AVHRR band wavelength ranges are 0.58 – 0.68 μm (R_1 – visible), 0.72 – 1.10 μm (R_2 – near infrared), 3.55 – 3.93 μm (R_3 – middle infrared), 10.3 – 11.3 μm (R_4 – thermal infrared) and 11.5 – 12.5 μm (R_5 – thermal infrared). This scheme is designed for applications using full spatial resolution HRPT (High-Resolution Picture Transmission) and LAC (Local Area Coverage) and reduced spatial resolution GAC (Global Area Coverage) data formats, particularly for NOAA 7 through 14.

The tests involved can be categorised based on surface types, i.e. ocean surfaces, vegetated land, arid land, and snow and ice. Each pixel in these categories undergoes a sequence of threshold tests to determine pixel status, i.e. fully cloudy, partially cloudy, cloud free, and snow–ice (Kriebel et al. 2003). The pixel identification is carried out in three stages: In stage 1, the tests are: the gross temperature using BT_5; the spatial coherence thermal test over sea surface based on the standard deviation thresholds on BT_4; the thin cirrus detection based on BT_4 – BT_5; the dynamic visible band test using R_1 and the dynamic ratio test using R_2/R_1 over land and water. In stage 2, the R_2/R_1 and the spatial coherence tests are repeated in order to identify the fully cloudy pixels among the partially cloudy pixels using slightly different thresholds (Kriebel et al. 2003). In stage 3, R_3 is used to identify snow-ice pixels. The main disadvantage of Apollo scheme is the tests make use only the five NOAA AVHRR bands, therefore may miss some clouds that cannot be detected within the bands' wavelength range.

(b) CLAVR scheme

CLAVR is a main cloud identification scheme for use of AVHRR global data processing (Stowe et al. 1999), aiming to work with AVHRR GAC (Global Area Coverage). It is particularly designed for NOAA 15 through 18, which are equipped with bands 3A (1.58 - 1.63 μm) and 3B (3.54 - 3.87 μm), so has better detection capability compared to Apollo scheme. CLAVR is designed to be clear-sky conservative (i.e. ensures that no cloudy pixel is identified as clear sky) and uses ancillary datasets (e.g. surface type maps, digital elevation maps and climate data) to set up thresholds. The tests include R_2 (over water) and R_1 (over land) as the gross contrast test, BT_4 to identify bright and cold pixels corresponding to clouds, R_2/R_1 for contrast test over water and land, R_{3A}/R_{3B} for albedo test over water and land, the $BT_{3B} - BT_5$ for cirrus detection test, $BT_3 - BT_5$ for uniform low stratus test, $BT_4 - BT_5$ for thin (large positive) and thick cloud test (near zero or negative difference), R_{3B} for opaque (below 1) and transparent (above 1) cloud test. Although seems better than Apollo scheme, the use of limited bands is still seen as the main limitation.

(c) SCANDIA and EUMETSAT scheme

The SCANDIA scheme (Karlsson 1989) is similar to the CLAVR and Apollo schemes in many ways, i.e. involved applying sequences of threshold tests using NOAA AVHRR bands. SCANDIA extra feature is that it groups a series of tests together rather apart from applying the individual threshold tests, i.e. the identification of a cloud pixel requires several threshold tests must be passed.

The EUMETSAT scheme (Dybbroe et al. 2005) is a more sophisticated scheme compared to SCANDIA, CLAVR and Apollo. The sheme uses dynamic thresholds that separate fully cloudy or cloud-contaminated from cloud-free pixels. The thresholds take into account the actual state of the atmosphere and surface and the sun–satellite viewing geometry using cloud-free radiative transfer model simulations. Cloud detection is done using sequences of grouped threshold tests that employ both spectral and textural features. Cloudy pixels are further divided into 10 different categories: 5 opaque cloud types, 4 semitransparent clouds, and 1 subpixel cloud category. However, the scheme does not use AVHRR band 2, which may be considered as a weakness.

(d) ISCCP scheme

The ISCCP scheme is the first project of the World Climate Research Programme and uses measurements from visible (0.65 ± 0.15 μm) and thermal infrared (11 ± 1 μm) wavelengths to detect cloud (Rossow and Garder 1993; Rossow and Schiffer 1999). The measurements are either from AVHRR GAC data or from the data obtained from geostationary satellites. Besides the conventional spectral-based approach, the ISCCP scheme also uses a temporal-based approach to separate cloudy and clear-sky pixels. In ISCCP, a cloud classification scheme based on height, pressure and optical thickness was introduced.

The major steps in the ISCCP scheme are:

(1) the gross spatial thermal contrast test – classifies pixels as cloudy if they are much colder than other pixels within a limited spatial domain,

(2) the gross temporal thermal contrast test – applied to a sequence of images over a 3-day interval and classifies a pixel as cloudy if it has sharply lower IR radiance compared to a day earlier or later,

(3) the generation of spatiotemporal statistics for both thermal and visible bands – conducted over 5-day time intervals,

(4) the identification of clear-sky thresholds using the results of the previous step and

(5) the classification of pixels into three categories: clear, cloudy, and marginally cloudy using the derived thresholds – the pixel is placed into the clear-sky (cloudy) category if visible and IR radiances pass the clear-sky (cloudy)

thresholds. If the radiances fall in between, the pixel is assigned to the marginally cloudy category.

The main weakness of the ISCCP is the use of only visible and thermal infrared wavelengths in detecting cloud, which may miss clouds detectable from other wavelength regions, e.g. near and middle infrared.

(e) MODIS scheme

The MODIS scheme can be regarded as the most comprehensive cloud detection scheme in terms of the number of spectral bands used, i.e. 22 out of 36 MODIS bands (i.e. in visible, near infrared and thermal infrared wavelengths) to maximise the cloud detection capability (Ackerman et al. 1998; Ackerman et al. 2010). Apart from the spectral information, it also uses other ancillary input such as topography and geometry of observation for each 1-km pixel, land/water and ecosystem maps, and daily operational snow/ice data products from the NOAA and National Snow and Ice Data Center. The MODIS cloud mask is a 48-bit cloud mask with flags specify the confidence level of clear-sky detection (confident cloudy, uncertain, probably clear, and confident clear), while other flags indicate high cloud type, shadow, thin cirrus, snow/ice, sun glint, and results from the other tests, including the 16 values of the cloud flags for all 250 m x 250 m sub-pixels within the 1 km x 1 km field of view. A cloud test may use a single band, ratio of bands or difference of bands. Each test returns a confidence level that a pixel is clear, ranging in value from 1 (high) to 0 (low). The tests are grouped into five categories based on their capability to detect similar cloud types: thick high clouds, thin cloud, low clouds, high thin cloud and high thin cirrus cloud. For a group, its confidence indicator is the smallest confidence level for the individual tests within that group. Other important criterion of the MODIS scheme is the inclusion of algorithm to detect cloud shadows. The Cloud shadow detection implemented in MODIS uses the spectral (not geometrical) approach and checks for cloud shadows once a confident clear-sky pixel is found. Cloud shadow is detected if reflectance in the 0.94-μm band (band 19) is less than 0.07, the ratio of reflectances at 0.87 and 0.66 μm (bands 2 and 1) are greater than 0.3, and the reflectance in the 1.2-μm band (band 5) is less than 0.2 (Ackerman et al 1998; Ackerman et al. 2006). The cloud mask was validated by using image interpretation and quantitative analysis. The former used visual inspection of the spectral and spatial features in a set of composite images, while the latter used pixel-to-pixel comparison with ground instruments or platform-based observations, which both show a good agreement with the cloud mask.

The MODIS cloud mask can be downloaded from the MODIS website (http://modis.gsfc.nasa.gov). For detecting cloud in daytime over land, only seven bits are involved:

(1) bit 14; BT_{35},
(2) bit 15; BT_{27}
(3) bit 16; R_{26} ,
(4) bit 18; $BT_{31} - BT_{32}$,
(5) bit 19; $BT_{31} - BT_{22}$,
(6) bit 20; R_1,

These MODIS cloud tests are divided into:

(1) Group 1; detection of thick high cloud using bits 14 and 15.
(2) Group 2; detection of thin cloud missed by Group 1 tests using bits 18 and 19.
(3) Group 3; detection of low cloud using bits 20.
(4) Group 4; detection of thin high cloud using bit 16.

For each test, a confidence level between 0 and 1 is assigned, where 0 represents high confidence of a cloudy condition and 1 represents high confidence of a clear condition. For a group, its confidence indicator is the minimum confidence level for the individual tests within that group, i.e.:

$$G_{i=1..5} = \min [F_i] \qquad \text{... (2.1)}$$

The final cloud mask confidence, Q, is a product of all individual tests:

$$Q = \prod_{i=1}^{N} F_i \qquad \text{... (2.2)}$$

If any test gives high confidence of a cloudy condition ($F_i = 0$), then the final cloud mask will indicate cloud ($Q = 0$).

In the latest version of MODIS cloud mask, the thresholds for day-time cloud detection over land is summarised in Table 2.2 (Ackerman et al. 2010).

Table 2.2: *Cloud criteria , test, its function and the threshold used in the MODIS cloud mask for day-time detection over land (Ackerman et al. 2010).*

Cloud criteria (see Table 2.1)	Test	Description	Threshold
Thick high clouds	BT_{35}	CO_2 slicing. Values smaller than threshold indicate ice cloud at middle and upper atmosphere	226 K
	BT_{27}	Values smaller than threshold indicate water low clouds	225 K
Thin high clouds	$BT_{31} - BT_{32}$	Values smaller than threshold indicate high cloud or cirrus cloud	2 K
Thick low clouds	$BT_{31} - BT_{22}$	Values smaller than threshold indicate low level water clouds	-11.0 K
Low clouds	R_1	Reflectance gross cloud test with vegetated land background. Values larger than threshold indicate cloud.	0.14
Thin high clouds	R_{26}	Values larger than threshold indicate thin cirrus cloud	0.03

(f) ACCA Scheme

ACCA is an automatic cloud cover assessment algorithm, developed in early 1980s for TM (Thematic Mapper) onboard Landsat 4 and 5 (Irish et al 2000). The first version of TM ACCA algorithm uses a single pass process that employs Bands 3, 5 and 6 radiance thresholds to detect cloudy pixels. The second version, the Landsat 7 ACCA algorithm uses five of eight ETM+ bands:

- Band 2 reflectance (R_2): 0.53 to 0.61 μm – green; 30 m resolution
- Band 3 reflectance (R_3): 0.63 to 0.69 μm – red; 30 m resolution
- Band 4 reflectance (R_4): 0.78 to 0.90 μm – near infrared; 30 m resolution

- Band 5 reflectance (R_5): 1.55 to 1.75 μm – middle infrared; 30 m resolution
- Band 6 brightness temperature (BT_6): 10.4 to 12.5 μm – thermal infrared; 60 m resolution.

Landsat 7 ACCA algorithm involves two passes. For pass one processing, the eight tests involved are:

- Filter 1: $0.08 > R_3$
- Filter 2: Normalized Snow Difference Index = $\frac{R_2 - R_5}{R_2 + R_5} < 0.7$
- Filter 3: $BT_6 < 300$ K
- Filter 4: $(1 - R_5) * R_6 < 225$
- Filter 5: $R_4/R_3 < 2$
- Filter 6: $R_4/R_2 < 2$
- Filter 7: $R_4/R_5 > 1$
- Filter 8: $R_5/R_6 > 210$ (warm clouds); $R_5/R_6 < 210$ (cold clouds)

Pixels that passed filter 1 through 7 are classified as clouds; Filter 8 further classifies the cloud pixels into warm or cold clouds. Pass two processing involves thermal analysis using band 6 exclusively, in which a thermal cloud signature is developed from the product of pass one and used to identify the remaining clouds in a scene. Finally, the last step involves processing the cloud mask for ambiguous pixels. Each non-cloud image pixel is examined and converted to cloud if at least 5 of its 8 neighbours are clouds.

From the analysis above, it is clear that MODIS cloud mask is the most comprehensive scheme, so will be adopted in our study to learn the spectral properties of cloud and then to detect cloud within satellite data. Subsequently, we will apply the MODIS analysis to Landsat data and finally its performance will be compared with the ACCA scheme.

Cloud shadow is a by-product of cloud that results from the projection of cloud and can cause a substantial impact to satellite data. In visible wavelengths, if undetected, it is likely to be classified as other classes; dark cloud shadows possess spectral

UNIVEI
OF SHEI

properties quite similar to those of water, while lighter shadows can be easily confused with dark vegetation. Therefore, cloud shadow needs to be detected and masked from remote sensing data before performing further processing. Ackerman et al. (2006) proposed a cloud shadow detection procedure based on spectral analysis of MODIS data. They proposed that cloud shadow can be indicated by R_{19} smaller than 0.07, R_2/R_1 larger than 0.3 and R_5 smaller than 0.2. The results were visually analysed and was sensibly matched with the location of the shadow. Later, Luo et al. (2008) proposed a method for detecting cloud shadow on MODIS data based on $Max(R_2,R_6)/R_3$ less than 1.5, R_1 less than 0.12, R_2 less than 0.24 and R_6 less than 0.24. They claimed that by using the method, most of the shadow pixels can be successfully removed from the data. Due to the simplicity and effectiveness, our study will make use of the Ackerman et al. (2006) method to remove cloud shadow; subsequently comparison with the Luo et al. (2008) method will be carried out.

2.5 Datasets and Methods

2.5.1 The MODIS Satellite

The MODIS instrument is the primary payload attached to two satellites, Terra and Aqua. Terra (Figure 2.7) was launched on December 18, 1999, and Aqua on May 4, 2002.

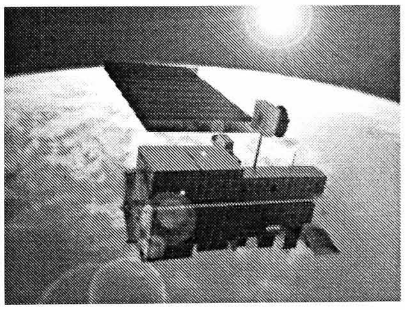


Figure 2.7: *Terra satellite (MODIS 2007).*

Table 2.3: *MODIS instrument specifications (MODIS 2007).*

Ownership:	National Aeronautics and Space Administrations (NASA), USA
Orbit:	705 km, 10:30 a.m. descending node (Terra) or 1:30 p.m. ascending node (Aqua), sun-synchronous, near-polar,

	circular
Scan Rate:	20.3 rpm, cross track
Swath Dimensions:	2330 km (cross track) by 10 km (along track at nadir)
Telescope:	17.78 cm diameter off-axis, afocal (collimated), with intermediate field stop
Size:	1.0 x 1.6 x 1.0 m
Weight:	228.7 kg
Power:	162.5 W (single orbit average)
Data Rate:	10.6 Mbps (peak daytime); 6.1 Mbps (orbital average)
Quantization:	12 bits
Spatial Resolution:	250 m (bands 1-2) 500 m (bands 3-7) 1000 m (bands 8-36)
Design Life:	6 years

MODIS instrument specifications are shown in Table 2.3. The main advantage of MODIS data is that it offers a wide range of spectral bands. There are 36 spectral bands covering the visible, near infrared and thermal infrared ranges of the electromagnetic spectrum. The primary use and the corresponding spectral information for all bands are summarised in Table 2.4.

Table 2.4: *Primary use and spectral information for MODIS bands (MODIS 2007).*

Primary Use	Band	Band Range[1]	Bandwidth[2]	Spectral Radiance[3]	Central Wavelength[4]
Land/Cloud/Aerosols Boundaries	1	0.620 – 0.670	41.8	21.8	0.659
	2	0.841 -0. 876	39.4	24.7	0.865
Land/Cloud/Aerosols Properties	3	0.459 – 0.479	17.6	35.3	0.470
	4	0.545 – 0.565	19.7	29.0	0.555
	5	1.230 – 1.250	24.5	5.4	1.240
	6	1.628 – 1.652	29.7	7.3	1.640
	7	2.105 – 2.155	52.9	1.0	2.130
Ocean Colour/ Phytoplankton/ Biogeochemistry	8	0.405 – 0.420	11.8	44.9	0.415
	9	0.438 – 0.448	9.7	41.9	0.443
	10	0.483 – 0.493	10.6	32.1	0.490
	11	0.526 – 0.536	11.8	27.9	0.531
	12	0.546 – 0.556	10.4	21.0	0.565
	13	0.662 – 0.672	10.1	9.5	0.653
	14	0.673 – 0.683	11.4	8.7	0.681
	15	0.743 – 0.753	10.0	10.2	0.750
	16	0.862 – 0.877	15.5	6.2	0.865
Atmospheric Water Vapour	17	0.890 – 0.920	35.7	10.0	0.905
	18	0.931 – 0.941	13.7	3.6	0.936
	19	0.915 – 0.965	46.3	15.0	0.940
Surface/Cloud Temperature	20	3.660 - 3.840	36.4	0.45(300K)	3.750
	21	3.929 - 3.989	182.6	2.38(335K)	3.959
	22	3.929 - 3.989	85.7	0.67(300K)	3.959
	23	4.020 - 4.080	88.2	0.79(300K)	4.050
Atmospheric	24	4.433 - 4.498	87.8	0.17(250K)	4.465
	25	4.482 - 4.549	93.7	0.59(275K)	4.515

Temperature					
Cirrus Clouds	26	1.360 - 1.390	94.3	6.00	1.375
Water Vapour	27	6.535 - 6.895	254.6	1.16(240K)	6.715
	28	7.175 - 7.475	325.3	2.18(250K)	7.325
Cloud Properties	29	8.400 - 8.700	369.2	9.58(300K)	8.550
Ozone	30	9.580 - 9.880	300.6	3.69(250K)	9.730
Surface/Cloud	31	10.780 - 11.280	510.3	9.55(300K)	11.030
Temperature	32	11.770 - 12.270	493.5	8.94(300K)	12.020
Cloud Top Altitude	33	13.185 - 13.485	13.335	4.52(260K)	13.335
	34	13.485 - 13.785	13.635	3.76(250K)	13.635
	35	13.785 - 14.085	13.935	3.11(240K)	13.935
	36	14.085 - 14.385	14.235	2.08(220K)	14.235

[1] Bands 1 to 36 are in μm

[2] Bandwidth values are in nm

[3] Spectral radiance values are in Wm^{-2} μm^{-1} sr^{-1}

[4] Central wavelength values are in μm

MODIS Level 1B (MODIS L1B) are the main data used in this study. There are four product files in the MODIS L1B product, summarised in Table 2.5 (MODIS Characterization Support Team 2006).

Table 2.5: *Summary of MODIS L1B products (MODIS Characterization Support Team 2006).*

Product Type		Product Content
MODIS/Terra	MODIS/Aqua	
MOD02QKM	MYD02QKM	Calibrated Earth View data at 250 m resolution
MOD02HKM	MYD02HKM	Calibrated Earth View data at 500 m resolution, including the 250 m resolution bands aggregated to 500 m resolution.
MOD021KM	MYD021KM	Calibrated Earth View data at 1 km resolution, including the 250 m and 500 m resolution bands aggregated to 1 km resolution.
MOD02OBC	MYD02OBC	On Board Calibrator (OBC) and Engineering Data

In this study, the MOD021KM product from MODIS Terra is used. These datasets were downloaded from the Level 1 and Atmosphere Archive and Distribution System (LAADS) website (NASA 2007). MOD021KM contains data in three forms: (1) Radiance (W $m^{-2}\mu m^{-1}sr^{-1}$) for the reflective bands (2) Radiance (W m^{-2} $\mu m^{-1}sr^{-1}$) for the emissive bands; and (3) Reflectance (dimensionless) for the reflective bands.

The relationship between the TOA reflectance, ρ and TOA radiance, L at the isotropic surface can be expressed as:

$$L = \frac{E_{\lambda}\mu_{s}\rho}{\pi} \qquad \text{... (2.3)}$$

where E_{λ} is the mean exoatmospheric solar irradiance at TOA (W m^{-2} μm^{-1}), μ_s is $\cos(\theta_s)$ and π is a constant equal to ~3.14159 (unitless); θ_s is the solar zenith angle.

The MODIS Level 1B also contains thermal data, which are recorded as TOA radiance, and can be converted to brightness temperature using the Planck function. Brightness temperature is defined as the temperature for an ideal black body with the observed radiance; it is the temperature a blackbody needs to have to emit radiation of the observed intensity at a given wavelength. From Planck's Law, the observed radiance is expressed as

$$L = \frac{2hc^{2}\lambda^{-5}}{\left(e^{\frac{hc}{k\lambda T}} - 1\right)} \qquad \text{... (2.4)}$$

where

L = radiance ($Wm^{-2}\mu m^{-1}sr^{-1}$)

h = Planck's constant (Js) = 6.626 x 10^{-34} Js

c = speed of light in vacuum (ms^{-1}) = 3 x 10^{8} ms^{-1}

k = Boltzmann gas constant (JK^{-1}) = 1.3806503 × 10^{-23} JK^{-1}

λ = band or detector centre wavelength (μm)

T = brightness temperature (K)

By inverting this formula, we can solve for brightness temperature, T:

$$T = \left(\frac{hc}{k\lambda}\right)\frac{1}{\ln\left(2hc^{2}\lambda^{-5}L^{-1} + 1\right)} \qquad \text{... (2.5)}$$

In a simpler form,

$$T = \left(\frac{c_2}{\lambda}\right) \cdot \frac{1}{\ln\left(\frac{c_1}{\lambda^5 L} + 1\right)} \qquad \text{... (2.6)}$$

where $c_2 = \frac{hc}{k} = 1.438 \times 10^4$ and $c_1 = 2hc^2 = 1.191 \times 10^8$.

In practice, this conversion can be carried out using built-in tools in image processing software, such as ENVI.

2.5.2 Methodology

The study area is Peninsular Malaysia, located within latitude 6°47' N, longitude 88°25' E (upper left), and latitude 1°21' N, longitude 106°20' E (lower right) as shown in Figure 2.8 that covers an area of about 140000 km^2.

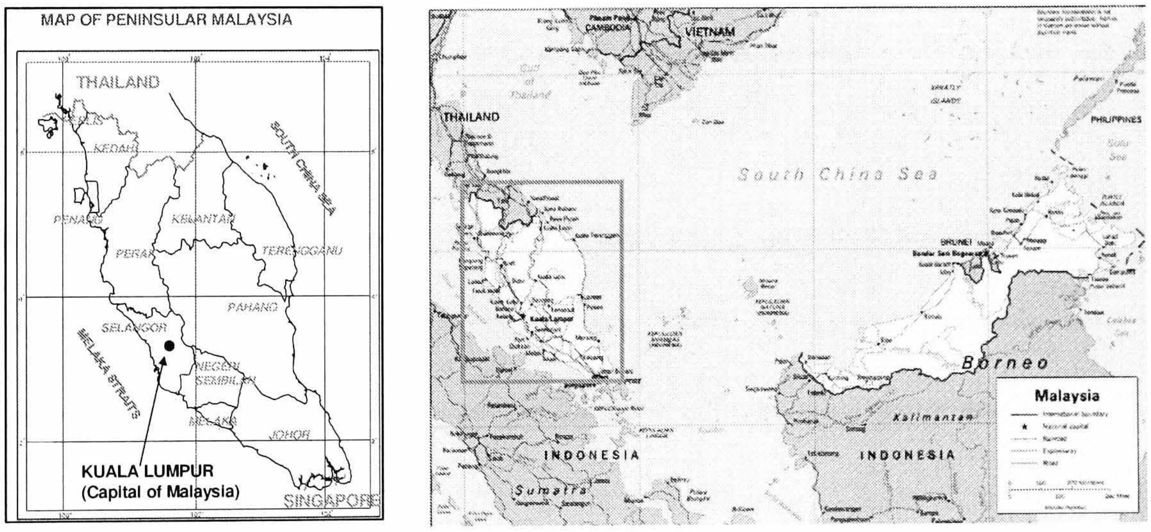
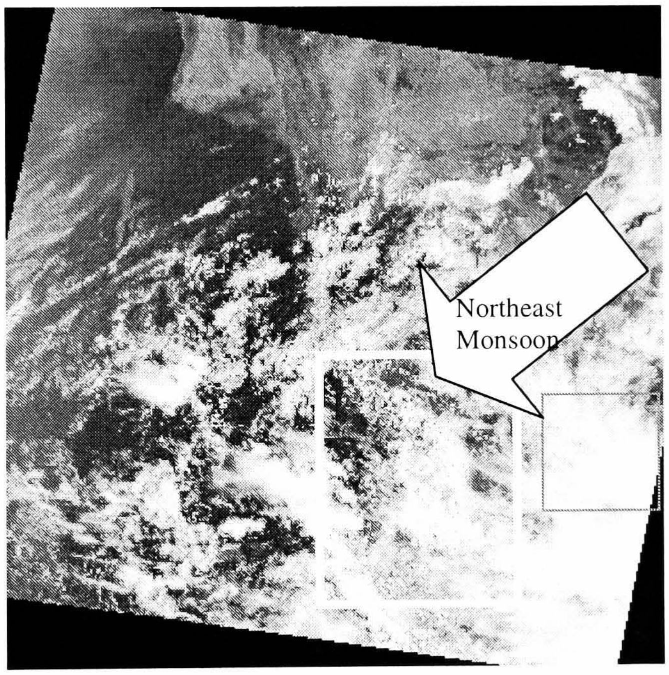


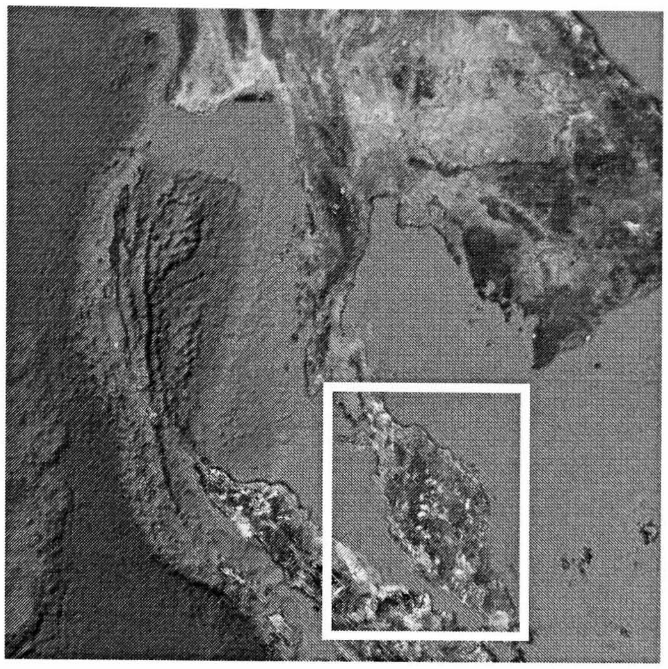
Figure 2.8: *Map of Peninsular Malaysia.*

A MODIS Terra dataset (i.e. MOD021KM.A2004030.0355) recorded on the 30 January 2004 at 03:55 UTC (11:55 a.m. local time; sun elevation angle 59.2°) was used because it was haze-free due to the Northeast Monsoon, which occurs from November to March every year. In Malaysia, the highest rain amount, which is associated with the high cloud amount, occurs during this period. Figure 2.9(a) shows bands 1, 4, and 3 of the MODIS dataset assigned to red, green and blue channels respectively; Malaysia is indicated by the area within the yellow box and Northeast Monsoon is demarcated by the arrow. For convenience, the same location from Google Map is shown in Figure 2.9(b). By comparing both images, it can be revealed that most parts of Malaysia are covered by clouds.

We will first carry out visual analysis of cloud to identify cloud pixels from the MODIS dataset. We will then carry out spectral analysis of cloud from both reflective and thermal MODIS bands; cloud and its shadow detection and masking will then be performed using the MODIS scheme on two MODIS datasets, i.e. 30 January 2004 and 15 February 2004. Subsequently, cloud detection and masking based on multitemporal basis are carried out for 2004 and 2005 to see cloud trends throughout these years. The MODIS analysis will later be applied to Landsat data, focusing on Klang district in the state of Selangor Malaysia, for use in later chapters.



(a)

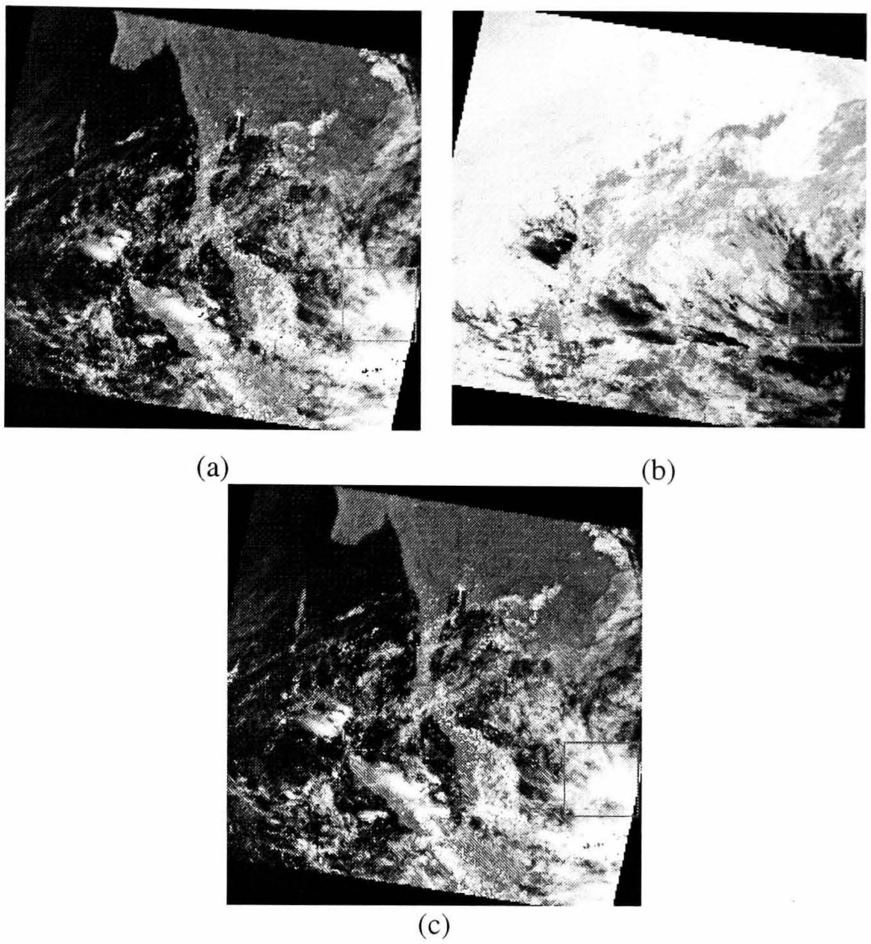


(b)

Figure 2.9 : *(a) MODIS Terra data dated 30 January 2004 (03:55 UTC); Malaysia indicated by the yellow box and the arrow is the Northeast Monsoon and (b) the same location from Google Map.*

Visual Analysis of Cloud from MODIS Data

Initially, we carried out visual analysis on MODIS Terra dataset dated 30 January 2004 using individual bands, which all 36 bands were individually displayed. For each of the 20 reflective bands (1 to 19 and 26), bright features (high reflectance), which were suspected to be cloud, were visually extracted. For the 16 thermal bands (20 to 25 and 27 to 36), the same procedure was carried out for dark features (low temperature). Figure 2.10 shows (a) MODIS band 2 and (b) band 31, in which the bright regions in the former correspond to very high reflectance resulting from the high scattering efficiency of cloud droplets, while the dark regions in the latter correspond to the very low brightness temperature of cloud. In Figure 2.10(c), bands 1, 2 and 3 displayed simultaneously as a colour composite image in order to enhance the difference between clouds and other features. With such a combination, cloud tends to appear as white, since it has high reflectance in these visible wavelength regions; this helps to 'double check' the first approach.



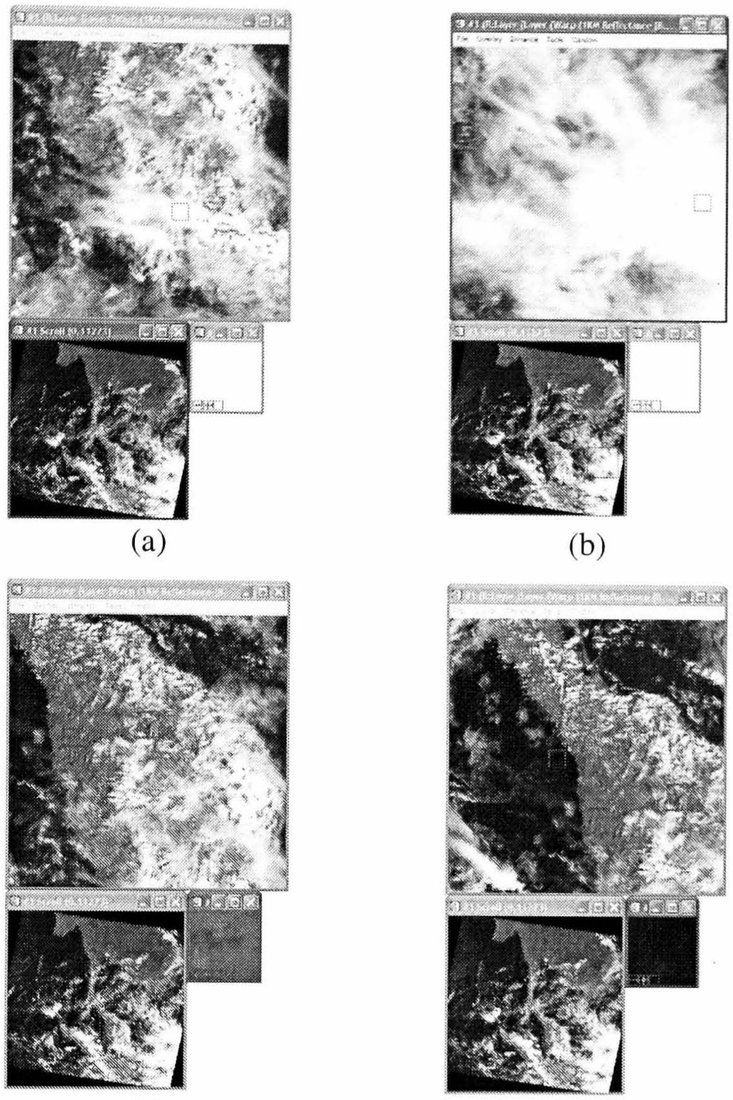
(a) (b)

(c)

Figure 2.10 : *MODIS Terra data for 30 January 2004: (a) MODIS band 2, (b) band 31 and (c) bands 3, 2 and 1 assigned to red, green and blue.*

Spectral Analysis

The visual analysis approach was used to sample 100 x 100 blocks of cloud over land and ocean pixels from the dataset. The locations of the sampling areas are shown in Figure 2.11(a) cloud over land, (b) cloud over ocean, (c) land and (d) ocean pixels. The image on the lower left is the full scene of MODIS Terra bands 3, 2 and 1 assigned to red, green and blue channels respectively from 30 January 2004; the top and the lower right images are the enlarged versions of the red box in the lower left and the top images respectively.

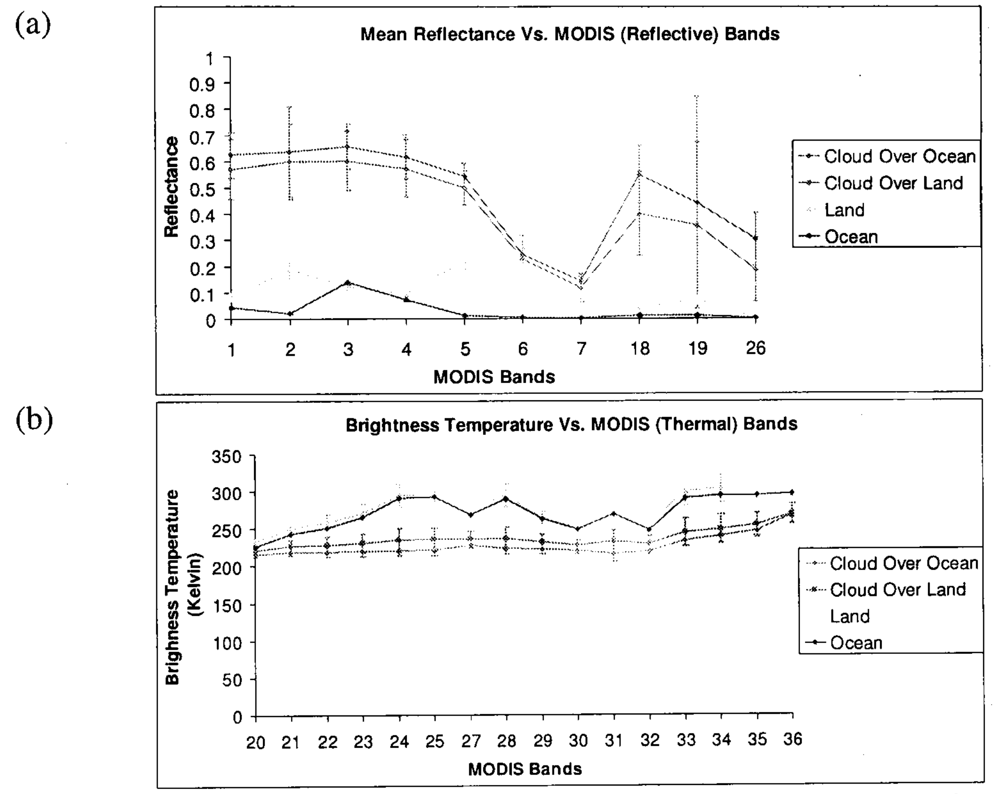


(a) (b)

(c) (d)

Figure 2.11: *Sampling for (a) cloud over land, (b) cloud over ocean, (c) land and (d) ocean pixels. The image on the lower left is the full scene of MODIS Terra bands 3, 2 and 1 assigned to red, green and blue channels respectively from 30 January 2004; the top and the lower right images are the enlarged versions of the red box in the lower left and the top images respectively.*

The reflectance curves for the reflective MODIS bands are shown in Figure 2.12(a). The negative reflectances for cloud in bands 8 to 17 are caused by saturation problems and have been omitted. For the remaining bands, cloud over land has lower reflectances because it tends to be thinner than cloud over ocean. Brightness temperature curves for the thermal MODIS bands are shown in Figure 2.12(b). These have the opposite trend to reflectance, with the brightness temperature of cloud over land being higher than cloud over the ocean. This is due to the fact that cloud over ocean is colder because it tends to be thicker than cloud over land (Ackerman et al. 2010). Much larger standard deviations in reflectance and brightness temperature are observed for clouds over the land than ocean due to the larger variations in surface reflectivity and emissivity respectively (Ackerman et al. 2010). Land has much lower reflectances than cloud due to the much less reflective surface properties and lower altitudes. Land has higher reflectances and brightness temperatures than ocean due to the lesser energy absorption and higher temperature respectively. Ocean has lower standard deviations in reflectance and brightness temperature due to the much uniform spectral properties.


(a)

Mean Reflectance Vs. MODIS (Reflective) Bands

(b)

Brightness Temperature Vs. MODIS (Thermal) Bands

Figure 2.12: *(a) Reflectance of cloud over the ocean and cloud over the land relative to non-cloud features. Vertical bars indicate standard deviations; (b) Same as (a) but for brightness temperature.*

We subsetted Malaysia from the full scene of MODIS dataset and masked the sea in white. Since this study focuses on land studies, the term 'cloud' in the following sections means cloud over the land. Based on MODIS scheme in Table 2.2, cloud detection is carried out using single reflective bands and thermal bands and brightness temperature differences.

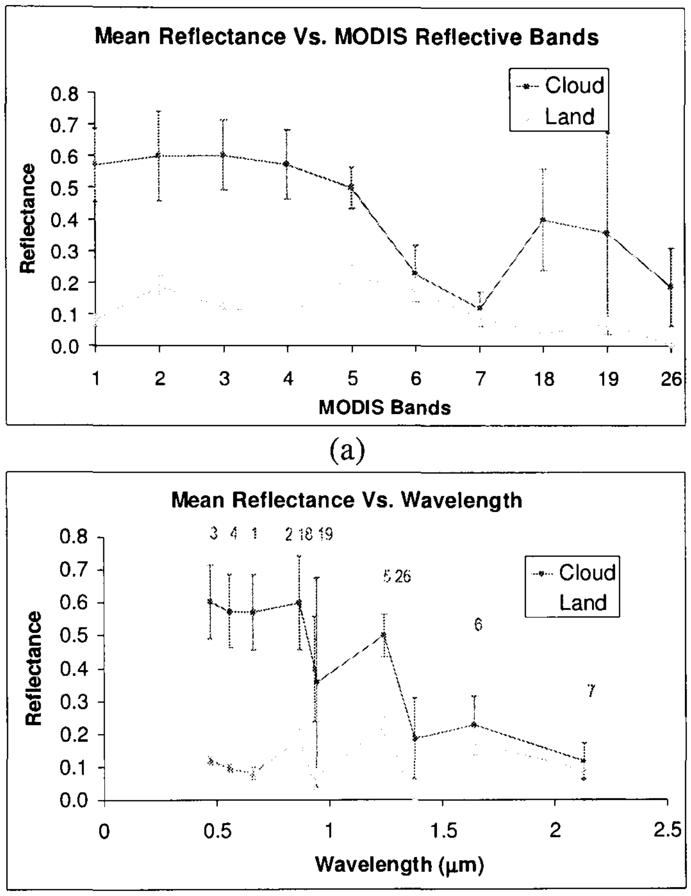
Cloud Detection Using Single Reflective Bands

TOA reflectance curves for cloud, land and ocean for all 20 MODIS reflective bands are plotted in Figure 2.13, and the mean reflectance for cloud, ocean and land against the MODIS bands and wavelengths, are shown in Table 2.6. Bands with negative values, due to saturation, have been omitted.

Table 2.6: *Mean reflectance for the MODIS reflective bands for 30 January 2004. R_k is reflectance for band k.*

30 January 2004			
MODIS Band (R_k)	Centre Wavelength (μm)	Mean R_k (dimensionless)	
		Cloud	Land
1 (R_1)	0.659	0.571	0.081
2 (R_2)	0.865	0.600	0.186
3 (R_3)	0.470	0.603	0.122
4 (R_4)	0.555	0.574	0.097
5 (R_5)	1.240	0.500	0.222
6 (R_6)	1.640	0.228	0.175
7 (R_7)	2.130	0.117	0.091
17 (R_{17})	0.905	0.095	0.139
18 (R_{18})	0.936	0.399	0.040
19 (R_{19})	0.940	0.357	0.072
26 (R_{26})	1.375	0.187	0.002

After removing all negative reflectance values (i.e. due to saturation) from the data, meaningful trends of spectral reflectance for cloud and land were revealed, as shown in Figure 2.13(a). Cloud exhibits much higher reflectance than land or ocean for bands 1 to 5, 18 and 19, but low reflectance values for bands 6, 7 and 26 (with a decreasing trend towards longer wavelengths). Figure 2.13(b) shows the reflectance plotted against wavelength with the corresponding band numbers given in red fonts, showing that cloud and land have distinctive spectral reflectance signatures. Both cloud and land exhibit a fluctuating trend. As most of the land is covered by vegetation, strong chlorophyll absorption occurs at wavelengths of 0.46 and 0.66 μm, which are often called the chlorophyll absorption band (Swain and Davis, 1978; Lillesand et al. 2004). Land reflectances increase from 0.66 μm to 0.86 μm because in this wavelength region, leaves typically reflect 40% to 50% of the incident energy respectively due to their internal structure (Lillesand et al. 2004). As all features contain water, the reflectance curves show absorption in the water absorption bands near wavelengths of 1.4 μm.



(a)

(b)

Figure 2.13: *Cloud spectral signature using reflectance data for 30 January 2004: (a) Plot of mean cloud reflectance without bands 8 – 17, and (b) Same as (a) but in term of wavelength to form the spectral signature of cloud and land, with MODIS band number in red font. Vertical bars indicate standard deviations.*

To separate between cloud and non-cloud, a threshold value from the MODIS cloud mask was used and the cloud masking results for R_1 and R_{26} reflectance tests are shown in Figure 2.14((a) and (b)). The raw data and masked data are shown in left and middle column, while the corresponding histogram, on the right column. For R_1 test, Figure 2.14(a(left)) clearly shows bright patches of opaque clouds in the east and south of Malaysia, while transparent clouds can be seen surrounding the opaque clouds. In Figure 2.14(a(right)), pixels with reflectance larger than the threshold were labelled as cloud and masked red. Pixels detected as cloud by R_1 test can be seen distributed throughout almost the whole Malaysia. These are low clouds, i.e. stratus, stratocumulus, cumulonimbus, cumulus and nimbostratus (see Table 2.1 and Table 2.2). The effectiveness of this test is due to the surface types, i.e. mainly vegetations, which posses much lower reflectance in 0.66 μm wavelength measurement; therefore separation between the cloud and cloud-free pixels can be done easily. For R_{26} test, Figure 2.14(b(left)) shows a much brighter but smaller cloud patches in the south and east of Malaysia and transparent clouds in between them. The Earth surface seems very dark due to the very low surface reflectance measured at 1.38 μm wavelength (i.e. near infrared), resulting in a high contrast between the clouds and their background. In Figure 2.14(b(right)), when mask is applied, more cloud pixels can be observed in the middle towards the north; clouds detected by this test are thin high clouds, i.e. cirrus, cirrostratus and cirrocumulus (see Table 2.1 and Table 2.2).

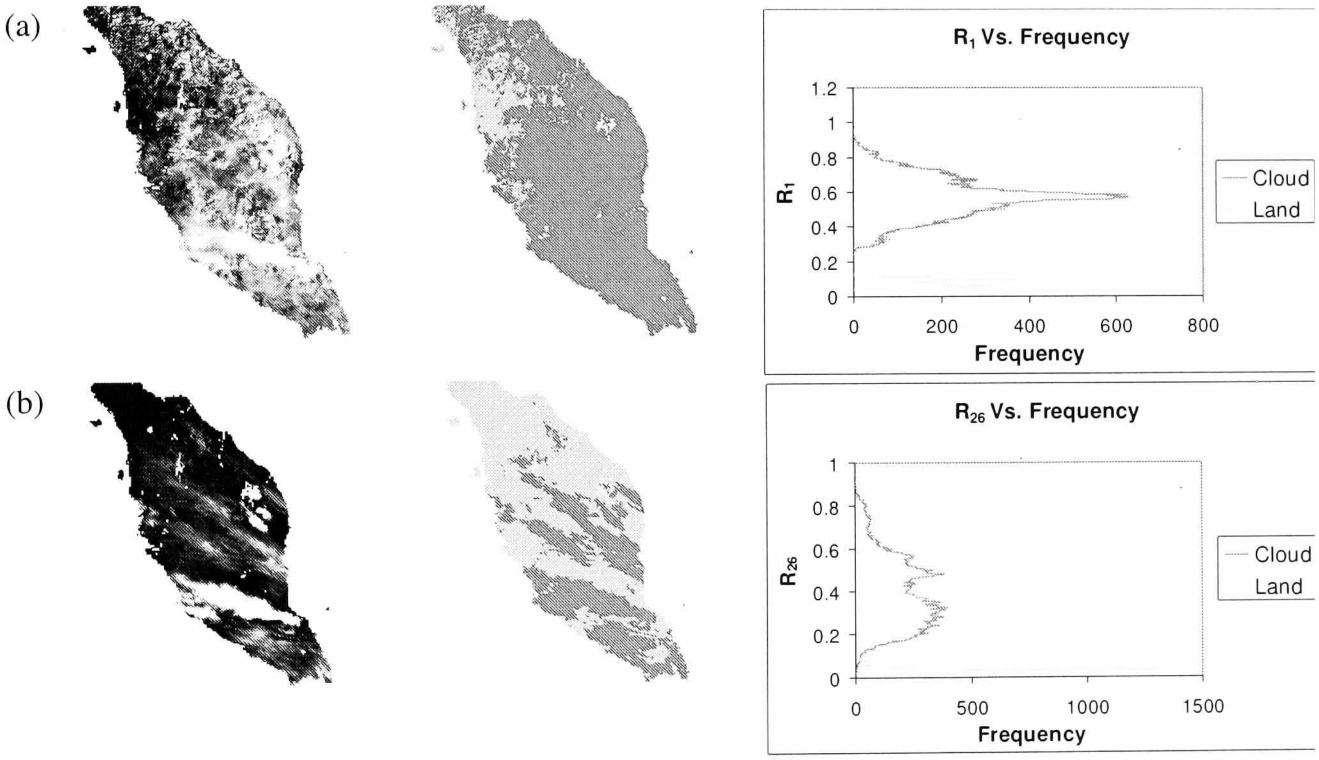


Figure 2.14: *(a) R_1 and (b) R_{26} test for 30 January 2004 before (left) and after (middle) applying the thresholds with the cloud pixels masked in red, and the corresponding histogram for cloud and land (right). Cloud-free and water body pixels are masked grey and white respectively.*

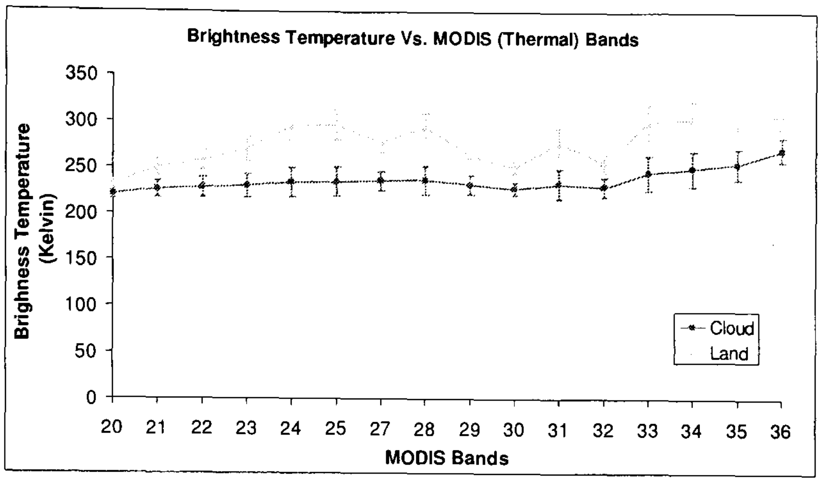
Cloud Detection Using Brightness Temperatures and Brightness Temperature Difference

Conversion from radiance to brightness temperature was carried out for all 16 thermal bands using Equation 1.5. The mean brightness temperatures values for each of the thermal MODIS bands are shown in Table 2.7.

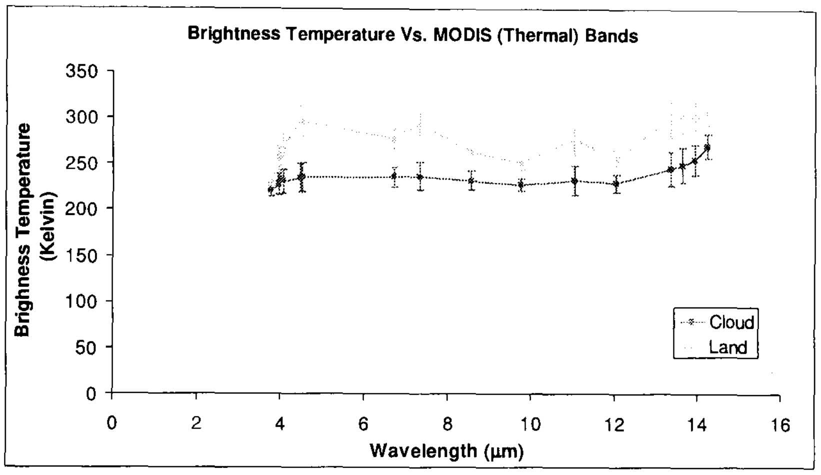
Table 2.7: *Cloud and land mean brightness temperature from MODIS thermal bands for 30 January 2004.*

30 January 2004			
MODIS Band (BT_k)	Wavelength (μm)	Mean BT_k (K)	
		Cloud	Land
20 (BT_{20})	3.750	269.964	306.490
21 (BT_{21})	3.959	254.896	304.237
22 (BT_{22})	3.959	249.560	303.843
23 (BT_{23})	4.050	245.129	299.746
24 (BT_{24})	4.465	229.749	256.047
25 (BT_{25})	4.515	232.711	277.244
27 (BT_{27})	6.715	227.489	250.893
28 (BT_{28})	7.325	232.098	263.732
29 (BT_{29})	8.550	237.042	294.706
30 (BT_{30})	9.730	235.807	277.727
31 (BT_{31})	11.030	235.607	296.679
32 (BT_{32})	12.020	234.682	294.901
33 (BT_{33})	13.335	230.818	270.227
34 (BT_{34})	13.635	228.322	258.166
35 (BT_{35})	13.935	226.383	249.858
36 (BT_{36})	14.235	220.722	231.596

Curves of brightness temperature for the thermal MODIS bands for cloud and land against band number and wavelength are shown Figure 2.15(a) and (b) respectively. Each point in (b) corresponds to that of (a) consecutively. Land exhibits a brightness temperature ranging from approximately 231 to 306 K shows a sharp increase between 4 and 5 μm, then a fluctuating trend at longer wavelengths. The cloud brightness temperature is nearly constant in the lower-numbered bands but increases at longer wavelengths.



(a)



(b)

Figure 2.15 : *Brightness temperature for cloud and land for MODIS Terra data dated 30 January 2004 plotted against (a) MODIS bands and (b) wavelength. Here, the points in (b) consecutively correspond to those in (a). Vertical bars indicate standard deviations.*

Brightness temperature tests using BT_{27} and BT_{35} were applied to the MODIS dataset from 30 January 2004. Figure 2.16 shows (a) BT_{27} and (b) BT_{35} tests for 30 January 2004 before (left) and after (right) applying the thresholds with the cloud pixels masked in red; cloud-free and water body pixels are masked grey and white respectively. Figure 2.16(a(left)) shows dark patches of cloud by BT_{27} in the south of Malaysia and much smaller patches can be seen in the east of Malaysia. In Figure 2.16(a(right)), the red masks are located about the same place where the black patches are found – almost no cloud is found elsewhere. Quite similar outcomes are shown by

BT_{35} Figure 2.16(b) due to the quite similar spectral response to cloud (Figure 2.15); both tests are sensitive to thick high clouds, e.g. cumulonimbus.

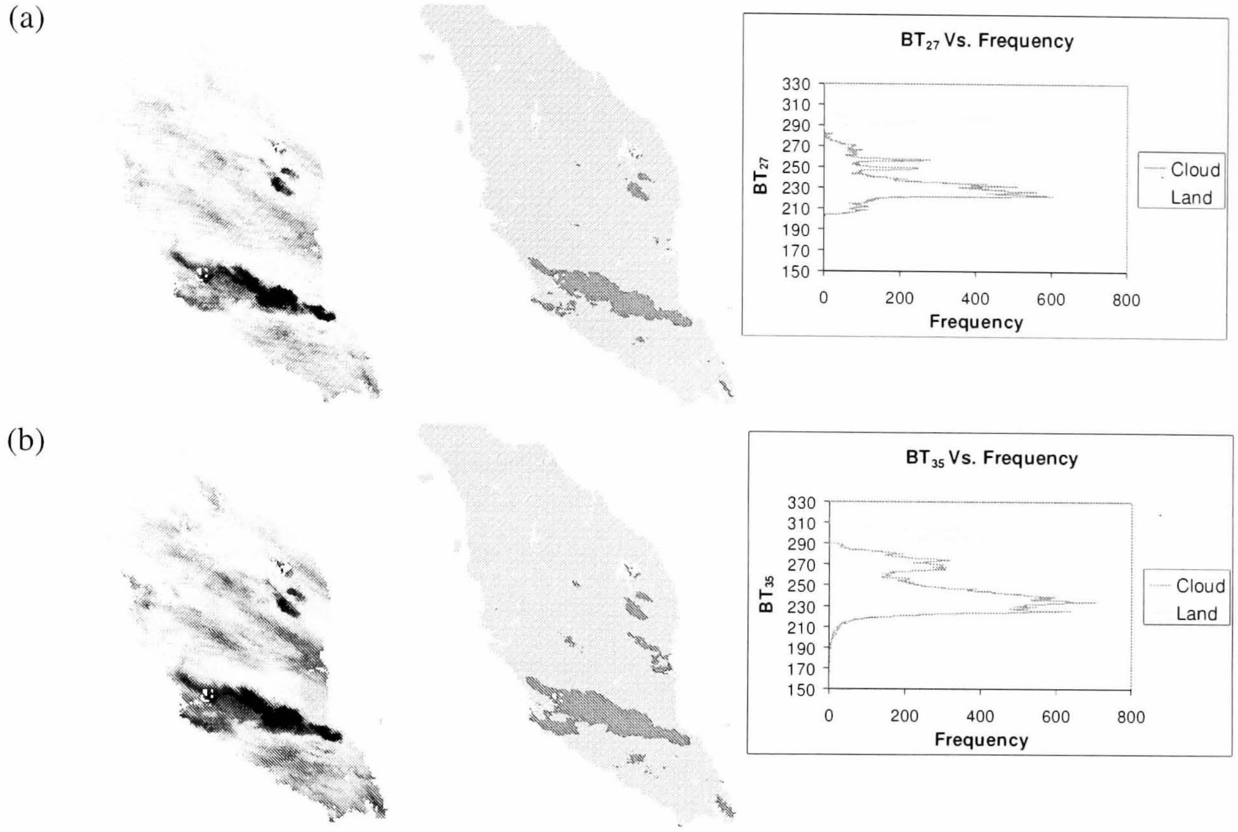


Figure 2.16: *(a) BT_{27} and (b) BT_{35} tests for 30 January 2004 before (left) and after (right) applying the thresholds with the cloud pixels masked in red. Cloud-free and water body pixels are masked grey and white respectively.*

Brightness temperature difference tests using BT_{31} - BT_{32} and BT_{31} - BT_{22} were applied to the MODIS dataset. For BT_{31} - BT_{32} test, cloud can hardly be seen by visual analysis of Figure 2.17(a(left)). In Figure 2.17(a(right)), when the mask was applied, most clouds are detected in the east of Malaysia. Clouds detected by this test were thin high clouds, e.g. cirrus, cirrostratus and cirrocumulus. In Figure 2.17(b(left)), for $BT_{31} - BT_{22}$ test, it seems that only a few cloud patches are visible in the middle and south of Malaysia; pixels detected as cloud can be seen throughout the whole Malaysia when the mask was applied (Figure 2.17(b(right))). Clouds detected by $BT_{31} - BT_{22}$ are thick low clouds, e.g. cumulonimbus and cumulus, which the former brings heavy downpour in Malaysia during the Northeast monsoon season.

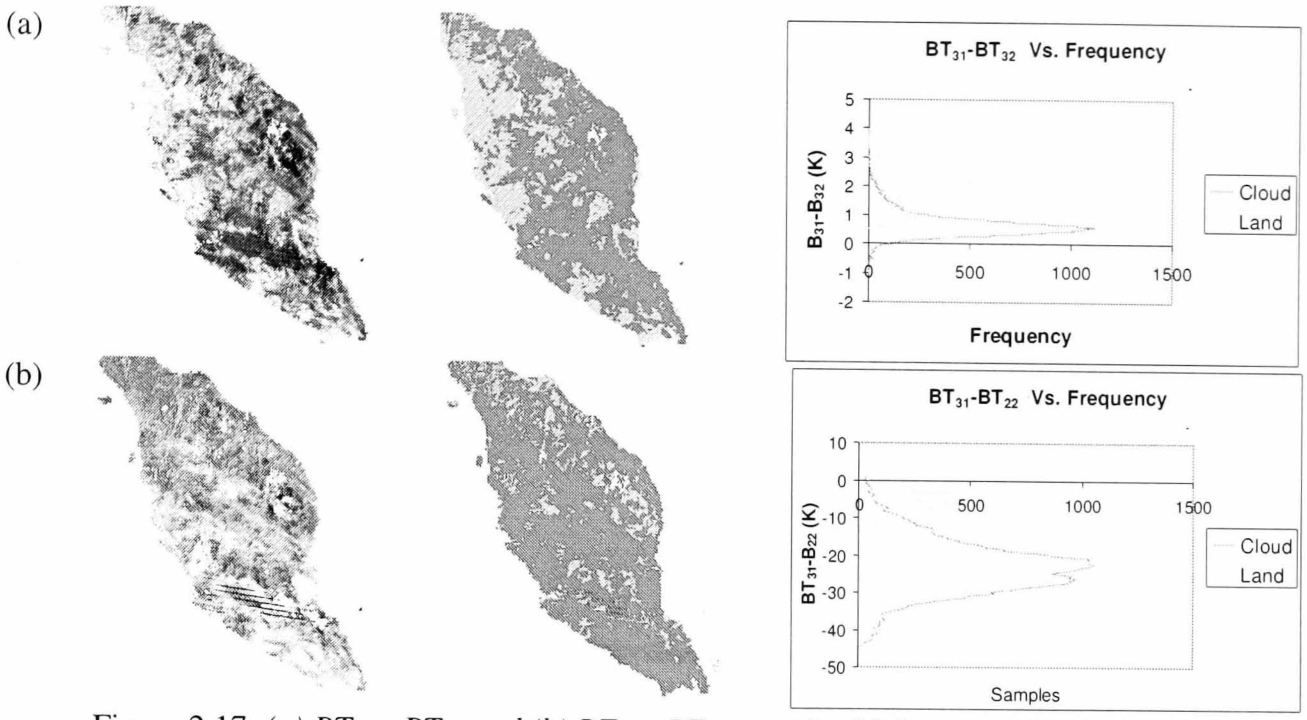


Figure 2.17: *(a) BT_{31} - BT_{32} and (b) BT_{31} - BT_{22} tests for 30 January 2004 before (left) and after (right) applying the thresholds with the cloud pixels masked in red. Cloud-free and water body pixels are masked grey and white respectively.*

It is useful to know the amount of cloud captured by each test so we can assess its effectiveness. Hence, for each test we calculate the amount of cloud in terms of area (km^2) and percentage land area (%). All cloud tests used and the amount of cloud captured on 30 January 2004 are given in Table 2.8. R_1 gives the largest area, i.e. 84% of the land or 121,549 km^2. This is followed by the BT_{31} – BT_{22} with 82% (119,968 km^2) and the BT_{29} – BT_{31} test with 70% (102,370 km^2). The least clouds are detected by the BT_{27}, 7% (9,549 km^2) and BT_{35}, 9% (13,522 km^2). The 84% captured by the R_1 test is due to the various types of cloud that are detectable in 0.66 μm wavelength; this owing to the much higher difference in cloud and vegetation spectral properties.

Table 2.8: *Cloud tests and area covered for 30 January 2004.*

Mask type	MODIS Test (based on band number)	Same as the second column, but based on wavelengths	Area	
			(km^2)	Percentage of cloud from land area (%)
Cloud mask	BT_{27}	$BT_{(6.7)}$	9549	6.6
	BT_{35}	$BT_{(13.9)}$	13522	9.3
	BT_{31} – BT_{32}	$BT_{(11)}$ – $BT_{(12)}$	90627	62.3
	BT_{31} – BT_{22}	$BT_{(11)}$ – $BT_{(3.9)}$	119968	82.5
	R_1	$R_{(0.66)}$	121549	83.6
	R_{26}	$R_{(1.38)}$	52406	36.0

The final spectral cloud masks were prepared from the six tests given above. Since the cloud masking performed here was meant for cloud conservative, i.e. ensures that no cloud-free pixel is identified as cloudy; therefore we selected the maximum confidence level for all tests (see Table 2.2) (Ackerman et al. 2010). A pixel was labelled as cloudy if it was identified as cloud by at least one test. By combining all the cloud tests, the final cloud mask for Malaysia is shown in Figure 2.18; cloud covers approximately 97% or 141000 km^2 of the land area.

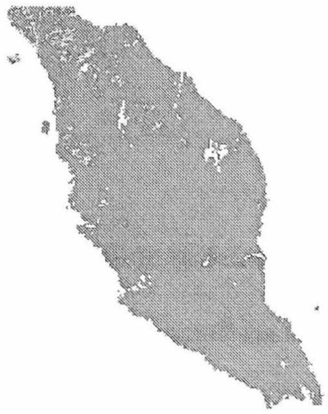


Figure 2.18: *The final cloud mask for Malaysia for 30 January 2004. Cloud pixels are masked red; cloud-free and water body pixels are masked grey and white respectively.*

To examine the mask in term of overlapping tests, we segmented the mask based on the number of tests that occured. Figure 2.19 shows the cloud mask for 30 January 2004 classified based on the number of overlapping tests. The colours (blue, cyan, yellow, magenta, maroon and green) are associated with the number of tests, while non-cloud and water pixels are masked grey and white respectively. It can be seen that the three-tests overlapping covers the largest area (36%) followed by the two-tests overlapping (24%) and the four-tests overlapping (16%), while the five-tests overlapping has the smallest area (4%). The six-tests overlapping (5%) occurs at the middle southern parts of Malaysia (southern Pahang, northern Johor and southern Selangor) – this indicates that several types of cloud occurred simultaneously over these areas. This is consistent with the fact that these areas received much higher rain (e.g. Muadzam Shah station in southern Pahang recorded more than 270 mm of mean rainfall) than other areas during January every year (Malaysian Meteorological Department 2010).

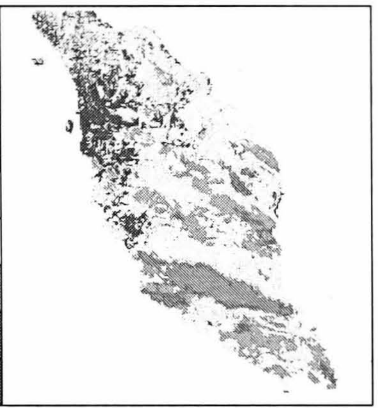
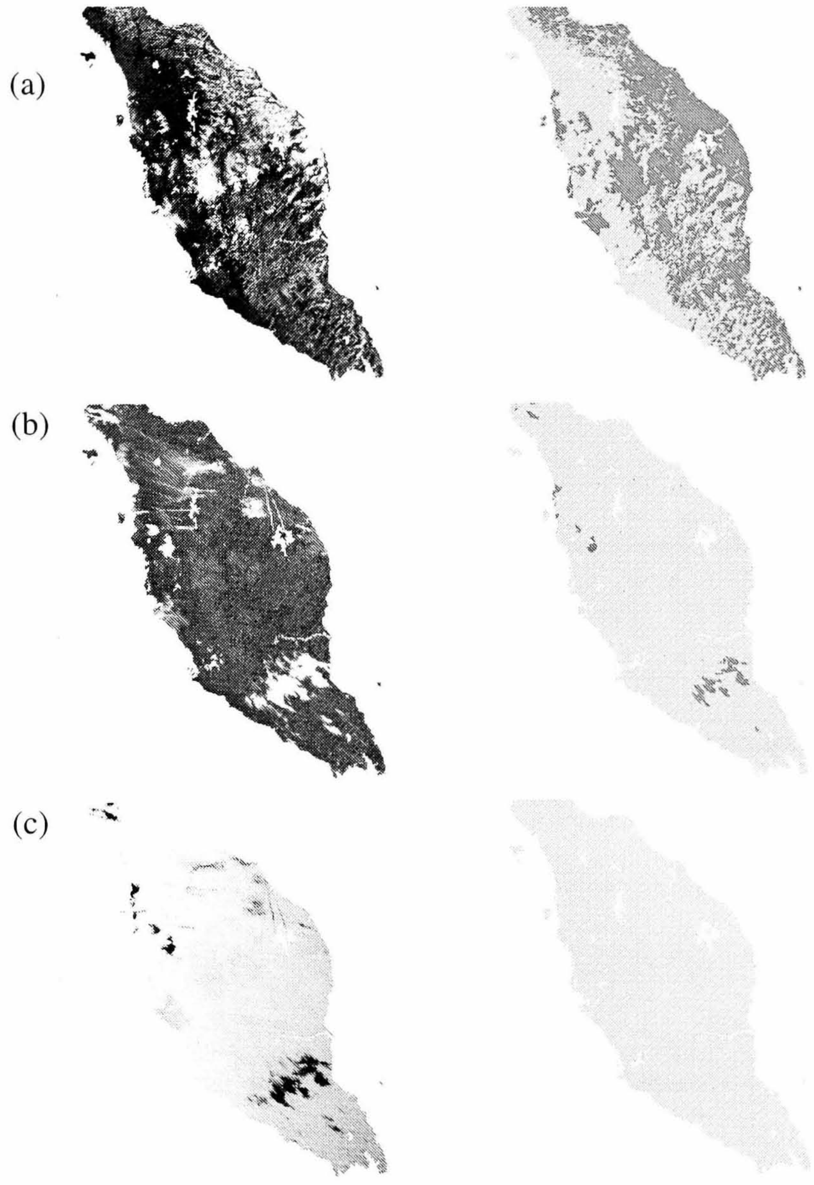
Number of test	Percent cloud from land area	Area (km^2)		
No cloud	3.4	4943	▨	
1	12.5	18174	■	
2	23.7	34458	□	
3	35.7	51905	□	
4	16.0	23263	▨	
5	3.7	5380	▨	
6	5.0	7270	▨	

Figure 2.19: *The cloud mask for 30 January 2004 classified based on the number of overlapping tests; the colours (blue, cyan, yellow, magenta, maroon and green) are associated with the number of tests, while non-cloud and water pixels are masked grey and white respectively.*

To evaluate the robustness of the cloud masking algorithm, the analysis was then applied to MODIS dataset from 15 February 2004, with sparser clouds. Similarly, all the tests are performed based on those shown in Table 2.2. Figure 2.20 shows the results of all the tests (a) R_1, (b) R_{26}, (c) BT_{35}, (d) BT_{27}, (e) BT_{31} - BT_{32}, (f) BT_{31} - BT_{22} before (left) and after (right) mask applied; cloud pixels are masked red while cloud-free and water body pixels are masked grey and white respectively. Test BT_{31} – BT_{32} detected the most cloud (74% or 107591 km^2) followed by R_1 and BT_{31} – BT_{22}, while no cloud is detected by BT_{35} test (Table 2.9). The final cloud mask for 15 February 2004 is shown in Figure 2.20(g). In overall, 83% or 121000 km^2 of land area was found covered with cloud and as expected, the eastern parts of Malaysia having more cloud than the western parts. This is about 14% less than that of 30 January 2004; February falls within the inter-monsoon season (i.e. the dry period in between the Northeast Monsoon and Southwest Monsoon), so is drier than January (Malaysian Meteorological Department 2010).

Table 2.9: *Cloud tests and area covered for 15 February 2004.*

Mask type	MODIS Test (based on band number)	Same as the second column, but based on wavelengths	Area	
			(km^2)	Percentage of cloud from land area (%)
Cloud mask	BT_{27}	$BT_{(6.7)}$	29	0.02
	BT_{35}	$BT_{(13.9)}$	0	0
	$BT_{31} - BT_{32}$	$BT_{(11)} - BT_{(12)}$	107591	74.0
	$BT_{31} - BT_{22}$	$BT_{(11)} - BT_{(3.9)}$	32277	22.2
	R_1	$R_{(0.66)}$	60775	41.8
	R_{26}	$R_{(1.38)}$	11922	8.2



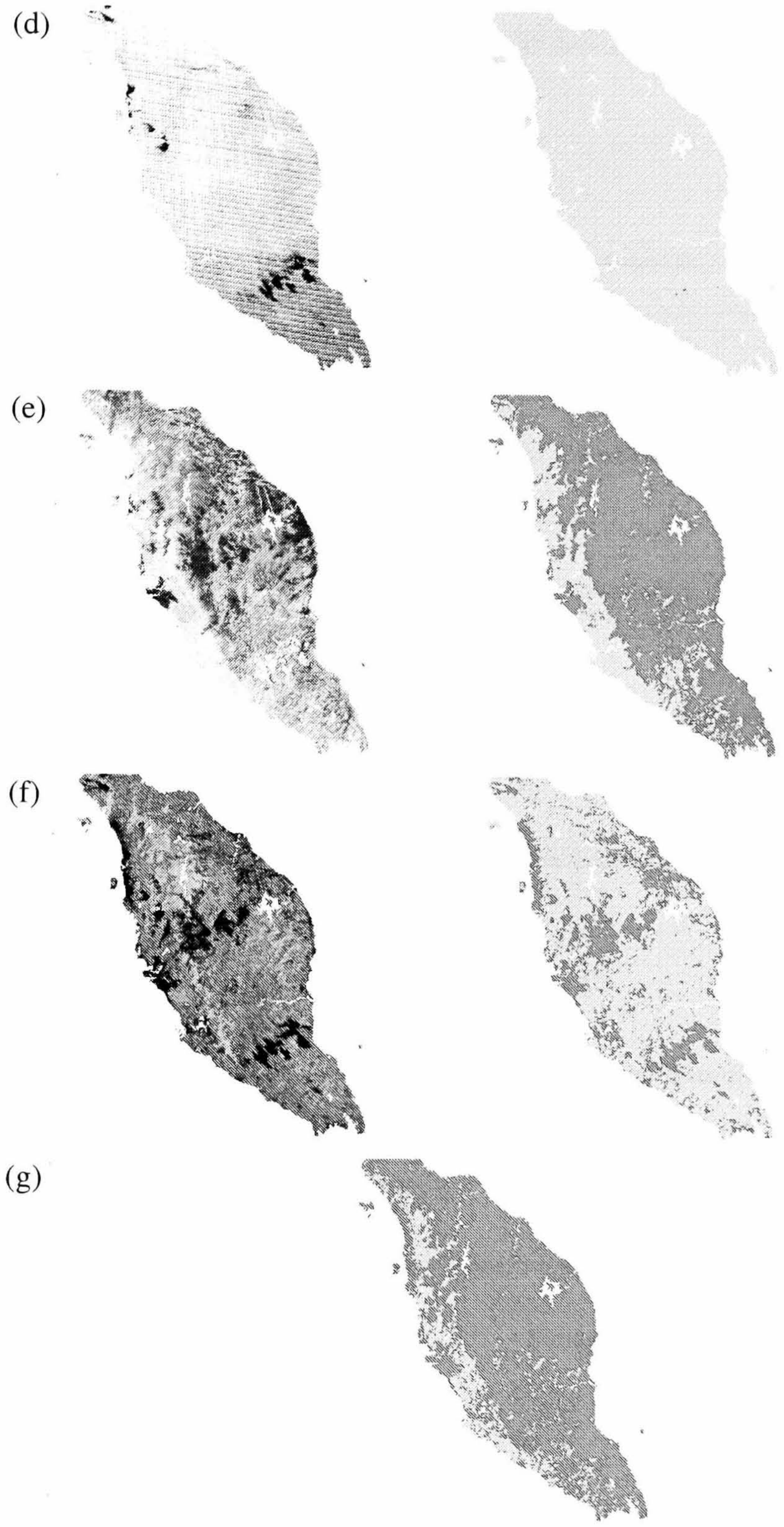


Figure 2.20: *Cloud masking results from all the tests (a) R_1, (b) R_{26}, (c) BT_{35}, (d) BT_{27}, (e) BT_{31} - BT_{32}, (f) BT_{31} - BT_{22} before (left) and after (right) mask applied. (g) the final cloud mask for 15 February 2004; cloud pixels are masked red, while cloud-free and water body pixels are masked grey and white respectively.*

Figure 2.21 shows the cloud mask for 15 February 2004 classified based on the number of overlapping tests for 15 February 2004. The colours (blue, cyan, yellow, magenta, maroon and green) are associated with the number of tests, while non-cloud and water pixels are masked grey and white respectively. Most of the cloud pixels were due to the single test (40%). The percentage decreases with the number of tests; only 0.4% pixels are associated with the four-tests overlapping. Unlike January dataset, there were no pixels detected as cloud by all the six or even five of the tests. This indicates the less cloud (and so as rain) that occurs in February due to the effects of the inter-monsoon season (Malaysian Meteorological Department 2010).

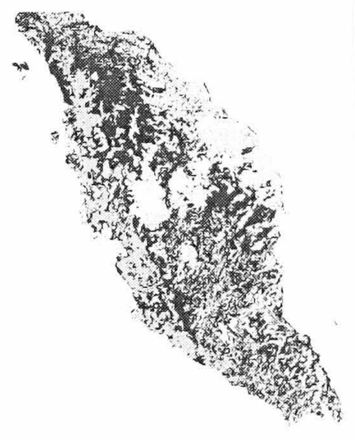
Number of test	Percent cloud from land area	Area (km^2)		
No cloud	17.4	25298	▒	
1	39.8	57867	■	
2	28.8	41873	□	
3	13.6	19774	□	
4	0.4	582	▒	
5	0	0	■	
6	0	0	■	

Figure 2.21: *The cloud mask for 15 February 2004 classified based on the number of overlapping tests; the colours (blue, cyan, yellow, magenta, maroon and green) are associated with the number of tests, while non-cloud and water pixels are masked grey and white respectively.*

Multitemporal Cloud Analysis

We further investigate the effectiveness of the cloud analysis by applying it to multitemporal datasets. The same procedure such as that of the 30th January 2004 dataset were applied to 24 other datasets from January 2004 to December 2005 at 0355 UTC (1155 LST). Figure 2.22 shows cloud masks generated for these datasets. It can be seen that cloud distribution changes dynamically with time; in overall, the cloud amount in 2005 seems to be more than 2004. This agrees with the fact that the total amount of rain received in 2005 was more than 2004 due to the effects of La Nina (wet spell) and El Nino (dry spell) respectively (Malaysian Meteorological Department 2010).

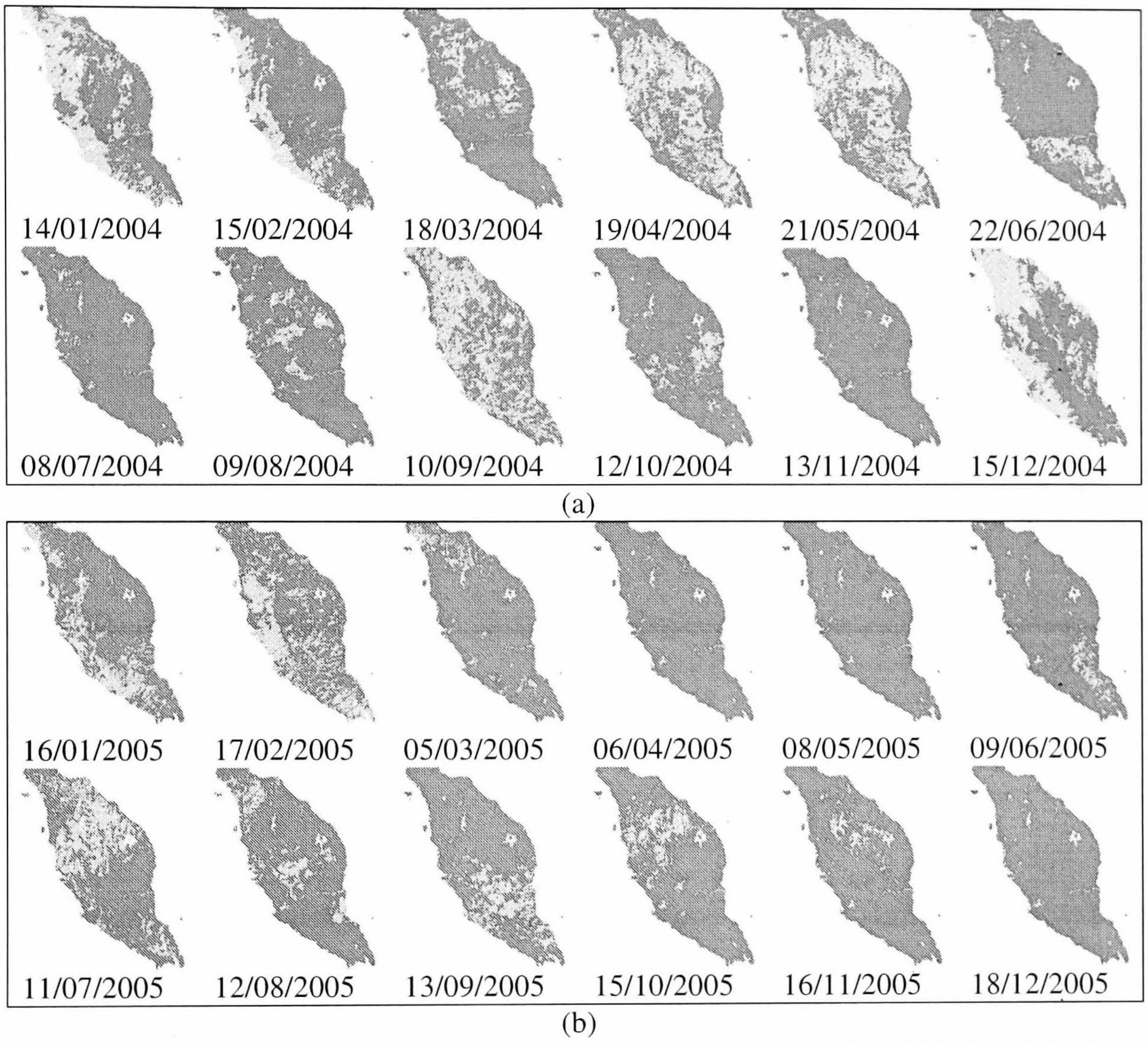


Figure 2.22: *Cloud masking for selected dates in (a) 2004, and (b) 2005; cloud pixels are masked red, while cloud-free and water body pixels are masked grey and white respectively.*

To see the cloud trend within this period, we plotted graph of area against the date of the datasets. Figure 2.23 shows cloud area against the date of the data from January 2004 to December 2005. It is noticeable that the 2005 datasets have more cloud than 2004 due to the effects of La Nina and El Nino respectively (Malaysian Meteorological Department 2010).

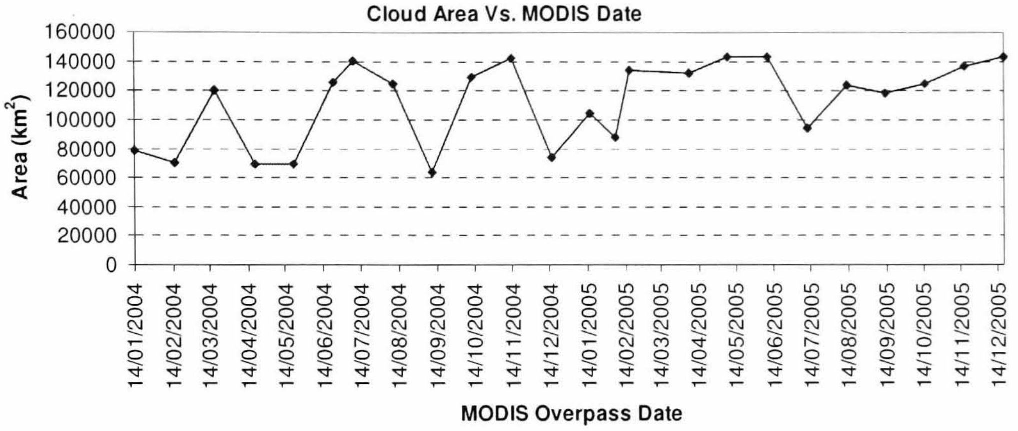


Figure 2.23: *Cloudy area versus different MODIS acquisition date from January 2004 to December 2005.*

We further examine the areas where clouds are prone to form by classifying the cloud based on its frequency of occurrence. This was carried out by overlapping the cloud masks in Figure 2.22 and then assigning colours to cloud pixels, based on the frequency of occurrence, for the year 2004 and 2005. Figure 2.24 shows the cloud area classified based on overlapping cloud pixels from the selected dates within 2004 and 2005; The colours are associated with the number of overlapping dates; non-cloud and water are masked grey and white respectively. The eastern parts of Malaysia seems to have more cloudy days than the western parts, in which consistence with the fact that the former is having more amount of annual rain than the later. It is also clear that the year 2005 is cloudier than 2004, in which is consistent with Figure 2.23, due to La Nina and En Nino respectively (Malaysian Meteorological Department 2010).

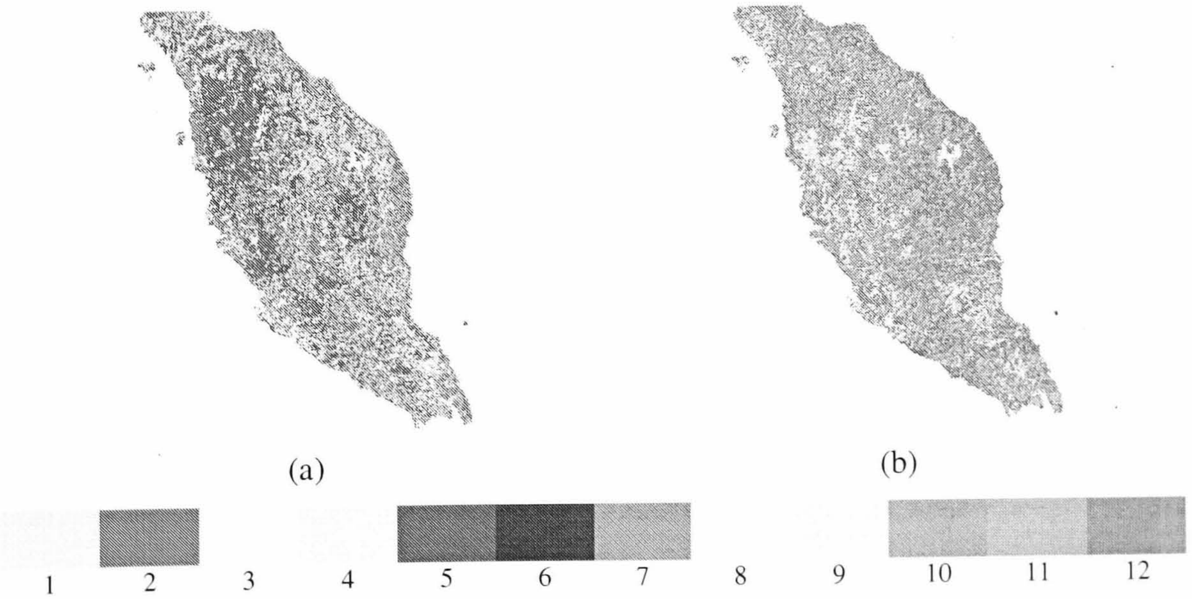


Figure 2.24: *Cloud area classified based on frequency of cloud occurrence from selected dates for (a) 2004 and (b) 2005. The colours are associated with the number of overlapping dates; non-cloud and water are masked grey and white respectively.*

Cloud Shadow Masking from MODIS Data

Cloud shadow masking was carried out based on $R_{19} < 0.07$ and $R_2/R_1 > 0.3$, and $R_5 < 0.2$; pixels were labelled as cloud shadow if they pass all these tests at once (Ackerman et al. 2006). Figure 2.25 shows (a) R_{19}, R_2/R_1 and R_5 assigned to red, green and blue respectively and (b) the final cloud shadow mask for Malaysia for 30 January 2004; cloud shadow pixels are masked yellow, while cloud-free and water body pixels are masked grey and white respectively. The colour composite image (left) does not tell much about the cloud shadow distribution. When the tests were applied, cloud shadow (masked yellow) can be seen in mostly in the northwest of Malaysia in Figure 2.25(right). Table 2.10 gives the area covered by the cloud shadow analysis on 30 January 2004; cloud shadow area is 2.5% of the land or 3674 km^2.

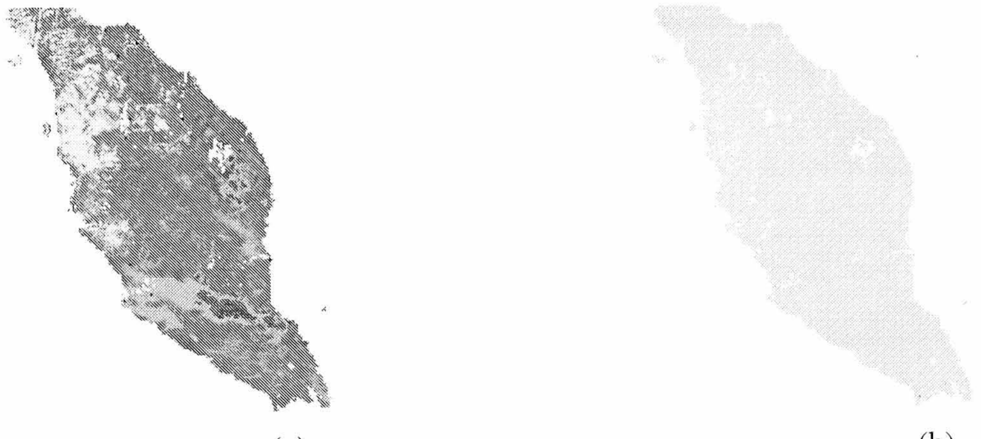


Figure 2.25: *(a) R_{19}, R_2/R_1 and R_5 assigned to red, green and blue respectively and (b) the final cloud shadow mask for Malaysia for 30 January 2004; cloud shadow pixels are masked yellow, while cloud-free and water body pixels are masked grey and white respectively.*

Table 2.10: *Cloud shadow test and area covered 30 January 2004.*

Mask type	MODIS Test (based on band number)	Same as the second column, but based on wavelengths	Area	
			(km^2)	Percentage from land area (%)
Cloud shadow mask	R_{19}; R_2/R_1; R_5	$R_{(0.94)}$; $R_{(0.87)}/R_{(0.66)}$; $R_{(1.2)}$	3674	2.5

A similar procedure is applied to dataset dated 15 February 2004; the cloud shadow mask is given in Figure 2.26(b) and the area is given in Table 2.9. More cloud shadows are found on 15 February (8.4% or 12213 km^2) compared to 30 January dataset because the severe cloud amount in the latter has prevent the cloud shadows to be visible from the satellite sensor.

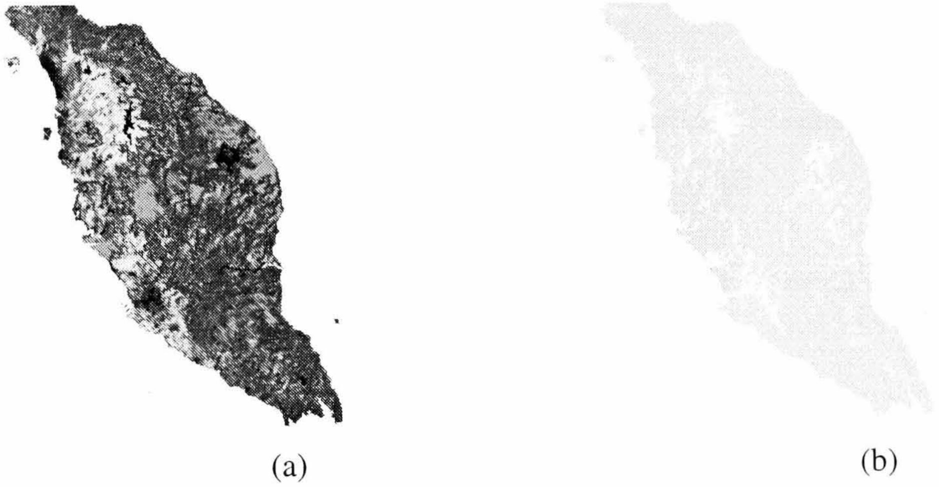


Figure 2.26: *(a) R_{19}, R_2/R_1 and R_5 in red, green and blue respectively and (b) the final cloud shadow mask for Malaysia for 15 February 2004; cloud shadow pixels are masked yellow, while cloud-free and water body pixels are masked grey and white respectively.*

Table 2.11: *Cloud shadow test and area covered 15 February 2004.*

Mask type	MODIS Test (based on band number)	Same as the second column, but based on wavelengths	Area	
			(km^2)	Percentage from land area (%)
Cloud shadow mask	R_{19}; R_2/R_1; R_5	$R_{(0.94)}$; $R_{(0.87)}/R_{(0.66)}$; $R_{(1.2)}$	12213	8.4

2.6 Application of the Cloud Analysis to Landsat data

The Landsat Satellite

The Landsat satellites have been providing optical data for almost 40 years. Landsat 1 – 3 launched in the 1970s and used Multispectral Scanner (MSS), while Landsat 4 – 5, launched in the 1980s, use Thematic Mapper (TM) as their main sensor. The latest Landsat 7, launched in 1999, uses the Enhanced Thematic Mapper (ETM+). Comparison between the specifications of these satellites is given in Table 2.12.

Landsat 5 was launched on March 1, 1984 with an expected lifetime of 5 years, and, after more than 26 years of operation, has provided the global science community with over 900,000 individual scenes and is the longest running satellite of the series (Figure 2.27). This study uses Landsat 5 TM for land cover classification, haze simulation and haze removal purposes (Chapters 3, 4 and 5).

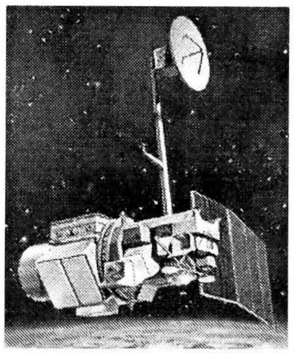


Figure 2.27: *Landsat 5 satellite (USGS 2010).*

Table 2.12: *Landsat satellite specifications (Markham et al. 2004).*

Landsat Satellite	Landsat 1 – 3	Landsat 4 – 5	Landsat 7
Spectral Bands	4 VNIR, 1 thermal (Landsat 3)	4 VNIR, 2 SWIR, 1 thermal	4 VNIR, 2 SWIR, 1 thermal, 1 panchromatic
Spatial Resolution (IFOV)	79 m – VNIR 240 m – thermal	30 m – VNIR, SWIR 120 m – thermal	30 m – VNIR, SWIR 60 m – thermal 15 m – panchromatic
Sampling	1.4 samples/IFOV along scan	1 samples/IFOV along scan	1 samples/IFOV along scan
Cross Track Coverage	185 km	185 km	183 km
Radiometric Resolution	6 bits (usually non-linearly compressed in bands 1 – 3 and decompressed to 7 bits on the ground)	8 bits	8 bits (2 gain states)

Radiometric Calibration	Internal lamps and shutter, Partial aperture solar (Landsat 1 – 3)	Internal lamps, shutter and black body	Internal lamps, shutter and black body, partial aperture solar, full aperture solar diffuser
Scanning Mechanism	Unidirectional Scanning	Bidirectional Scanning with Scan Line Corrector	Bidirectional Scanning with Scan Line Corrector
Period of operation	Landsat 1: 1972 – 1975 Landsat 2: 1975 – 1982 Landsat 3: 1978 – 1983	Landsat 4: 1982 – 2001 Landsat 5: 1984 – present	Landsat 7: 1999 – present
Main sensor	MSS	MSS TM	ETM
Altitude	917	705 km	705 km
Repeat Cycle	18 days	16 days	16 days
Equatorial Crossing	9:30 AM +/- 15 minutes	9:45 AM +/- 15 minutes	10:00 AM +/- 15 minutes
Type	Sun synchronous, near polar	Sun synchronous, near polar	Sun synchronous, near polar
Inclination	99.2° 99.1° (Landsat 3)	98.2°	98.2°

Landsat 5 TM level 1 data come in Product Generation System (LPGS) format and need to be converted into a physically meaningful common radiometric unit, representing the at-sensor spectral radiance. The Level 1 Landsat 5 TM data received by users are in scaled 8-bit numbers, Q_{cal}, or also known as digital number (DN). Conversion from Q_{cal} to spectral radiance, L_λ, can be done by using the following equation (Chander et al. 2009):

$$L_\lambda = \frac{(L_{\max\lambda} - L_{\min\lambda})}{(Q_{cal\max} - Q_{cal\min})}(Q_{cal} - Q_{cal\min}) + L_{\min\lambda} \qquad \dots (2.7)$$

where

L_λ = Spectral radiance at the sensor's aperture (W/ m^2 sr μm)

Q_{cal} = Quantized calibrated pixel value (DN)

$Q_{cal\min}$ = Minimum quantised calibrated pixel value corresponding to $L_{\min\lambda}$ (DN)

$Q_{cal\max}$ = Maximum quantised calibrated pixel value corresponding to $L_{\max\lambda}$ (DN)

$L_{\min\lambda}$ = Spectral at-sensor radiance that is scaled to $Q_{cal\min}$ (W/ m^2 sr μm)

$L_{\max\lambda}$ = Spectral at-sensor radiance that is scaled to $Q_{cal\max}$ (W/ m^2 sr μm)

Q_{calmin} and Q_{calmax} are 1 and 255 respectively. Table 2.13 shows $L_{min\lambda}$, $L_{max\lambda}$ and the mean exoatmospheric solar irradiance (E_{λ}).

Table 2.13: *Landsat TM spectral range, post-calibration dynamic ranges and the mean exoatmospheric solar irradiance (Chander et al. 2009).*

Band	Spectral range	Centre wavelength	$L_{min\lambda}$	$L_{max\lambda}$	E_{λ}
	(μm)		(W/ m^2 sr μm)		
1	0.452 – 0.518	0.485	-1.52	169	1983
2	0.528 – 0.609	0.569	-2.84	333	1796
3	0.626 – 0.693	0.660	-1.17	264	1536
4	0.776 – 0.904	0.840	-1.51	221	1031
5	1.567 – 1.784	1.676	-0.37	30.2	22.0
6	10.45 – 12.42	11.435	1.2378	15.3032	N/A
7	2.097 – 2.223	2.223	-0.15	16.5	83.44

Scene-to-scene variability can be reduced by converting the at-sensor spectral radiance to TOA reflectance, also known as in-band planetary albedo. By performing this conversion, the cosine effect of different solar zenith angles due to the time difference between data acquisitions is removed, different values of the exoatmospheric solar irradiance arising from spectral band differences are compensated and the variation in the Earth–Sun distance between different data acquisition dates is corrected. The TOA reflectance can be computed by using (Chander et al. 2009):

$$\rho_{\lambda} = \frac{\pi L_{\lambda} d^2}{E_{\lambda} \cos(\theta_s)} \qquad \text{... (2.8)}$$

where

ρ_{λ} = Planetary TOA reflectance (unitless)

π = Mathematical constant equal to ~3.14159 (unitless)

L_{λ} = Spectral radiance at the sensor's aperture (W m^{-2} sr^{-1} μm^{-1})

d = Earth–Sun distance (astronomical units)

E_{λ} = Mean exoatmospheric solar irradiance (W m^{-2} μm^{-1})

θ_s = Solar zenith angle (degrees) d can be generated from the Jet Propulsion Laboratory (JPL) Ephemeris at http://ssd.jpl.nasa.gov/?horizons or can be obtained from the literature (e.g. Chander et al. (2009)). In this study, conversion to at-sensor spectral radiance and TOA reflectance is performed using ENVI software.

The relationship between Landsat bands and MODIS cloud bands is shown in Table 2.14. It can be seen that 8 MODIS cloud bands overlap with Landsat bands. Due to the much narrower bandwidth, a Landsat band can overlap with more than one MODIS bands.

Table 2.14: *Relationship between MODIS cloud bands and Landsat bands. Shaded area indicates irrelevancy.*

MODIS					**Landsat**				
Spatial Resolution (m)	Band No.	Band Range (µm)	Centre wavelength (µm)	Band width (nm)	Spatial Resolution (m)	Band No.	Band Range (µm)	Centre wavelength (µm)	Band width (nm)
250	1	0.620 – 0.670 (Red)	0.659	41.8	30	3	0.626 – 0.693 (Red)	0.660	67.0
	2	0.841 – 0.876 (Near infrared)	0.865	39.4	30	4	0.776 – 0.904 (Near infrared)	0.840	128.0
500	3	0.459 – 0.479 (Blue)	0.470	17.6	30	1	0.452 – 0.518 (blue)	0.485	66.0
	4	0.545 – 0.565 (Green)	0.555	19.7	30	2	0.528 – 0.609 (Green)	0.569	81.0
	5	1.230 – 1.250 (Near infrared)	1.240	24.5					
	6	1.628 – 1.652 (Mid infrared)	1.640	29.7	30	5	1.55 – 1.75 (Mid infrared)	1.676	200.0
	7	2.105 – 2.155 (Mid infrared)	2.130	52.9	30	7	2.08 – 2.35 (Mid infrared)	2.223	270.0
1000	19	0.915 – 0.965 (Near infrared)	0.940	46.3					
	22	3.929 - 3.989 (Mid infrared)	3.959	85.7					
	26	1.360 - 1.390 (Near infrared)	1.375	94.3					
	27	6.535 - 6.895 (Mid infrared)	6.715	254.6					
	31	10.780 – 11.280 (Thermal infrared)	11.030	510.3	120	6	10.40 – 12.50 (Thermal infrared)	11.435	2100.0
	32	11.770 – 12.270 (Thermal infrared)	12.020	493.5					
	35	13.785 - 14.085 (Thermal infrared)	13.935	300.0					

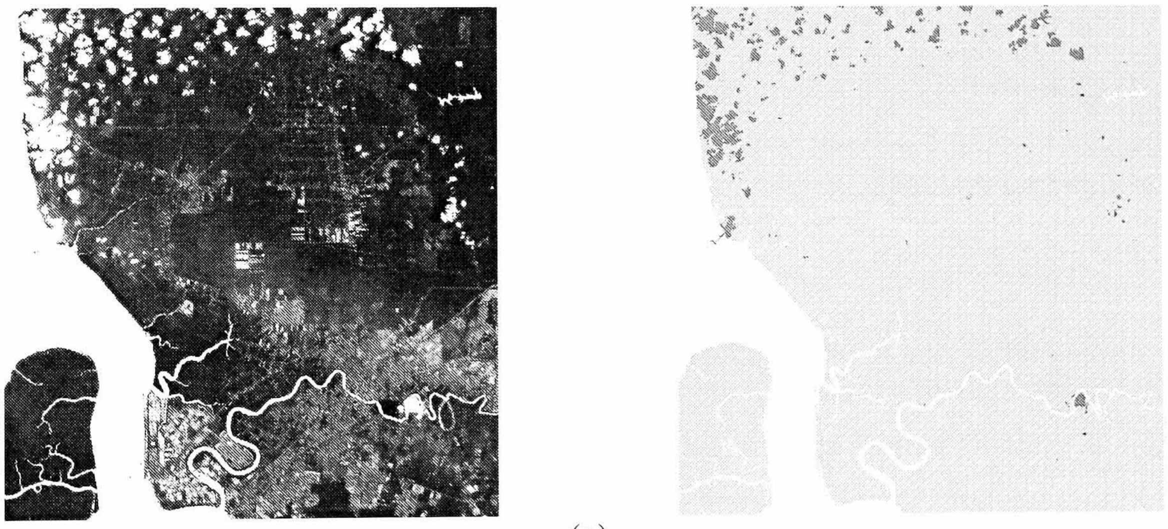
Results of the Applications of the Cloud Analysis to Landsat Data

By analysing Table 2.9 and Table 2.14, in term of spectral characteristics, for visible wavelengths, it can be seen that R_3 of Landsat with centre wavelength (0.660) closely matches with R_1 of MODIS (0.659 μm centre wavelength) so Landsat R_3 can be used to simulate the MODIS R_1 test in order to detect cloud. For thermal infrared wavelengths, only BT_6 is available on Landsat so it will be used to simulate the BT_{35} test of MODIS. After exhaustive testing of a variety of thresholds to separate cloud and non-cloud within Landsat data, we settled on 0.23 and 291 K for R_3 and BT_6 respectively; pixels values greater than 0.23 and less than 291 K in R_3 and BT_6 respectively will be classified as cloud. By combining both tests, a pixel will be flagged as cloudy if it is detected as cloud by at least one of the tests. Similarly, for cloud shadow, we analysed Table 2.11 and Table 2.14 and found that the ratio of R_4/R_3 of Landsat matches the ratio of R_2/R_1 of MODIS cloud shadow. Besides that, R_4 of Landsat will be used to simulate R_{19} test in MODIS. We found that the same thresholds as used in MODIS also suited Landsat data; pixels values less than 0.07 in R_4 and greater than 0.3 in R_4/R_3 were to be flagged as cloud shadow. Table 2.15 shows cloud and cloud shadow tests and their thresholds for Landsat and those equivalents in MODIS.

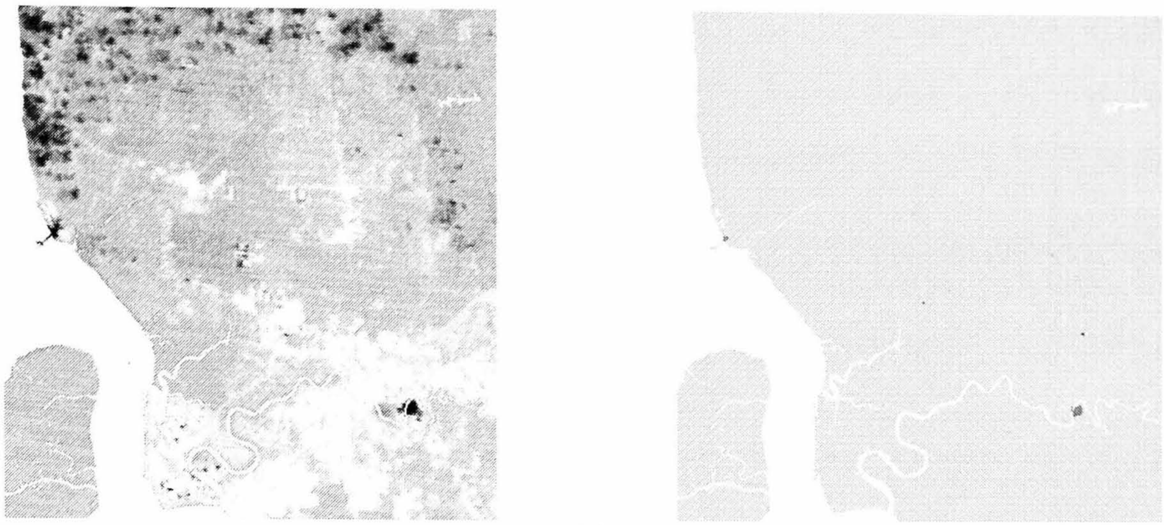
Table 2.15: *Tests and thresholds for Landsat data.*

Mask type	MODIS Tests (subscript refers to MODIS band)	Tests Determined for Landsat Data (subscript refers to Landsat band)
Cloud	$R_1 > 0.14$ or $BT_{35} < 226$ K	$R_3 > 0.23$ or $BT_6 < 291$ K
Cloud shadow	$R_{19} < 0.07$ and $R_2/R_1 > 0.3$	$R_4 < 0.07$ and $R_4/R_3 > 0.3$

For the purpose of this thesis, cloud masking will be carried out on Landsat data for Klang in Selangor, Malaysia, which located within longitude 101° 10' E to 101°30' E and latitude 2°99' N to 3°15' N which covers an area of approximately 540 km^2. Initially, cloud masking was carried out on data from 2 April 1994 by using the tests and thresholds as given in Table 2.15. Figure 2.28 shows (a) R_3 and (b) BT_6 in raw form (left) and with cloud mask (right) for 2 April 1994; cloud pixels are masked red, while cloud-free and water body pixels are masked grey and white respectively. In Figure 2.28(a)(left), due to the very high reflectance, cloud can be seen as white patches in the Northern parts of the image; in the middle and southern parts, cloud patches are seen quite similar to other bright features (e.g. bare land and urban). After red mask is applied to the cloud and grey mask as non-cloud, a much clearer view of cloud was obtained; some cloud patches can be seen in the middle and southern parts of the image Figure 2.28(a)(right). In Figure 2.28(b)(left), cloud, due to its very low temperature, appears as black patches in the northern and middle of the image; not much cloud is detected by the BT_6 test as seen in Figure 2.28(b)(right). More cloud pixels are detected by R_3 (2.8% or 15 km^2) than BT_6 (0.1% or 1 km^2) due to the better separation capability between cloud and non-cloud in reflective compared to thermal wavelengths (see Figure 2.13 and Figure 2.15).



(a)



(b)

Figure 2.28: *(a) R_3 and (b) BT_6 in raw form (left) and with cloud mask (right) for 2 April 1994; cloud pixels are masked red, while cloud-free and water body pixels are masked grey and white respectively.*

Pixels that were detected as cloudy by any of the tests were labelled as cloud pixels; they were found amounting 2.8% (15 km^2) from the land area, where 0.1% overlaps occur between cloud pixels detected by R_3 and BT_6. By combining the tests, pixels detected as cloud by at least one of the test were flagged as cloudy; the final cloud mask is shown in Figure 2.29.

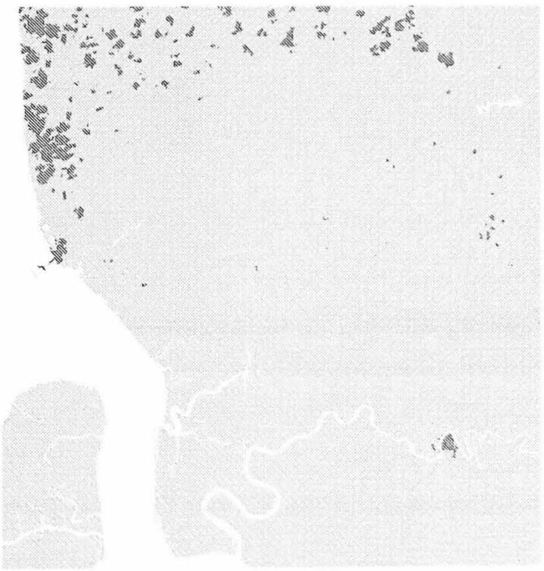


Figure 2.29: *The final cloud mask for Landsat data from 2 April 1994; cloud pixels are masked red, while cloud-free and water body pixels are masked grey and white respectively.*

For cloud shadow, R_4 and R_4/R_3 test were used simultaneously based on the thresholds given in Table 2.15. Figure 2.30 shows the outcomes of applying R_4 and R_4/R_3 tests in colour composite (left) and (b) the resulting cloud shadow mask for Landsat data from 2 April 1994; cloud shadow pixels are masked yellow, while cloud-free and water body pixels are masked grey and white respectively. Cloud shadow pixels were found 2.4% (13 km^2) from the land area.

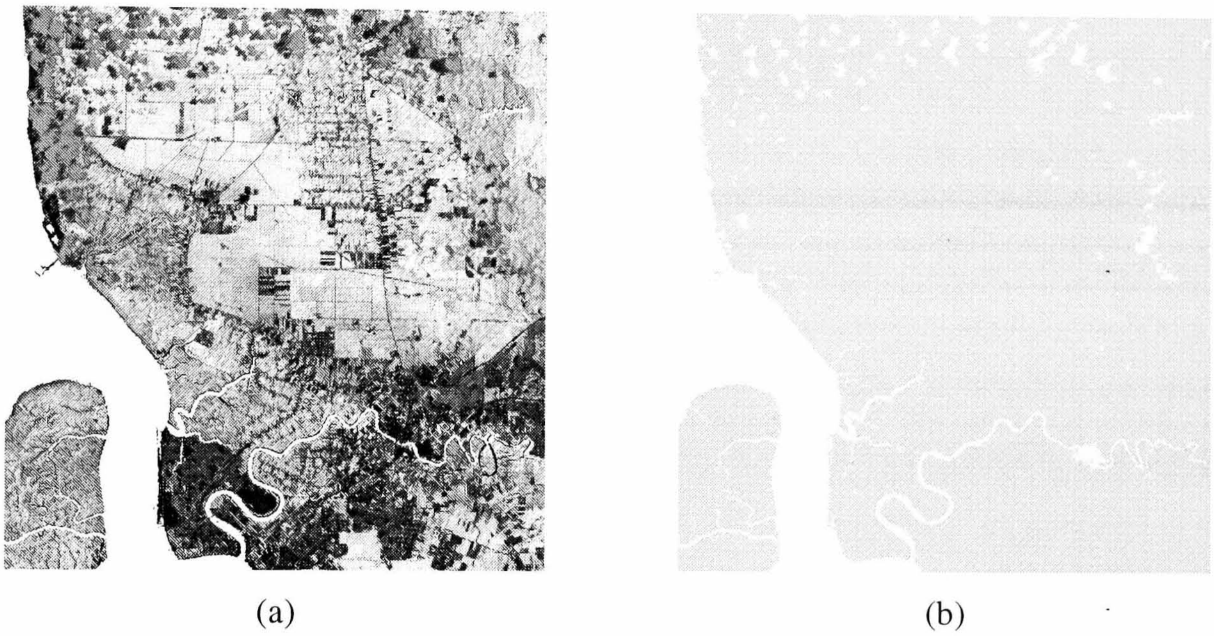


Figure 2.30: *Result of applying R_4 and R_4/R_3 in colour composite (left) and (b) with cloud shadow mask for Landsat data from 2 April 1994; cloud shadow pixels are masked yellow, while cloud-free and water body pixels are masked grey and white respectively.*

The outcomes from cloud and cloud shadow masks were combined and masked black; the combined mask is about 5.2% (28 km^2) from the land area (Figure 2.31).

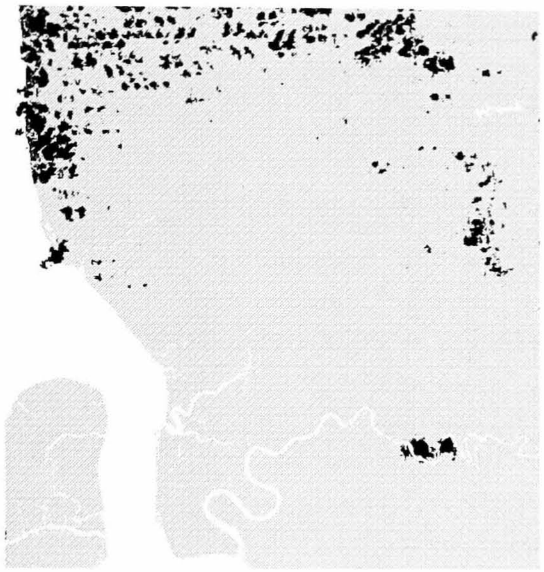
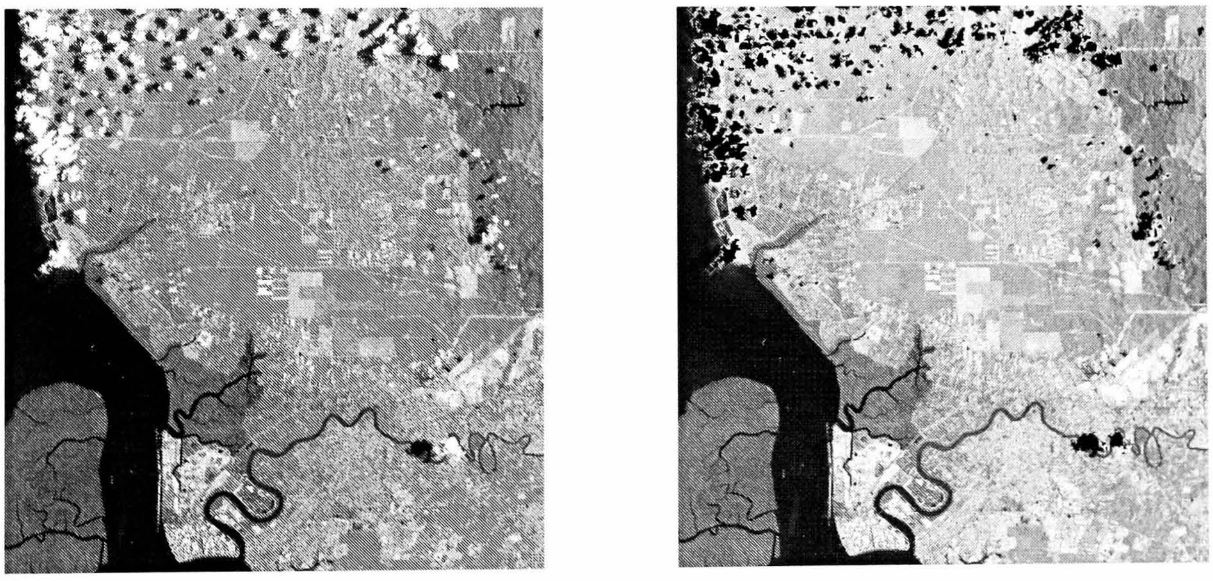


Figure 2.31: *The combined cloud and cloud shadow mask for Landsat data from 2 April 1994; cloud and cloud shadow pixels are masked black, while cloud-free, cloud shadow-free and water body pixels are masked grey and white respectively.*

Figure 2.32 shows the Landsat bands 4, 5 and 3 from 2 April 1994 assigned to red, green and blue (a) before and (b) after cloud and its shadow masked black. Visually most cloud and its shadow were successfully removed from the data.

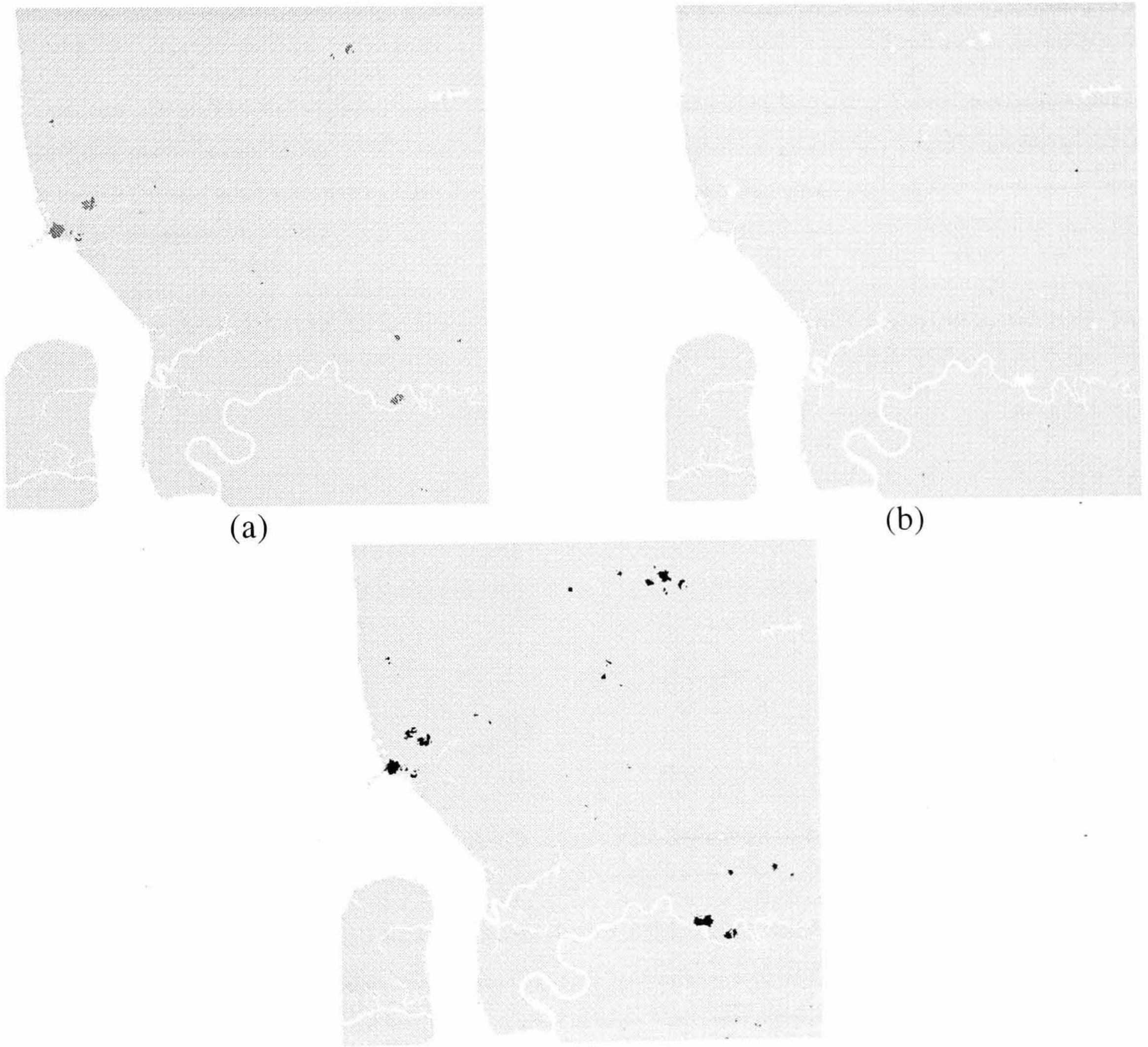


(a) before

(b) after

Figure 2.32: *Landsat data from 2 April 1994 (a) before and (b) after masking of cloud and its shadow; cloud and its shadow are masked black.*

The cloud analysis was then applied to Landsat data from 11 February 1999 with sparser cloud; this data will be used as the main data for the subsequent chapters of this thesis. Figure 2.33 shows (a) the cloud mask (red), (b) the cloud shadow mask (yellow) and (c) the combination of (a) and (b) (black) for Landsat data from 11 February 1999. The total cloud area was 0.24% or 1.3 km^2 from the land; cloud detected by R_3 was 0.2% or 1.2 km^2, while BT_6, 0.1% or 0.6 km^2 with about 0.06% overlapping between the two tests. For cloud shadow, the amount was 0.23% or 1.2 km^2 from the land area. Total cloud and cloud shadow (0.5% or 2.5 km^2). Figure 2.34 shows the data (a) before and (b) after masking of cloud and its shadow (masked black).



(a)

(b)

(c)

Figure 2.33: *(a) cloud mask (red patches), (b) cloud shadow mask (yellow patches) and (c) combination of (a) and (b) (black patches) for Landsat data from 11 February 1999.*

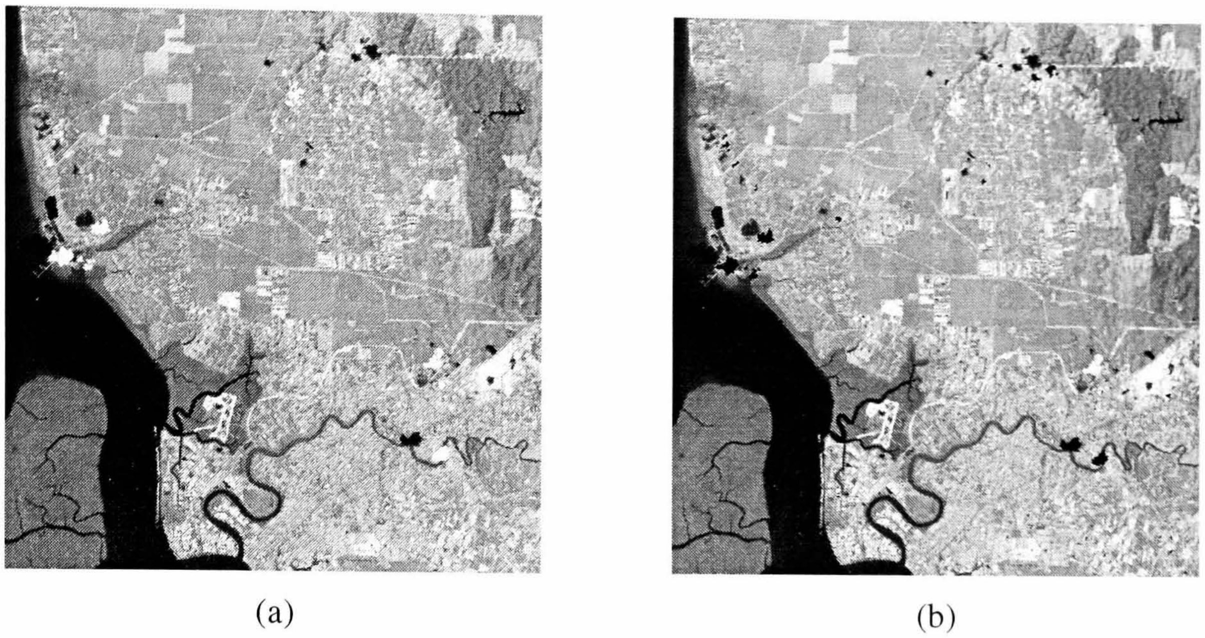
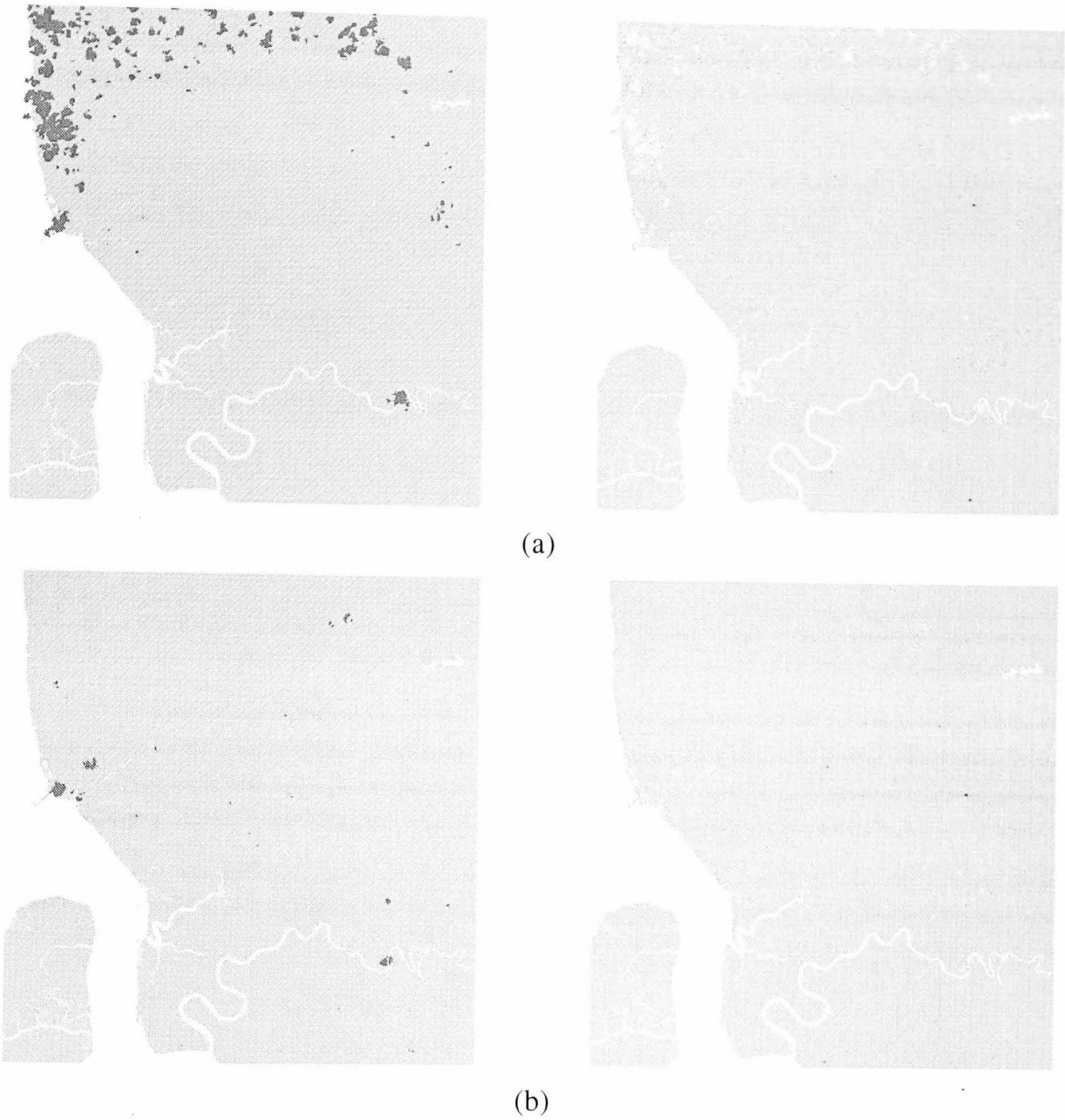


Figure 2.34: *Landsat data from 11 February 1999 before and after masking of cloud and its shadow; cloud and its shadow are masked black.*

Validation works were carried out in two parts, i.e. visual and quantitative analysis. For visual analysis, the cloud masking results were qualitatively compared with the ACCA scheme. Figure 2.35 shows cloud mask produced using our masking method (left) and ACCA scheme (right) from (a) 2 April 1994 and (b) 11 February 1999. The cloud was masked red for the cloud analysis and green for the ACCA scheme; non-cloud and water pixels were masked grey and white respectively. For 2 April 1994, as can be seen in Figure 2.35(a), the methods fairly agree between each other; only very small amount of cloud in the middle of the image that is not detected by the cloud analysis but detected by the ACCA method. This is due to the use of more tests in ACCA, so it tends to detects more cloud than the cloud analysis. For 11 February 1999, where the cloud patches are sparser, a more consistent outcome from both methods were obtained (Figure 2.35(b)).

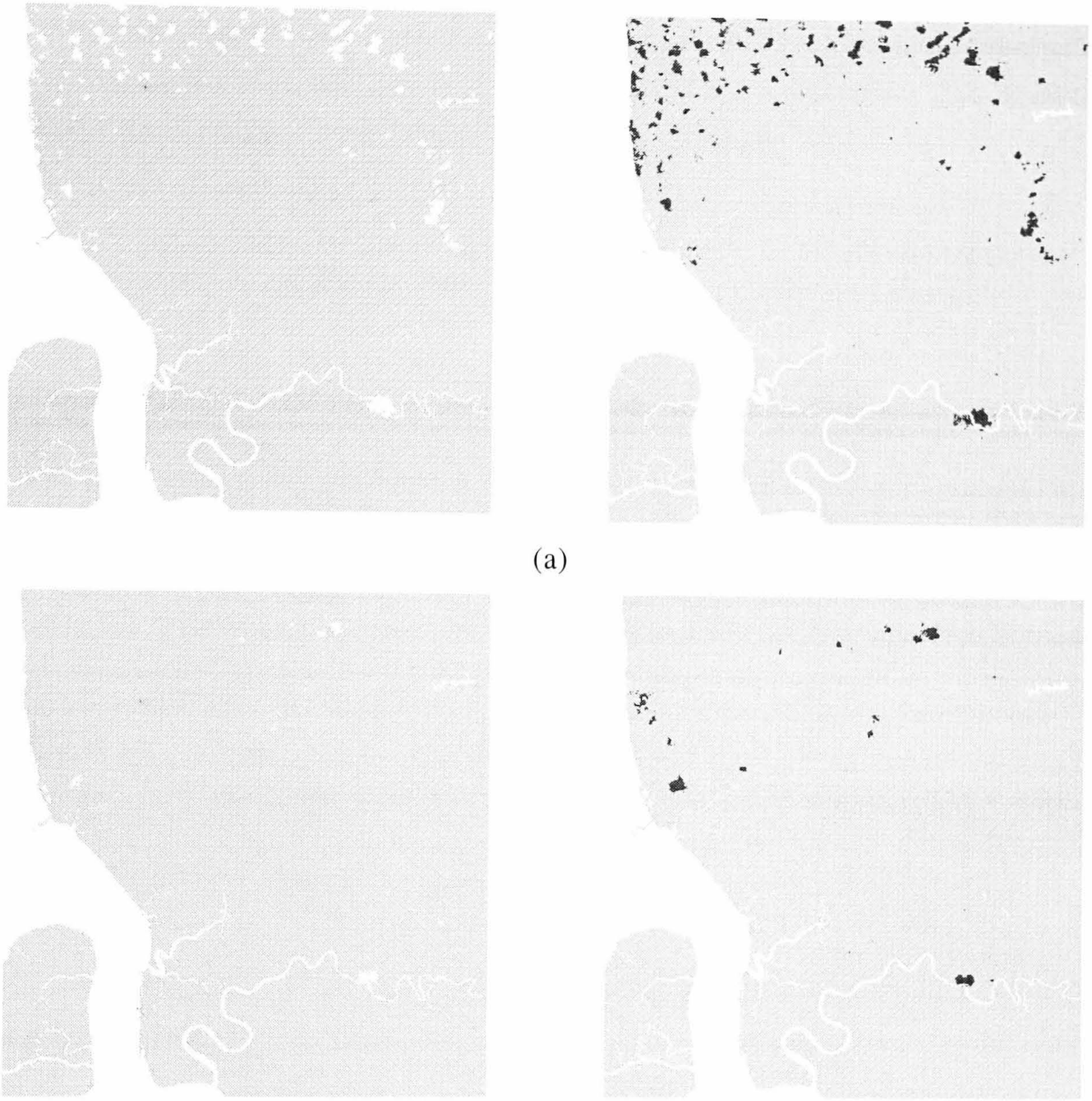

(a)

(b)

Figure 2.35: *Cloud mask produced using our masking method (left) and ACCA scheme (right) from (a) 2 April 1994 and (b) 11 February 1999. Cloud pixels are masked red for the cloud masking method and green for the ACCA scheme; non-cloud and water pixels are masked grey and white respectively.*

For cloud shadow, validation was made by visually compared with the Luo et al. (2008) method. Figure 2.36 shows cloud shadow mask produced using our masking method (left) and Luo et al. (2008) scheme (right) from (a) 2 April 1994 and (b) 11 February 1999. For 2 April, the outcomes from both methods are comparatively consistent. For 11

February 1999, more patches of cloud shadow were detected near the Northwestern coastal areas by Luo et al. (2008) method compared to the cloud shadow analysis.



(a)

(b)

Figure 2.36: *Cloud shadow mask produced using our masking method (left) and Luo et al. (2008) scheme (right) from (a) 2 April 1994 and (b) 11 February 1999. Cloud shadow pixels are masked yellow for our method and blue for Luo et al. (2008) method; non-cloud shadow and water pixels are masked grey and white respectively.*

In term of quantitative analysis, the cloud mask produced using the cloud analysis was compared with the ACCA scheme through a confusion matrix. Table 2.16 shows confusion matrix between our cloud mask and ACCA scheme for 2 April 1994 based on (a) pixels and (b) percentages. A quite high agreement was obtained for which 87.2% and 99.6% of the pixels were detected as cloud and non-cloud respectively by both methods, giving an overall accuracy of 99.2% and kappa coefficient 0.86.

Table 2.16: *Confusion matrix between our cloud mask and ACCA scheme for 2 April 1994 based on (a) pixels and (b) percentages.*

ACCA scheme (Pixels)

		Not cloud	Cloud	Total
Cloud analysis (Pixels)	Not cloud	507663	1916	509579
	Cloud	2130	13095	15225
	Total	509793	15011	524804

(a)

ACCA scheme (Percent)

		Not cloud	Cloud	Total
Cloud analysis (Percent)	Not cloud	99.58	12.76	97.10
	Cloud	0.42	87.24	2.90
	Total	100	100	100

(b)

Overall Accuracy = 99.23%

Kappa Coefficient = 0.862

For 11 February 1999, 81.2% and 100% pixels were detected as cloud and non-cloud respectively, giving an overall accuracy of 100% and kappa coefficient of 0.79 (Table 2.17). In overall, the Landsat data from 2 April 1994 and 11 February 1999 give an overall accuracy and kappa coefficient of more than 90% and 0.7, indicating a high agreement between the cloud analysis and the ACCA scheme. The difference is mainly due to the more tests used in the ACCA scheme compared to the cloud analysis.

Table 2.17: *Confusion matrix between our cloud mask and ACCA scheme for 11 February 1999 based on (a) pixels and (b) percentages.*

		ACCA scheme (Pixels)		
		Not cloud	Cloud	Total
Cloud analysis (Pixels)	Not cloud	523336	223	523559
	Cloud	282	963	1245
	Total	523618	1186	524804

		ACCA scheme (Percent)		
		Not cloud	Cloud	Total
Cloud analysis (Percent)	Not cloud	99.95	18.80	99.76
	Cloud	0.05	81.20	0.24
	Total	100	100	100

Overall Accuracy = 99.90%

Kappa Coefficient = 0.792

The cloud shadow mask produced using the analysis was compared using the Luo et al. (2008) scheme. Table 2.18 shows the confusion matrix between the shadow analysis and Luo et al. (2008) scheme for 2 April 1994 based on (a) pixels and (b) percentages. For 2 April 1999, approximately 91.3% and 100% pixels were detected as cloud shadow and non-cloud shadow respectively by both methods, giving an overall accuracy of 99.4% and kappa coefficient of 0.86.

Table 2.18: *Confusion matrix between our shadow mask and Luo et al. (2008) scheme for 2 April 1994 based on (a) pixels and (b) percentages.*

		Luo et al. scheme (Pixels)		
		Non-cloud shadow	Cloud shadow	Total
Cloud shadow analysi	Non-cloud shadow	511432	965	512397
	Cloud shadow	2323	10084	12407
	Total	513755	11049	524804

		Luo et al. scheme (Percent)		
		Non-cloud shadow	Cloud shadow	Total
Cloud shadow analysi	Non-cloud shadow	99.55	8.73	97.64
	Cloud shadow	0.45	91.27	2.36
	Total	100	100	100

Overall Accuracy = 99.37%

Kappa Coefficient = 0.857

For 11 February 1999, approximately 81 % and 100% pixels were detected as cloud shadow and non-cloud shadow respectively by both methods, giving an overall accuracy of 99.4% and kappa coefficient of 0.845 (Table 2.19). For dates, the overall accuracy and kappa coefficient was more than 90% and 0.8 respectively, indicating a high agreement between the cloud analysis and the Luo et al. (2008) scheme.

In overall, the cloud and cloud shadow analysis give a high agreement with the ACCA and the Luo et al. (2008) scheme respectively. Subsequently, the masked Landsat data from 11 February 1999 will be used as the main data in classification analysis in Chapter 3.

Table 2.19: *Confusion matrix between our shadow mask and Luo et al. (2008) scheme for 11 February 1999 based on (a) pixels and (b) percentages.*

		ACCA scheme (Pixels)		
		Non-cloud shadow	Cloud shadow	Total
Cloud shadow analysi	Non-cloud shadow	519632	961	520593
	Cloud shadow	154	4057	4211
	Total	519786	5018	524804

		ACCA scheme (Percent)		
		Non-cloud shadow	Cloud shadow	Total
Cloud shadow analysi	Non-cloud shadow	99.97	19.15	99.20
	Cloud shadow	0.03	80.85	0.80
	Total	100	100	100

Overall Accuracy = 99.79%

Kappa Coefficient = 0.878

2.7 Summary and Conclusions

1. Haze has a lower standard deviation and less reflective than cloud; as haze gets severe, it scatters more solar radiation and eventually becomes as reflective as cloud. Hence, very thick haze has standard deviation and reflectance similar to cloud.

2. Spectral analysis based on MODIS scheme is the most suitable for Malaysia due to allowing the optimal used of its rich bands.

3. Cloud masking using MODIS analysis over Malaysia shows a comparable outcome with climatological observations.

4. When applied to two scenes of Landsat data, the cloud and shadow analysis shows a high agreement with ACCA and Lou et al. scheme respectively.

Chapter 3

Land Cover Classification using Remote Sensing Data

3.1 Introduction

The primary objective of this thesis is to assess the effects of haze on our ability to recover information about land cover and land use, and to develop and test methods to reduce its negative impact. Our particular interest is in mitigating the effects of haze on land cover classification, though the outcome is also relevant to other remote sensing applications, e.g. precision farming, etc. A number of land covers in Malaysia are considered, involving those of commercial and non-commercial values, e.g. oil palm, rubber, coconut, industry, forest, urban, industry, etc. Such efforts are important for realising the Malaysian government's vision in preparing Malaysia to be a fully developed country by the year 2020 (Malaysian Prime Minister Office 2010).

In order to quantify the effects of haze and our ability to remove it, we therefore need to define a set of classification methods and performance criteria against which to measure these effects and to assess how they are changed by the correction methods described later.

A large number of classification methods are available, and a brief review is given in Section 3.2. In this review, we describe the main features of the methods, but our principal aim is to select the methods most appropriate to the studies of haze in the later chapters. Our criteria for this selection include:

- simplicity, i.e. the practicality of using a large amount of data. This should involve a smaller number of procedures but should produce reasonably accurate and standard results,
- the ability to select important land covers with an acceptable accuracy, i.e. each pixel will be assigned to the correct land cover on the ground – the performance

of the method should not be easily affected by factors such as the complexity of land covers, topographic conditions, etc. and

- objectivity, i.e. not involving tuning by a user to improve performance – the generated classification works straight away without needing any adjustment in terms of the number of classes, training pixels, etc.

In practice, these criteria lead us to consider the use of Maximum Likelihood (ML), which is a supervised method (Section 3.3). In order to facilitate the use of this method, we can analyse its behaviour from a single image from 11 February 1999 (Section 3.4). This image contains the main land covers of Malaysia and has clear sky conditions (free from haze and little cloud cover), and therefore meets the purpose of our study, i.e. to provide a base map for use in studying the effects of haze on land cover classification and how this can be corrected (i.e. does not involve change detection).

A critical issue for classification is accuracy and in Section 3.4 we discuss how this can be defined and how we can measure it, given the available satellite and ground data. Since this is the fundamental issue for the later assessment of the effects of haze and their correction, we will provide an extended analysis of the suitability of our data in order to arrive at meaningful estimates of accuracy

This analysis in this chapter serves several important purposes, viz. to classify the land covers, assess classification accuracy, relate the spectral correlation with the classifications of the land cover types, investigate the roles of covariance and mean structure in separating different classes and investigate the decision boundary of the classes. Section 3.5 summarises our findings and provides the context for the haze analysis in Chapters 4 and 5.

3.2 Literature Review

There are thousands of papers on land classification, so in this review we will focus on two of the most important issues, viz. classification methods and classification accuracy (Jensen 1996; Lillesand et al. 2004; Lu and Weng 2007). Studies on such issues have actively carried out in many parts of the world but this is not the case for the tropics and countries like Malaysia.

3.2.1 Classification Methods

Classification approaches can be grouped in several ways, such as supervised and unsupervised, parametric and non-parametric, hard and soft (i.e. fuzzy) classification or per-pixel, subpixel and per-field (Mather 2004; Canty 2006; Lu and Weng 2007). For convenience, we will group classification approaches as per-pixel, subpixel and per-field.

Per-pixel Classification

This is the oldest and most frequently used approach; it ensures that each pixel within an image is assigned to a class. Per-pixel classification algorithms can be supervised or unsupervised. Supervised classification is knowledge-driven, while unsupervised classification is data-driven, i.e. the former uses the knowledge about the study area in order to classify it, while the latter uses the knowledge to label the clusters to land covers after the clustering processes end.

In supervised classification, land cover classes are defined and reference data are used as training samples. The signatures generated from the training samples are used to train the classifier in classifying the satellite data into a thematic map. In unsupervised classification, clustering-based algorithms are used to partition the image into a number of spectral classes based on the statistical information inherent in the image. Since no prior definition of the classes is used, the users are responsible for labelling and merging the clusters into meaningful classes. Examples of supervised classification classifiers are ML, minimum distance and Mahalanobis distance for those using parametric classifiers (e.g. those assuming the data has a Gaussian distribution with parameters, viz. the covariance matrix and mean vector, estimated from training samples), while parallelepiped, neural networks, decision tree classifiers and support vector machines use non-parametric classifiers (i.e. they do not make any assumptions about the data and do not use any parameters to calculate cluster separation). Examples of unsupervised classification classifiers are ISODATA and K-means. In land cover mapping, per pixel classification based on supervised methods is often preferred to unsupervised methods.

The parallelepiped classifier, known as the 'box decision rule', is often considered to be the simplest supervised algorithm (Campbell 2002). This algorithm makes use of the ranges of values within the training data to define regions within a multidimensional data space. The Mahalanobis distance uses statistics for each class but assumes that all class covariances are equal. All pixels are classified to the closest region of interest (ROI) class, depending on the distance threshold specified by users; some pixels may be unclassified if they do not meet the threshold (Richards 1999). The minimum distance classifier employs the central values of the spectral data that forms the training data set to classify pixels. The neural network classification is a self adaptive method that is able to estimate the posterior probabilities, which provide a basis for establishing the classification rule (Zhang 2000). A decision tree classifier makes use of a series of binary decisions to determine the correct category for each pixel. The decisions can be based on any available characteristic of the dataset. The support vector machine method involves a learning process based on structural risk minimisation, which can minimise classification error without the need to assume data distribution (Mountrakis et al. 2011). It is capable of handling data with a limited training sample. However, it often linked to high computational requirements and processing times. An ML classifier is a powerful classification technique based on the maximum likelihood decision rule and depends on the quality of training samples, which are usually determined based on ground-verified land cover maps and knowledge of the area. Due to its practicality, and its ability to discriminate between land covers effectively, objectivity and easy availability through the use of most image processing software (Lu and Weng 2007) (e.g. ENVI, ERDAS and PCI Geomatics), numerous remote sensing data users worldwide, including those in

Malaysia, use ML to classify land covers in their projects or research studies (Fuller et al. 2005; Fuller et al. 1994).

Subpixel Classification

Subpixel classification approaches have been developed as a better solution for mixed pixels problems, i.e. the existence of more than one class in a pixel, especially when coarse spatial resolution data are used. Such approaches require a fuzzy representation, in which each pixel is composed of multiple and partial memberships of all candidate classes. The most popular approaches are the fuzzy-set technique (Zhang and Kirby 1999; Zhang and Foody 2001) and spectral mixture analysis (SMA) classification (Rashed et al. 2001; Lu et al. 2003).

In SMA, each pixel is evaluated as a linear combination of a set of endmember spectra. The output is in the form of fraction images, with one image for each endmember spectrum, representing the area proportions of the endmembers within a pixel. It has been demonstrated that SMA is helpful for improving classification accuracy and is important for improving area estimation of land use and land cover classes based on coarse spatial data. However, its main shortcoming is that it is rather difficult to assess the accuracy of subpixel classification (Lu and Weng 2007), which cannot be measured in a straightforward way using the confusion matrix technique (i.e. each pixel being associated with one class), which will be used to investigate the effects of haze in Chapter 4. Moreover, in the Malaysian context, the subpixel classification approach is less preferable due to the constraints in expertise, facilities and cost. Remote sensing applications (e.g. land cover mapping, precision farming) are still rely heavily on per-pixel classification.

Per-field Classification

Per-field classifiers are designed to deal with the problem of environment heterogeneity; i.e. high spectral variation within the same land cover class. They make use of land parcels (i.e. known as 'fields') as individual units. This is also known as a segmentation approach. It requires the use of a geographical information system (GIS) than can integrate both raster (i.e. satellite data) and vector data. The vector data are used to subdivide an image into parcels, on which classification processes are based on. This avoids interclass spectral variations (Lu and Weng 2007). Nevertheless per-field classifications are frequently affected by factors such as the spectral and spatial properties of remote sensing data, the size and shape of the fields, the definition of field boundaries and the land covers chosen (Janssen and Molenaar 1995; Lu and Weng 2007). In addition, difficulties in handling vector and raster data can affect the use of the per-field classification approach. Another per-field approach is to use object-oriented classification, which does not require the use of vector data (Lu and Weng 2007). This involves two consecutive stages, i.e. image segmentation and classification. The former merges pixels based on objects and the latter classifies the objects rather than the pixels. The most commonly used object-oriented classification is eCognition (Benz et al. 2004; Wang et al. 2004). However, the main shortcoming of this method is that land surface objects are often difficult to acquire (Smith and Fuller 2001). Also, it is not relevant in the Malaysian context and for achieving the aims of this thesis.

3.2.2 Classification Accuracy

Classification accuracy is one o the key parameters required to judge the quality of land cover classification and can be defined as the degree to which the derived image classification conforms to the 'truth' (Campbell 2002). Two of the most important components in accuracy assessment are analysis of reference data and sampling design (Stehman 1999).

Analysis of Reference Data

Studies have shown that the most widely used technique to analyse reference data is to use a confusion or error matrix (Congalton 1991). A confusion matrix works by comparing classification result with reference information, while accuracy is conveyed in terms of percentage of overall classification accuracy, producer accuracy and user accuracy (Congalton 1991). The acceptable of overall accuracy is 85%, with no class less than 70% accurate (Thomlinson et al. 1999). Kappa statistics have been used as early as the 1980s as an additional classification accuracy measure to compensate for chance agreement (Congalton 1991). In Chapter 4 and 5, we will show that the confusion matrix technique is very useful in investigating the effects that haze has on land cover classification.

Since then, researchers worldwide have been heavily relying on these measures (i.e. measures of overall, producer and user accuracy) due to their robustness and simplicity in assessing the quality of land cover classifications. Hence, not many promising assessment techniques have been developed. However, in 2001, Koukoulas and Blackburn proposed a way of calculating the classification success index (CSI) using a confusion matrix that takes into account errors of omission (producer accuracy minus one) and commission (user accuracy minus one). CSI was initially proposed for use in studies of forested environments and especially in natural or semi-natural landscapes, where the variety of species and spatial heterogeneity makes land cover classification complicated. An individual classification success index (ICSI) was established to account for the classification success of a specific class, while a group classification success index (GCSI) was used to measure classification success for the main classes in the study area. An index of 0.8 was considered to be adequate for successful classification. Koukoulas and Blackburn (2001) claim that their technique is an important research tool rather than just an indicator of the errors that accumulate during the classification process. Our study will make use of CSI and ICSI as extended measures for assessing the performance of classifications.

Sampling Design

The collection of reference pixels can be performed using interpretation of higher resolution imagery or hardcopy maps with adequate ground truth knowledge of the study area (San Miguel-Ayanz and Biging 1996) and on-site collection using a global positioning system (GPS) (Lillesand et al. 2004). Due to logistics and time, the former is more preferable than the latter. When selecting samples within study area, the minimum number of samples required per class is 50. If the types of land use and land cover exceed12, the minimum number of samples needs to be increased to 75 or 100 (Congalton 1991; Lillesand et al. 2004). The samples can be in form of pixels, clusters of pixels, or polygons. Sampling designs frequently considered include simple random sampling, systematic sampling, stratified sampling and cluster sampling (Congalton 1991; Stehman 1997).

In *random sampling*, locations for sample collection are selected randomly, using a random number generator or a table of random digits to ensure that every member of the population has an equal chance of being selected for the sample (Stehman 2000). This method ensures that the allocation of sample locations is not biased and does not require any prior information about the field site. The main problem with simple random sampling is that it tends to undersample classes with small areas. In *systematic sampling*, the chosen samples are distributed in a regular pattern, such as a grid. The starting pixel is chosen randomly. Sampling is then carried out in every *K*th pixel in both horizontal and vertical directions from the starting pixel for a square grid. A different sampling interval may be chosen for the horizontal and vertical directions to form a rectangular grid. The advantages of this technique are that it is simple and has good spatial coverage. The main drawback of systematic sampling is the absence of an unbiased estimator of the variance (Gallego 2004). In *stratified random sampling*, a simple random sample of pixels is selected for each stratum (Stehman et al. 2007). The strata are usually land cover classes and the size of samples collected from each stratum takes into account the size of that stratum. This is the most commonly employed sampling design. However prior information about the land covers within the study area is required. This can normally be obtained from maps and satellite data. *Cluster sampling* involves taking a group of samples from a predetermined number of random locations. It employs two types of sampling unit, i.e. a primary sampling unit and a secondary sampling unit. The cluster often consists of a block of pixels (e.g. 3 by 3 or 5 by 5). The disadvantage of cluster sampling is that the standard error formulae are more complex than those required for simple random sampling, due to the need to account for the lack of independence among the secondary sampling units within a cluster (Stehman 1997).

3.2.3 Implementation of Land Cover Classification

Studies of classification of remote sensing data have long been carried out by numerous researchers worldwide, with more efforts made regionally than globally. Many regional studies have been carried out in places such as Europe (Thompson et al. 1998) and America (Jia and Richards 1994; Guerschman et al. 2003; Low and Choi 2004) due to the existence of up-to-date remote sensing facilities as well as ground truth information. There is also an increasing interest in carrying out such studies in climate-affected regions such as Africa (Wang et al. 2010) and highly populated regions such as India (Thenkabail et al. 2005) and China (Liu et al. 2011). Nonetheless, not much effort has been expended in tropical countries such as Malaysia (Baban and Yusof 2001; Ismail and Jusoff 2008), despite the recent promising developments in remote sensing capabilities in such countries (Yusoff et al. 2002).

Two studies that were undertaken in Malaysia are cited here. Baban and Yusof (2001) used ML classification to map landuse/cover distribution on a mountainous tropical island, Langkawi. An unnamed unsupervised classification using Landsat bands 3, 4 and 5 was initially performed to aid the selection of the training pixels for the study area. ML classification was then carried out on eight classes, namely, inland forest, mangrove forest, rubber, paddy fields, mixed horticulture, grassland, urban and water. The overall classification accuracy was 90% with individual class accuracy ranging from 74% for rubber to 100% for paddy. Another study was conducted by Ismail and Jusoff (2008), where ML classification was used to classify five forms of land use and land covers in Pahang, Malaysia, viz. primary forest, logged over forest, agriculture crops, water and cleared lands. The classification accuracy of the classified images was assessed by comparing the classes with the corresponding reference pixels (i.e. obtained using visual interpretation of satellite data and land use maps) by using a confusion matrix technique. The result was acceptable for both studies. Here, the reference pixels were obtained using stratified random sampling approach based on visual interpretation of satellite data and land use maps. The overall accuracy of the classification was 89% with a kappa coefficient of 0.8.

In Great Britain, one of the earliest initiatives for national land cover mapping was the Land Cover Map of Great Britain (LCMGB), initiated in 1990 by the UK Institute of Terrestrial Ecology. LCMGB raster dataset was the first comprehensive land cover of Britain to be mapped using satellite data. It was produced using a per pixel supervised ML classification of Landsat TM consisting of 25 land cover classes (Fuller et al. 1994). Later, its updated version, LCM2000, was produced using per-pixel supervised ML classification, combined with ancillary geographical data and containing 26 land cover classes. The accuracy of LCM2000 and LCMGB is assessed using a confusion matrix in comparison with field surveys, selected based on a stratified random sampling scheme, where the overall levels of accuracy obtained were 85%, and 80% respectively (Fuller et al. 2003).

Thompson et al. (1998) compared ML classification and ISODATA clustering methods for coasts and river corridors along the East coast of England using all 14 bands of Compact Airborne Spectrographic Imager (CASI). The 12 classes considered were water, bare earth/river banks, urban, arable, pasture, haycut, lowland rough vegetation, deciduous wood, coniferous wood, upland grass, heather/grass mix, heather, burnt heather, upland bog and bare rock. Training and reference pixels were sampled based on visual interpretation of satellite data itself and land cover maps. However, details of the sampling approaches used were not stated. The results are presented as classification maps, confusion matrices and feature space images. They show that ML classification produced excellent results in separating inland cover types while ISODATA clustering was considered to be an acceptable alternative, due to it involving less user input rather than dependence on a priori information in the study area.

In the USA, Paola and Schowengerdt (1995) carried out a detailed comparison of the back-propagation neural network and ML classification, using Landsat TM bands 1, 2, 3, 4, 5 and 7, for urban land use in Tucson and Oakland in California. 13 classes were considered, viz. tarmac, building, grass, foothills natural vegetation, sand, desert scrub, bare soil, urban residential, asphalt, riparian vegetation, dense urban and shaded foothills natural vegetation. For each class, training pixels were extracted from a training region (i.e. about the same size) and were defined through visual interpretation of the Landsat image and knowledge of the study area. Reference pixels were determined using the same approach and a confusion matrix was then employed for accuracy assessment. Analyses were conducted in terms of classification accuracy, class mean, class standard deviation, density plots and decision space analysis. It was found that the neural network method could classify areas with highly mixtures of land cover compared more effectively than the ML, However, this method did consume much more computing time.

Low and Choi (2004) performed a hybrid classification for land use and land cover mapping by using Landsat 7 ETM+ data over the Atlanta metropolitan area, in the largest city of the state of Georgia, USA. The land use and land cover classes within the study area are urban/industry, settlement, cleared land, crop land, forest and water. In their approach, ISODATA clustering was initially used to aid the selection of training pixels, followed by a supervised fuzzy classification. Accuracy assessment was carried out using a confusion matrix with reference pixels based on the visual interpretation of aerial photographs. No details concerning the sampling approach were given. The hybrid classification was compared with: (a) ISODATA clustering, (b) ML classification and (c) supervised fuzzy classification. The hybrid classification was found to be slightly better in terms of classification accuracy than the ISODATA clustering, but the ML and supervised fuzzy classification produced much lower levels of accuracy.

In Japan, Yoshida and Omatu (1994) used a neural network approach, i.e. a back-propagation algorithm, to classify land use and land cover in Tokushima city using Landsat TM bands 3, 4 and 5, and compared their results with those obtained by ML classification. Nine classes were considered, viz. the dark part of the forest, bare land, inhabited districts, roads, forests or grassy places, rivers or seas, farms, clouds and shadows of clouds. In order to select training pixels, Kohonen's self-organizing feature map and geographical information were used. A confusion matrix was subsequently used to assess the classification accuracy. However the approach to collect the reference pixels used in the confusion matrix was not stated. The neural network classifications show a better overall accuracy compared to ML classification, but more effort and time were required, particularly in determining the number of output layers in Kohonen's method and the categories and numbers of neurons at the hidden layer by the BP algorithm.

In Turkey, Erbek et al. (2004) examined the performance of two artificial neural network classifiers for land use classification using Landsat TM bands 2, 3 and 4, viz. multi layer perceptron and learning vector quantization. The study area was near Istanbul (approximately 270 km^2), a rapidly growing metropolis with a wide range of land use activities. Separate sets of training and reference pixels were selected, based on visual interpretation of the Landsat data and aerial photographs of the study area. However, the sampling approach used was not stated. The performance of these classifiers was compared to the ML classification for six classes, viz. green area, bare soil, urban areas, water, highway and industrial areas. In terms of overall accuracy and its Kappa coefficient, the ML classification was better than the learning vector quantization neural network but worse than the multi layer perceptron neural network classification. However, Erbek et al. (2004) claim that neural network classification using both classifiers required a much longer time than ML classification.

In East Africa, Otukei and Blaschke (2010) assessed land cover change in the Pallisa District, Eastern Uganda from 1986 to 2001 using Landsat TM and ETM+ datasets. They employed several classification methods, viz. decision trees, support vector machines and ML classification algorithms and compared their classification accuracy. Training and reference pixels were selected, based on knowledge of the study area as well as visual interpretation techniques by which subsequently classification accuracy was evaluated using a confusion matrix. However, the sampling techniques used were not discussed. The highest classification accuracy and Kappa coefficient were shown by the decision tree method, followed closely by the ML and support vector machine methods. No effort was made to further analyse the classifications using other means, e.g. band correlations and decision boundaries.

In China, Liu et al. (2010) carried out a mixed-label analysis classification, based on the k-nearest neighbour (K-NN) using a nonparametric regression algorithm, and compared it with ML, neural network and minimum distance classifications. The classification analysis was carried out using simulated and real data (i.e. Landsat TM) of Dongguan in the Pearl River Delta, China (i.e. covering 296 x 299 pixels) and involved six classes, i.e. urban, forest, water, grass, agriculture and developing land. Training and reference pixels were selected using random stratified sampling based on visual interpretation of high-resolution satellite data and collection of ground truth data. The mixed-label analysis classification was found to be producing the highest overall accuracy and Kappa coefficient, followed by the neural network, ML and minimum distance classifications. As with ML, the accuracy of the mixed-label analysis classification was mainly influenced by the quality of the training data. However, the major setback was that it required longer than other methods.

Liu et al. (2011) used an integrated fuzzy and ML classification method, known as fuzzy topology-based ML classification, to classify land use and land cover in Xuzhou City, China. Landsat bands 1 – 5 and 7 were used to classify the study area into four classes, i.e. building, woodland, water and farmland. By using this method, each class in the image is treated as a fuzzy set in a fuzzy space to give a natural representation of objects. The fuzzy class is then decomposed into two parts; an interior and a boundary. The interior represented the core of a class and the boundary represented an overlapping area between classes. The two parts were eventually combined by using the properties of spatial connectivity in fuzzy topology. Training and reference pixels were selected randomly based on visual interpretation of satellite data and land use maps of the study area. Accuracy assessment was then performed using a confusion matrix. The fuzzy topology-based ML yielded higher classification accuracy and coefficients than the conventional ML classification. Liu et al. (2011) assumed pixel uncertainty to be one of the main sources of error in such classification. However, no further discussion was carried out concerning this issue.

In Israel, Rozenstein and Karnieli (2011) compared several land use and land cover classification approaches using Landsat TM bands 1 – 5 and 7, in the northern Negev. Six classification methods were employed: ISODATA, integration of ISODATA and DSS (decision support systems), ML classification, integration of ML classification and DSS, hybrid (combination of ISODATA and ML) classification, integration of hybrid classification and DSS, where the classes involved were urban or built-up land, agricultural fields, rangeland and mixed rangeland, forest, water bodies such as reservoirs and barren land. Training pixels were obtained by digitising polygons on high-resolution orthophotos of Israel, and then projecting them onto the satellite image. Reference pixels were selected by using stratified random sampling based on the ISODATA cluster map. The integration of hybrid classification and DSS yielded the highest classification accuracy and Kappa coefficient. Rozenstein and Karnieli (2011) remarked that the incorporation of DSS could increase the classification accuracy by 5 to 10%. However, this depends on the availability of quality ancillary data. This is often a problem, particularly when mapping large areas, especially in developing countries.

A quite different study was carried out by Wilkinson (2005) who examined a compilation of 15 years of peer-reviewed experiments on satellite data classification to assess the degree of progress being made in land cover mapping through developments in classification algorithms and systems approaches (e.g. postclassification analysis). The results of over 500 reported classification experiments were quantitatively analysed in terms of types of classifier (neural network and nonneural approaches), classification accuracy, the number of classes, and resolution of the satellite data and test areas. The outcome of the study reveals that no significant upward trend was shown in the hundreds of experiments analysed in the study over the past 15 years. It was concluded that improvements in the techniques are too small to have had any appreciable effect on classification. From this, we can infer that the performances of conventional classifiers, such as ML, are as effective as advanced classifiers, such as artificial neural networks, fuzzy-sets and expert systems.

Hence, in our study, we employed ML classification using Landsat data for Klang in Selangor, Malaysia. ML classification is used as it is still the preferable classification method in national land cover mapping (e.g. LCMGB and LCM2000) (Fuller et al. 1994). The use of ML is also justified by the fact that recently developed methods do not show significant improvement in classification accuracy in determining the quality of land cover map (Wikham 2004). The choice of Landsat data is due to the fact that it is still a preferable data for national land cover mapping (e.g. LCM2000 and NLCD2001) and local applications (e.g. Low and Choi (2004), Erbek et al. (2004), Otukei and Blaschke (2010), Liu et al. (2010), Liu et al. (2011) and Rozenstein and Karnieli (2011)). An Accuracy assessment was carried out by means of the well known confusion matrix technique (Wilkinson 2005; Liu et al. 2007) and reference data were selected using stratified random sampling (Jensen 1996; Lillesand et al. 2004). Subsequently, the performance of ML was measured by making use of classification accuracies (Wilkinson 2005; Song et al. 2001). This was further verified by the assessment technique proposed by Koukoulas and Blackburn (2001). Other quantitative analyses, e.g. band correlations and decision boundaries (Paola and Schowengerdt 1995), were also considered. Klang in Selangor Malaysia was selected as the study area due to having important land covers in Malaysia (Baban and Yusof 2001), and also because the area is not too complex, therefore suitable for use in haze removal analysis in subsequent chapters (Chapters 4 and 5).

3.3 General Classification Concepts

In remote sensing, classification is the process of assigning a pixel to a particular type of land cover. Classification uses data (typically a measurement vector or feature vector ω) from a space borne or airborne acquisition system. It aims to assign a pixel associated with the measurement ω at position $\mathbf{x}$ to a particular class i, where $1 \leq i \leq M$ and M is the total number of classes. The classes are defined from supporting data, such as maps and ground data for test sites. Two types of classification are commonly used, supervised and unsupervised. Supervised classification starts from a known set of classes, learns the statistical properties of each class and then assigns the pixels based on these properties. Unsupervised classification is a two-step operation of grouping pixels into clusters based on the statistical properties of the measurements, and then labelling the clusters with the appropriate classes.

As supervised classification classifies pixels based on known properties of each cover type, it requires representative land cover information, in the form of training pixels. Signatures generated from the training data will be in a different form, depending on the classifier type used. For ML classification the class signature will be in the form of class mean vectors and the covariance matrices. However, the disadvantage is that the derived classes may not be statistically separable.

On the other hand, in terms of unsupervised classification, the clustering process produces clusters that are statistically separable, giving a natural grouping of the pixels. Landcover information is then used in the following labelling process where clusters are assigned to classes based on the available landcover information. This has the disadvantages that (1) a cluster may represent a mixture of different landcover types and (2) a single landcover may be split into several clusters. Furthermore, the assignment of clusters to classes (the labelling process) requires manual input using available knowledge, and needs to be carefully performed after the clustering, in order to correctly label the clusters.

The probability distributions of the data may take a variety of forms, but very frequently they are assumed to be Gaussian (Normal). When each class obeys a multivariate normal distribution for N spectral dimensions (i.e. the number of bands used), we can define the probability that feature vector a $\boldsymbol{\omega}$ occurs in a specified class i as:

$$P(\omega \mid i) = (2\pi)^{-\frac{N}{2}} \left(|C_i|\right)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\omega - \mu_i)^t C_i^{-1} (\omega - \mu_i)\right) \qquad \text{... (3.1)}$$

where,

$$C_i = \left\langle (\omega_j - \mu_i)(\omega_j - \mu_i)^t \right\rangle \approx \frac{1}{Q_i} \sum_{j=1}^{Q_i} \left\{ (\omega_j - \mu_i)(\omega_j - \mu_i)^t \right\}$$

where μ_i is the class mean vector, C_i is the class covariance matrix for class i, Q_i is the number of pixels in class i, ω_j is the feature vector of the j^{th} pixel and $|.|$ is determinant. This assumption is likely to be suitable for data that comes directly from spectral band measurements, but should not be used if the feature vector contains more general types of data, e.g. band ratios, without first testing its validity.

3.3.1 Maximum A Posteriori and ML Classification

The most commonly used supervised classification method is ML. It is based on a more general approach derived from Bayes' theorem, which states that the a posteriori distribution $P(i|\omega)$, i.e., the probability that a pixel with feature vector ω belongs to class i, is given by:

$$P(i \mid \omega) = \frac{P(\omega \mid i)P(i)}{P(\omega)} \qquad \text{... (3.2)}$$

where $P(\omega|i)$ is the likelihood function, $P(i)$ is the a priori information, i.e., the probability that class i occurs in the study area and $P(\omega)$ is the probability that ω is observed, which can be written as:

$$P(\omega) = \sum_{i=1}^{M} P(\omega \mid i)P(i) \qquad \text{... (3.3)}$$

where M is the number of classes. $P(\omega)$ is often treated as a normalisation constant to ensure $\sum_{i=1}^{M} P(i \mid \omega)$ sums to 1. Pixel **x** is assigned to class i by the rule:

$$x \in i \quad \text{if } P(i|\omega) > P(j|\omega) \quad \text{for all } j \neq i \qquad \text{... (3.4)}$$

Maximum a Posteriori (MAP) classification is possible by using Equation 3.6 if we have the prior information P(i). This is the most powerful use of the Bayes Theorem. If we do not know P(i), it is common to assume a uniform prior:

$$P(i) = P(j) \quad \forall \quad i, j \qquad \text{... (3.5)}$$

Hence, P(i) can be neglected and Equation 3.3 becomes:

$$P(i \mid \omega)\, \alpha \, \frac{P(\omega \mid i)}{P(\omega)} \qquad \text{... (3.6)}$$

The absence of prior information is the distinction between ML and MAP classification. Maximising $P(i \mid \omega)$ is equivalent to maximising the likelihood function $P(\omega \mid i)$, i.e. ML:

$$x \in i \quad \text{if} \quad P(\omega|i) > P(\omega|j) \qquad \text{for all } j \neq i \qquad \text{... (3.7)}$$

ML often assumes that the distribution of the data within a given class i obeys multivariate Gaussian distribution. It is then convenient to define the log likelihood (or discriminant function):

$$g_i(\omega) = \ln P(\omega \mid i) = -\frac{1}{2}(\omega - \mu_i)^t C_i^{-1}(\omega - \mu_i) - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(|C_i|) \qquad \text{... (3.8)}$$

Since log is a monotonic function, Equation 3.7 is equivalent to:

$$x \in i \quad \text{if} \quad g_i(\omega) > g_j(\omega) \quad \text{for all } j \neq i\,. \qquad \text{... (3.9)}$$

Each pixel is assigned to the class with the highest likelihood or labelled as unclassified if the probability values are all below a threshold set by the user (Lillesand et al. 2004). The general procedures in ML are as follows:

1. The number of land cover types within the study area is determined.

2. The training pixels for each of the desired classes are chosen using land cover information for the study area. For this purpose, the Jeffries-Matusita (JM) distance can be used to measure the class separability of the chosen training pixels. For normally distributed classes, the JM separability measure for two classes, J_{ij}, is defined as follows (Richards, 1999):

$$J_{ij} = \sqrt{2\left(1 - e^{-\alpha}\right)} \qquad \text{... (3.10)}$$

where α is the Bhattacharyya distance and is given by (Richards, 1999):

$$\alpha = \frac{1}{8}\left(\mu_i - \mu_j\right)^t \left[\frac{\left(C_i + C_j\right)}{2}\right]^{-1} \left(\mu_i - \mu_j\right) + \frac{1}{2}\ln\left(\frac{\left|\frac{C_i + C_j}{2}\right|}{\sqrt{\left|C_i\right|\left|C_j\right|}}\right) \qquad \text{... (3.11)}$$

J_{ij} ranges from 0 to 2.0, where $J_{ij} > 1.9$ indicates good separability of classes, moderate separability for $1.0 \leq J_{ij} \leq 1.9$ and poor separability for $J_{ij} < 1.0$ (ENVI 2006).

3. The training pixels are then used to estimate the mean vector and covariance matrix of each class.

4. Finally, every pixel in the image is classified into one of the desired land cover types or is labelled as unknown.

In ML classification, each class is enclosed in a region in multispectral space where its discriminant function is larger than that of all other classes. These class regions are separated by decision boundaries, where the decision boundary between class i and j occurs when:

$$g_i(\omega) = g_j(\omega) \qquad \text{... (3.12)}$$

For multivariate normal distributions, this becomes:

$$-\frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\mu}_i)^t C_i^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu}_i)-\frac{N}{2}\ln(2\pi)-\frac{1}{2}\ln(|C_i|)- \\ \left(-\frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\mu}_j)^t C_j^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu}_j)-\frac{N}{2}\ln(2\pi)-\frac{1}{2}\ln(|C_j|)\right)=0 \qquad \text{... (3.13)}$$

which can be written as:

$$-(\boldsymbol{\omega}-\boldsymbol{\mu}_i)^t C_1^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu}_i)-\ln(|C_i|)+(\boldsymbol{\omega}-\boldsymbol{\mu}_j)^t C_j^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu}_j)+\ln(|C_j|)=0 \qquad \text{... (3.14)}$$

This is a quadratic function in N dimensions. Hence, if we consider only two classes, the decision boundaries are conic sections (i.e. parabolas, circles, ellipses or hyperbolas).

3.4 Methodology

In this study, ML classification was applied to our study area (Klang in Selangor, Malaysia), which covers approximately 540 km^2 within longitude 101° 10' E to 101°30' E and latitude 2°99' N to 3°15' N. The satellite data comes from bands 1, 2, 3, 4, 5 and 7 of Landsat-5 TM dated 11th February 1999, while the supporting data is a land cover map from October 1991 of the study area. The map, with a 1:50,000 scale, was produced by ARSM using SPOT data dated 26 February and 10 June 1991 and was supplemented by Landsat data (i.e. date not stated) and a ground truth survey carried out on October 1991. Although there is a relatively lengthy time gap between the Landsat data and the landcover map, the study area is known to be a non-intensively developing zone, with no major changes in land cover.

Visual interpretation of the Landsat data (Figure 3.1(b)), aided by the land cover map (Figure 3.1(a)), was carried out and 9 main classes were identified, viz. coastal swamp forest, dryland forest, oil palm, rubber, industry, cleared land, urban, coconut and bare land (Figure 3.1(b)) (sediment plumes refer to off-shore sediment from erosion caused by natural and man-made alteration of the landscape (Gupta, 1996)).

Coastal swamp forest covers most of Klang Island (in the south-west of the image) and coastal regions in the south-west of the scene. Most of the dryland forest can be recognised as a large straight-edged region in the north-east. Oil palm is the most important commercial crop and can be found in the centre towards the north-west, while rubber is unevenly distributed in the north and south-east of the scene. Oil palm plantations, mostly managed by FELDA (the Federal Land Development Authority, Malaysia) are far more abundant than rubber plantations due to higher demand and a better price in the global markets (Simeh and Ahmad, 2001). Urban areas fill the lower middle of the scene, from the coastal region and inland. Industry can be recognised in the brighter patches near the urban areas, especially in the southwest and northeast. The relatively large urban and industry areas reflect the fact that Klang town and Klang port play an important role in stimulating the surrounding areas economically. Cleared land is spread all over the scene and is indicated by line-like shapes and patches of no particular shape. In the ML classification, regions of interest (ROIs) associated with the training pixels for 9 classes of land cover were determined based on the land cover map.

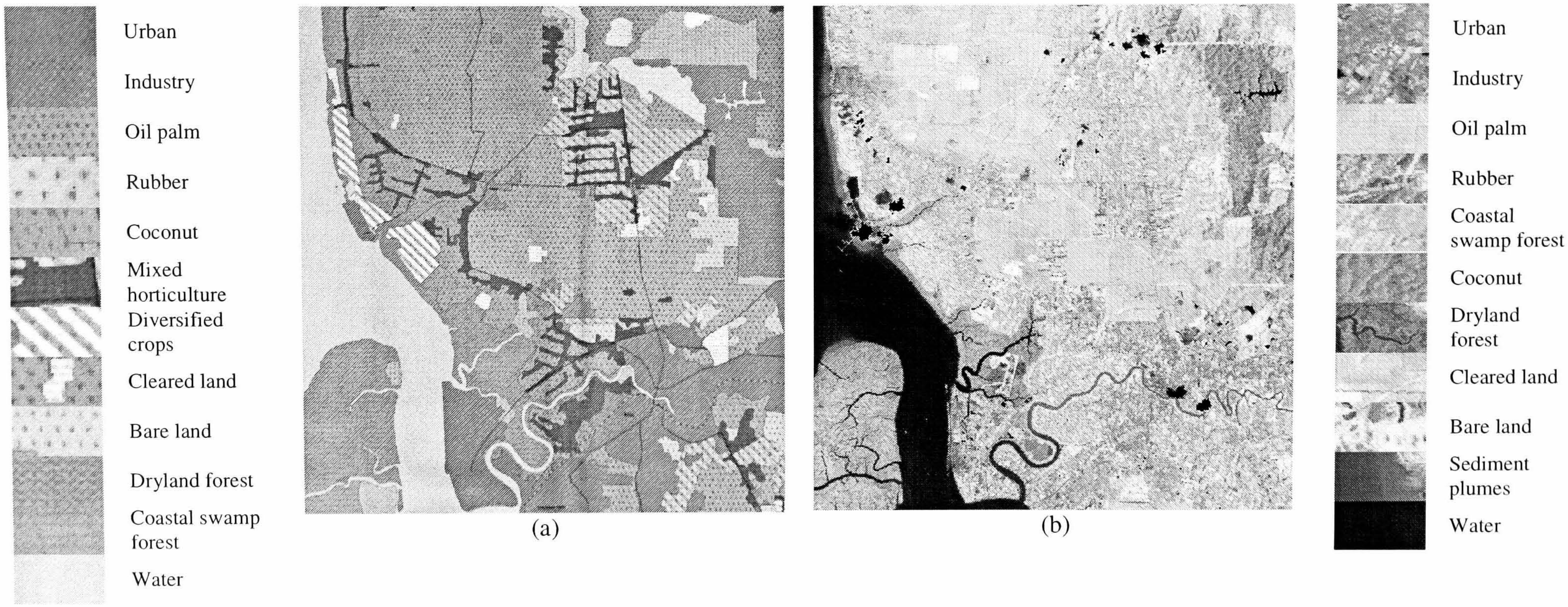
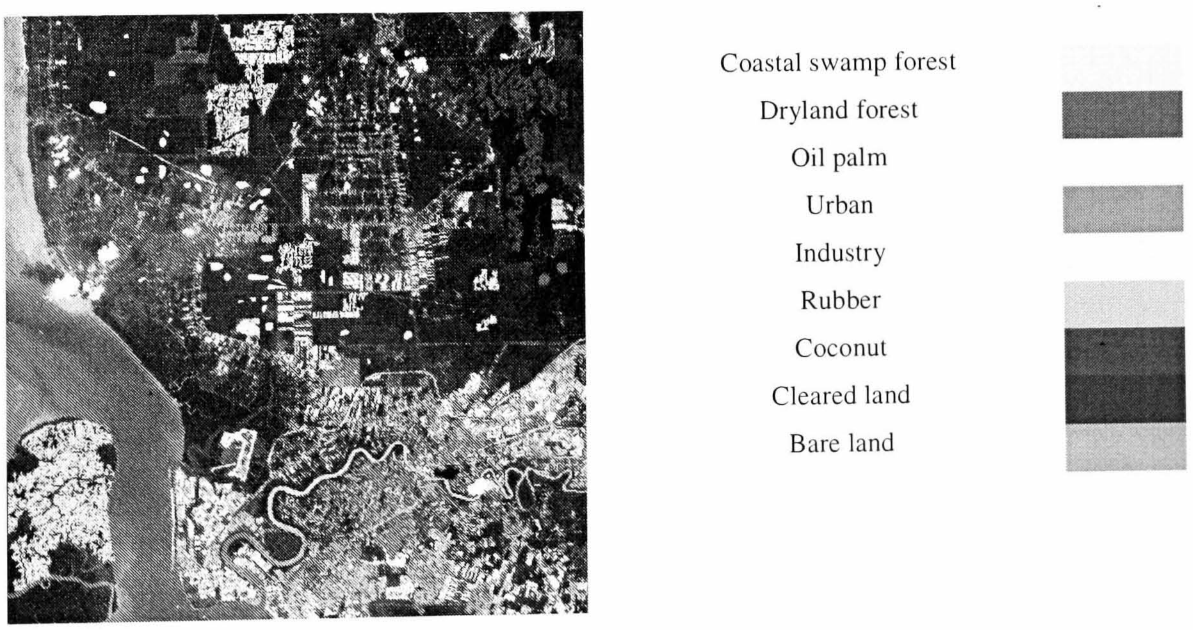


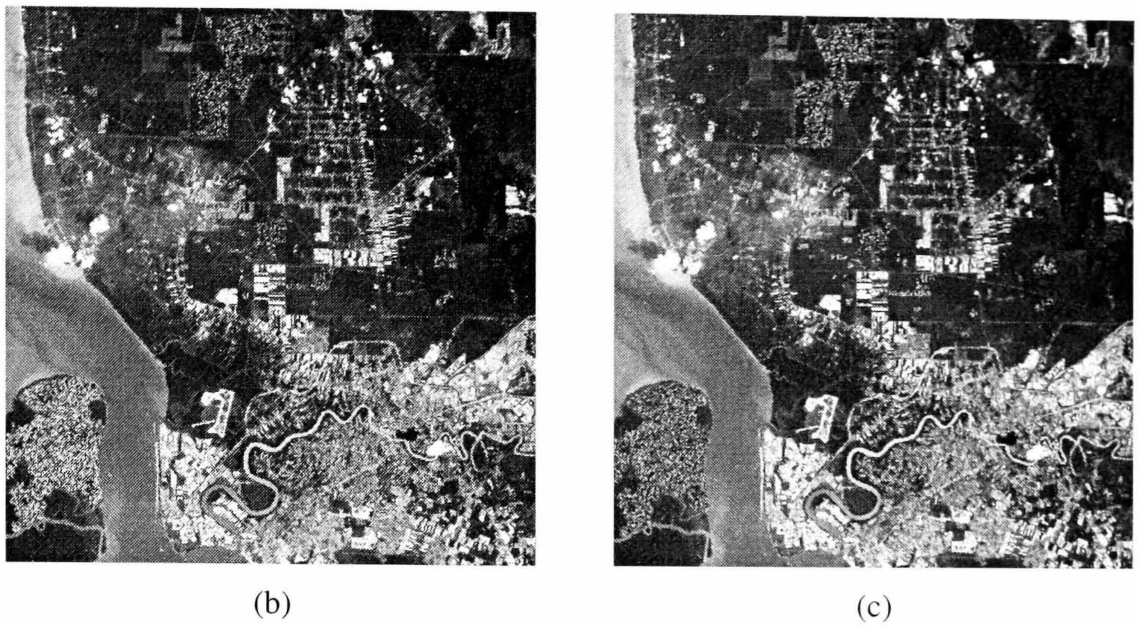
Figure 3.1: *The study area from (a) the land cover map and (b) the Landsat-5 TM with bands 5 4 and 3 assigned to the red, green and blue channels, with cloud and its shadow masked in black.*

3.4.1 ML Classification

Sampling was carried out by means of stratified random sampling technique by making use of built-in functions in the ENVI software. This technique involves dividing the population (the entire classification image) into homogeneous subgroups (the ROI for individual classes) and then taking a simple random sample in each subgroup. The ROI was determined by choosing one or more polygons for each class based on visual interpretation of the land cover map and Landsat data (Figure 3.1). This was assisted by region growing tools from the ENVI software. With the region growing tool, pixels within the polygons were grown to neighbouring pixels based on a threshold, i.e. the number of standard deviations away from the mean of the drawn polygons. Approximately 30% of the pixels within the ROI of each class were selected to be training pixels, using a random sampling technique. Figure 3.2 shows the locations of (a) the original sampling pixels (b) those chosen for training pixels to be used in classification and (c) reference pixels for accuracy assessment. The numbers of training pixels are: rubber (196), coastal swamp forest (4452), dryland forest (1849), oil palm (3148), industry (105), cleared land (375), urban (693), coconut (465) and bare land (94).



(a)



(b) (c)

Figure 3.2: *Locations for (b) training pixels and (c) reference pixels; the colours within the images are associated with the land cover classes within the study area as shown in the colour table in (a).*

The class separability of the chosen training pixels was determined by means of the JM distance (see 3.3.1), which is shown for all class pairs in Table 3.1. Fifty-two pairs have a JM distance of between 1.9 and 2.0, indicating good separability, three from 1.0 to 1.9 indicating moderate separability and none less than 1.0 indicating poor separability. The worst separability, possessed by the oil palm – coconut pair (1.553), was expected since both have very similar spectral characteristics. For each class, these training pixels provide values from which to estimate the means and covariances of the spectral bands used.

Table 3.1: *The separabilities measured by Jeffries–Matusita distance for the training pixels.*

	Coastal swamp forest	Dryland forest	Oil palm	Rubber	Cleared land	Coconut	Bare land	Urban	Industry
Coastal swamp forest	0.000	-	-	-	-	-	-	-	
Dryland forest	2.000	0.000	-	-	-	-	-	-	
Oil palm	2.000	1.985	0.000	-	-	-	-	-	
Rubber	2.000	1.942	2.000	0.000	-	-	-	-	
Cleared land	1.999	1.997	1.952	1.981	0.000	-	-	-	
Coconut	1.984	1.999	1.553	2.000	1.965	0.000	-	-	
Bare land	2.000	2.000	2.000	2.000	1.997	2.000	0.000	-	
Urban	2.000	2.000	2.000	1.999	1.703	2.000	2.000	0.000	
Industry	2.000	2.000	2.000	2.000	1.930	2.000	2.000	1.955	0.000

The outcome of ML classification, after assigning the classes with suitable colours, is shown in Figure 3.3: coastal swamp forest (green), dryland forest (blue), oil palm (yellow), rubber (cyan), cleared land (purple), coconut (maroon), bare land (orange), urban (red) and industry grey. Clouds and their shadows are masked black, while non-land classes, i.e. water and sediment plumes are masked white. Although being similar, coastal swamp forest and dryland forest can be clearly seen in the south-west and north-east of the classified image, as indicated by the land cover map (see Figure 3.1). Oil palm and urban dominate the northern and southern parts respectively. Rubber appears as scattered patches that mostly are surrounded by oil palms. Coconut can be seen in the coastal area in the north-west of the image. Industry mostly occupies areas near the Klang port, in the south. A quite large area of bare land can be seen in the east, while cleared land can be seen mostly in the north, south and south-east of the image. The class areas in terms of percentage (with respect to the whole image) and square kilometres are given in Table 3.2. The three biggest classes are oil palm (133 km^2), cleared land (103 km^2) and urban (37 km^2), while the smallest class is bare land (8 km^2). The classified land areas add up to a total of 453 km^2, i.e. 84% from the whole image.

Table 3.2: *Classes determined by ML classification with corresponding areas in percentage and square kilometres.*

Class	Area (%)	Area (km^2)
Coastal swamp forest	6.5	35.3
Dryland forest	5.4	29.4
Oil palm	24.5	132.6
Rubber	3.3	17.8
Cleared land	19.2	103.5
Coconut	6.9	37.3
Bare land	1.5	7.8
Urban	10.8	58.4
Industry	5.7	30.9

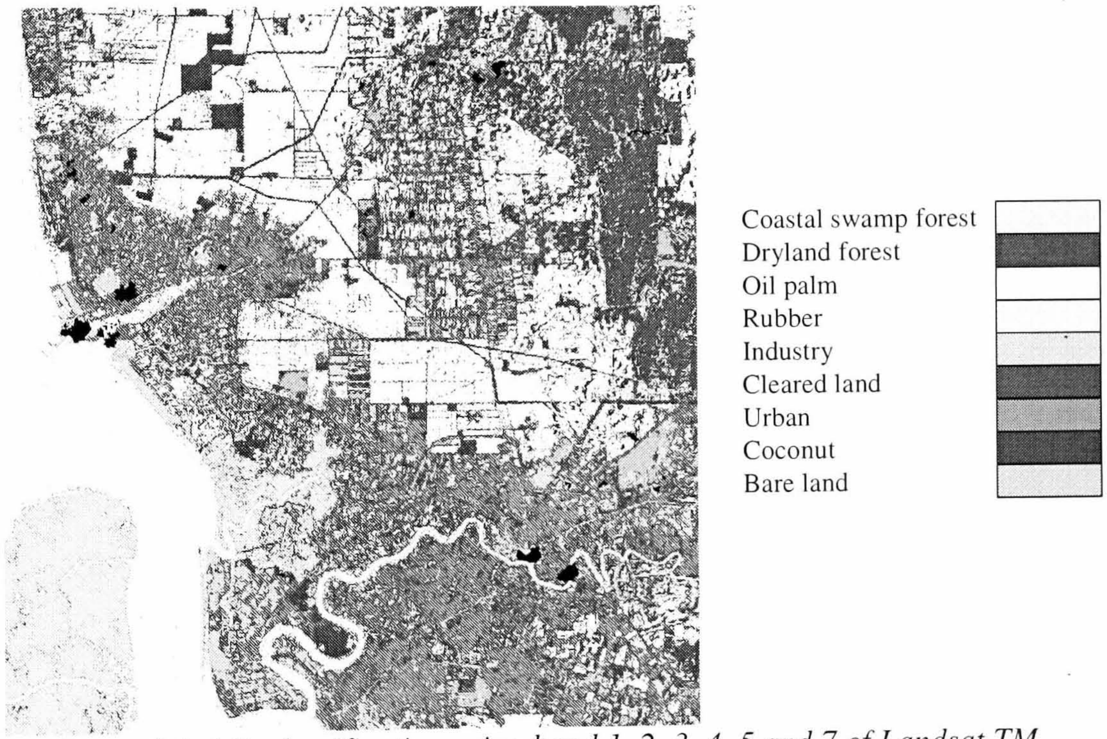


Figure 3.3: *ML classification using band 1, 2, 3, 4, 5 and 7 of Landsat TM.*

Accuracy assessment of the ML classification is determined by means of a confusion matrix (sometimes called an error matrix), which compares, on a class-by-class basis, the relationship between reference data (ground truth) and the corresponding results of a classification (Lillesand et al. 2004). Such matrices are square, with the number of rows and columns being equal to the number of classes, i.e. 9.

For each class, a different set of the pixels (i.e. those not overlapping with the training pixels) were chosen to be reference pixels. They were selected by making use of the stratified random sampling technique: rubber (230), coastal swamp forest (5175), dryland forest (2194), oil palm (3665), industry (125), cleared land (347), urban (811), coconut (564) and bare land (111) (Figure 3.2 (c)).

Table 3.3 shows the confusion matrix for the ML Classification. The diagonal elements in Table 3.3(b) represent the percentage of correctly assigned pixels and are also known as the producer accuracy. Producer accuracy is a measure of the accuracy of a particular classification scheme and shows the percentage of a particular ground class that has been correctly classified. The minimum acceptable accuracy for a class is 70% (Thomlinson et al. 1999). This is calculated by dividing each of the diagonal elements in Table 3.3 (a) by the total of the column in which it occurs:

$$\text{Producer accuracy} = \frac{c_{aa}}{c_{\bullet a}} \qquad \text{... (3.15)}$$

where,

c_{aa} = element at position ath row and ath column
$c_{\bullet a}$ = column sum

Table 3.3 (c) shows the producer accuracy for all the classes. It can be seen that all classes possess producer accuracy higher than 90%. Bare land gives the highest (100%) and cleared land the lowest (91%) figures. The low accuracy of figures for cleared land is mainly because 3% and 2% of its pixels were classified as coconut and oil palm, while 1% each as industry and rubber respectively; i.e. the small roads and spaces between trees were misclassified as cleared land due to their having quite similar spectral properties.

User accuracy is another measure of how well the classification has performed. This indicates the probability that the class to which a pixel is classified from an image actually representing that class on the ground (Story and Congalton 1986; Congalton 1991). This is calculated by dividing each of the diagonal elements in the confusion matrix by the total of the row in which it occurs:

$$\text{User accuracy} = \frac{c_{aa}}{c_{a\bullet}} \qquad \ldots (3.16)$$

where, $c_{a\bullet}$ = row sum

Coastal swamp forest, dryland forest, oil palm, bare land and urban show a user accuracy of more than 90%. Cleared land possesses the lowest accuracy, i.e. 77%, while coconut and industry account for between 80% and 90%. The low accuracy of cleared land is because the cleared land (3%) and oil palm (3%) pixels are classified as coconut.

A measure of behaviour of the ML classification can be determined by the overall accuracy, which is the total percentage of pixels correctly classified, i.e.:

$$\text{Overall accuracy} = \frac{\sum_{a=1}^{U} c_{aa}}{Q} \qquad \ldots (3.17)$$

where Q and U represent the total number of pixels and classes respectively. The minimum acceptable overall accuracy is 85% (Thomlinson et al. 1999; McCormick 1999; Scepan 1999; Wulder et al. 2006).

The Kappa coefficient κ is a second measure of classification accuracy which incorporates the off-diagonal elements as well as the diagonal terms to give a more robust assessment of accuracy than overall accuracy. This is computed as (Jensen 1996):

$$\kappa = \frac{\sum_{a=1}^{U} \frac{c_{aa}}{Q} - \sum_{a=1}^{U} \frac{c_{a\bullet} c_{\bullet a}}{Q^2}}{1 - \sum_{a=1}^{U} \frac{c_{a\bullet} c_{\bullet a}}{Q^2}} \qquad \text{... (3.18)}$$

where $c_{a\bullet}$ = row sum and $c_{\bullet a}$ = column sum . The ML classification yielded an overall accuracy of 97.8% and Kappa coefficient 0.97, indicating very high agreement with the ground truth.

To further validate the accuracy achieved, we extended this analysis by performing the assessment technique proposed by Koukoulas and Blackburn (2001), in terms of the classification success index (CSI) and individual classification success index (ICSI). CSI is defined as the sum of average user and producer accuracy minus one:

$$\mathrm{CSI} = \left[\frac{\sum_{a=1}^{U} (\mathrm{UA}_i + \mathrm{PA}_i)}{U} - 1 \right] \qquad \text{... (3.19)}$$

Where UA_i and PA_i represent user accuracy and producer accuracy for class i . The CSI for the ML classification was 0.9 (Table 3.3(d)).

ICSI is the CSI for specific class and is defined as the sum of producer and user accuracy, minus one for a particular class:

ICSI for class i can be calculated using:

$$\mathrm{ICSI} = [\mathrm{UA}_i + \mathrm{PA}_i - 1] \qquad \text{... (3.20)}$$

Five classes, i.e. coastal swamp forest, dryland forest, bare land, oil palm and rubber, show a ICSI of more than 0.9, while that of cleared land and coconut is less than 0.8 (Table 3.3(d)).

Table 3.3(a): *Confusion matrix for ML classification in pixels.*

Overall Accuracy = 97.72%

Kappa Coefficient = 0.97

		Ground Truth (Pixels)									
	Class	Coastal swamp forest	Dryland forest	Oil palm	Cleared land	Coconut	Bare land	Urban	Industry	Rubber	Total classified pixels
ML Classification (pixels)	Coastal swamp forest	5156	0	0	0	0	0	0	0	0	0
	Dryland forest	0	2180	2	0	0	0	0	0	0	0
	Oil palm	0	4	3504	6	12	0	0	0	0	0
	Cleared land	0	0	22	310	8	0	59	1	77	1
	Coconut	0	1	126	9	533	0	0	0	0	0
	Bare land	0	0	0	1	0	111	0	0	0	0
	Urban	0	0	3	7	6	0	744	0	0	0
	Industry	6	1	0	3	0	0	8	124	1063	0
	Rubber	0	6	0	3	0	0	0	0	223	232
	Total ground truth pixels	5162	2192	3657	339	559	111	811	125	230	13186

Table 3.3(b): *Confusion matrix for ML classification in percentages.*

		Ground Truth (%)									
	Class	Coastal swamp forest	Dryland forest	Oil palm	Cleared land	Coconut	Bare land	Urban	Industry	Rubber	Total classified pixels
ML Classification (%)	Coastal swamp forest	99.88	0	0	0	0	0	0	0	0	39.1
	Dryland forest	0	99.45	0.05	0	0	0	0	0	2.17	16.59
	Oil palm	0	0.18	95.82	1.77	2.15	0	0	0	0	26.74
	Cleared land	0	0	0.6	91.45	1.43	0	7.27	0.8	0.87	3.05
	Coconut	0	0.05	3.45	2.65	95.35	0	0	0	0	5.07
	Bare land	0	0	0	0.29	0	100	0	0	0	0.85
	Urban	0	0	0.08	2.06	1.07	0	91.74	0	0	5.76
	Industry	0.12	0.05	0	0.88	0	0	0.99	99.2	0	1.08
	Rubber	0	0.27	0	0.88	0	0	0	0	96.96	1.76
	Total ground truth pixels	100	100	100	100	100	100	100	100	100	100

Table 3.3(c): *Producer accuracy for the classes.*

Class	Producer Accuracy	
	(Pixels)	(%)
Coastal swamp forest	5156/5162	99.88
Dryland forest	2180/2192	99.45
Oil palm	3504/3657	95.82
Cleared land	310/339	91.45
Coconut	533/559	95.35
Bare land	111/111	100
Urban	744/811	91.74
Industry	124/125	99.2
Rubber	223/230	96.96

Class	User Accuracy	
	(Pixels)	(%)
Coastal swamp forest	5156/5156	100
Dryland forest	2180/2187	99.68
Oil palm	3504/3526	99.38
Cleared land	310/402	77.11
Coconut	533/669	79.67
Bare land	111/112	99.11
Urban	744/760	97.89
Industry	124/142	87.32
Rubber	223/232	96.12

Table 3.3(d): *CSI and ICSI for the classes.*

Class	User Accuracy	Producer Accuracy	ICSI
Coastal swamp forest	1	0.9988	0.9988
Dryland forest	0.9968	0.9945	0.9913
Oil palm	0.9938	0.9582	0.952
Cleared land	0.7711	0.9145	0.6856
Coconut	0.7967	0.9535	0.7502
Bare land	0.9911	1	0.9911
Urban	0.9789	0.9174	0.8963
Industry	0.8732	0.992	0.8652
Rubber	0.9612	0.9696	0.9308
CSI = 0.9			

3.4.2 Accuracy Analysis

ML with 9 classes has an overall accuracy 97.7% and a Kappa coefficient of 0.97) (Table 3.3(a)).

In terms of individual classes, in descending order, the producer accuracies (Table 3.3(c)) of the classes are bare land (100%), coastal swamp forest (99.88%), dryland forest (99.45%), industry (99.2%), rubber (96.96%), oil palm (95.82%), coconut (95.35%), urban (91.74%) and cleared land (91.45%).

The CSI and ICSI by Koukoulas and Blackburn (2001) (see Section 3.2.2), is also considered. The CSI for ML was found to be 0.9 (i.e. exceeding 0.8 - the index for an acceptable classification (Koukoulas and Blackburn 2001)). The ICSI for the individual classes were coastal swamp forest (1), dryland forest (0.99), oil palm (0.95), cleared land (0.69), coconut (0.75), bare land (0.99), urban (0.89), industry (0.87) and rubber (0.93). Only cleared land and coconut showed an ICSI of less than 0.8. However, these classes are less important economically compared to the rest of the classes. Overall, the analyses show that the ML classification is a satisfactory and therefore can be used as a base map for studying the effects of haze in Chapter 4.

3.4.3 Correlation Matrix Analysis

As discussed in Section 3.3, classification uses the covariance of the bands. Nonetheless, covariance is not intuitive; more intuitive is the correlation, $\rho_{k,l}$, i.e. covariance normalised by the product of the standard deviations of bands, k and l:

$$\rho_{k,l} = \frac{C_{k,l}}{\sigma_k \sigma_l} = \frac{E\left((I_k - \mu_k)\,(I_l - \mu_l)\right)}{\sigma_k \sigma_l} \qquad \text{... (3.21)}$$

where $C_{k,l}$ is the covariance between bands k and l, σ_k and σ_l are the standard deviations of the measurements in bands k and l respectively, E is the expected value operator, and I_k and I_l and μ_k and μ_l are the intensities and means of bands k and l respectively. When using more than two bands, it is convenient to use a correlation matrix, where the element in row m and column n that correspond to band k and l is given by $\rho_{k,l}$. If $m = n$, then $\rho_{k,l} = 1$, so this will be the value of the diagonal elements of the matrix. Otherwise, if $m \neq n$, $\rho_{k,l}$ lies between -1 and 1.

In order to analyse the correlation matrices, plots of correlation versus band pairs for all classes from ML are plotted (Figure 3.4). Each coloured curve represents a correlation between a specific band (given by a specific colour) and all bands (on the x-axis).

Landsat bands 1, 2 and 3 are located within a very close wavelength range of the visible spectrum, with their centre wavelengths differing only by about 0.1 µm. Measurements made from these bands normally exhibit similar responses and therefore are highly correlated. Poor correlations may result from mixed pixel problems (the existence of more than one class in a pixel). Correlations between lower-numbered bands (i.e bands 1, 2 and 3) and higher-numbered bands (i.e. bands 4, 5, and 6) are much lower because involving bands with non-adjacency wavelengths.

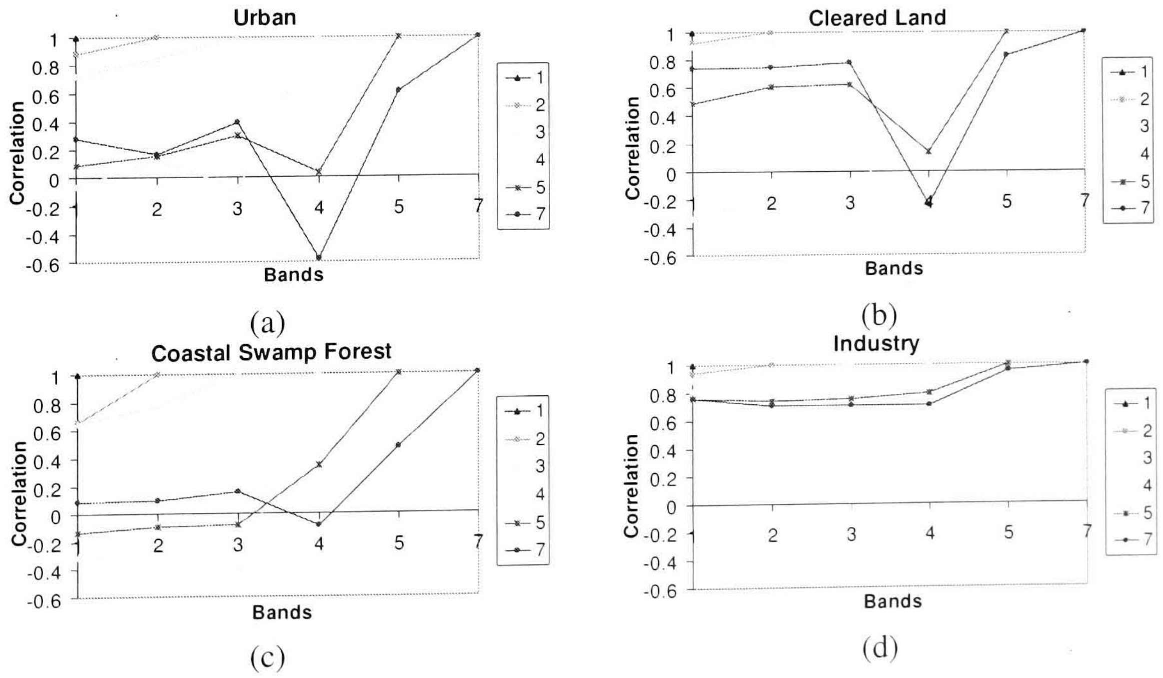
A high correlation is shown by industry (with very high reflectances) due to the strong relationships of variation between the brightness of pixels and mean brightness in all bands (1, 2, 3, 4, 5 and 7).

For dryland and coastal swamp forest, it is apparent that correlations involving bands 1, 2 and 3 are always quite high. This is because these band combinations are always correlated when measuring reflectance from green-vegetation types; band 1 is ideal to discriminate vegetation from soil, band 2 detects green reflectance from healthy vegetation and band 3 detects chlorophyll absorption. However, for coastal swamp forest, negative correlations can be seen for pairs involving bands 4 and 5, which are very sensitive to forest stand timber volume (Gemmel 1995). This is consistent with the fact that the distinct vegetation species in both forests have different spectral properties observed from bands 4 and 5. The different spectral properties are associated mainy with the tree species composition, forest stand structures and vegetation vigour (Lu et al. 2004).

Dryland forest has a stronger pair a 7:4 correlation than the coastal swamp forest because of the stronger relationship between the timber volume and ratio of bands 7 / 4 (Ahern et al. 1991). Most timber in Malaysia comes from the dry land forest type as the tree structure is much bigger than that of the coastal swamp forest. It mainly comes from the dipterocarp forest species, which include Anisoptera, Dipterocarpus, Dryobalanops, Hopea, Shorea and Parashorea (Suzuki 2005). Since bands 5 and 7 are located in the near and mid infrared region respectively, they are sensitive to water in leaves. Hence, they are well correlated with each other.

When compared with individual class accuracies (Table 3.3(c)), bare land (100%), rubber (96.96%), coconut (95.35%), industry (99.2%) and dryland forest (99.45%) have positive correlations for all pairs. Overall, industry has higher positive correlations for all pairs in comparison to bare land, but the former has a lower classification accuracy compared to the latter, while coastal swamp forest with the second highest accuracy has a mixed correlation trend. Thus there is no clear relationship between the positiveness of the correlation and classification accuracy.

In conclusion, land covers have unique band correlation trends that explain the relationship s between measurements from different bands. However it was found that there is no clear relationship between the correlation trends and the classification accuracy of a land cover.

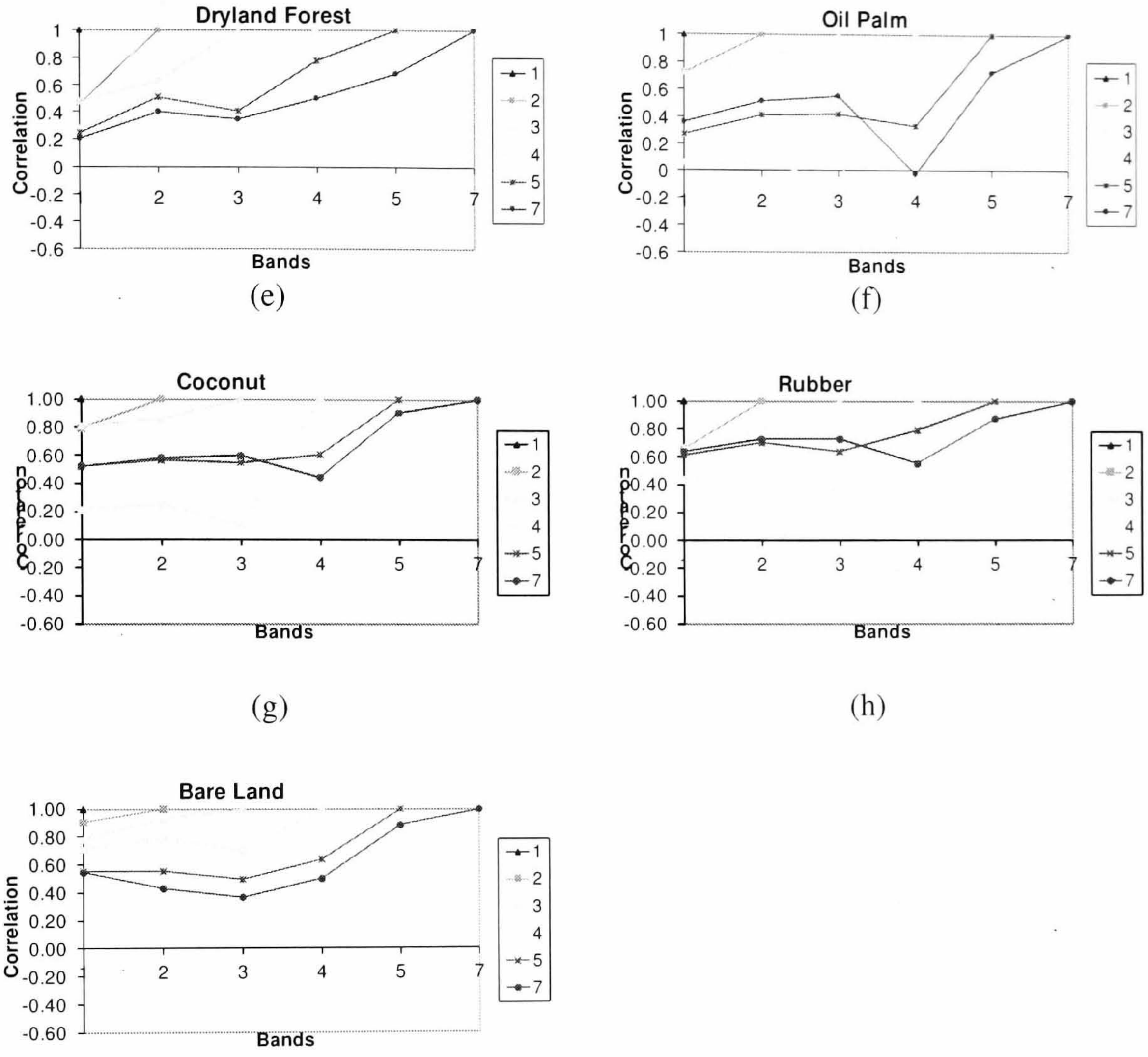


(a)

(b)

(c)

(d)



(e) (f)

(g) (h)

(i)

Figure 3.4: *Correlations between band pairs from the ML classification for the classes.*

3.4.4 Mean and Standard Deviation Analysis

It is interesting that, despite being very similar, both forests can still be separated quite effectively by the ML. Figure 3.5 shows the means of coastal swamp forest and dryland forest classes, which are almost the same, particularly in bands 1, 2, 3 and 4. The quite low DNs in band 3 are due to the absorption of the red light by vegetation within the forests (i.e. also known as chlorophyll absorption band). Since vegetation has high reflectance in band 4 (near IR region), they have quite big DNs (bright). The quite different DN in bands 5 and 7 between the forests is due to the quite different moisture conditions of the vegetations occupying the forests. The largest difference in mean occurred in band 5 due to the sensitivity to the variation of moisture conditions between the forests. This is followed by bands 1 and 7.

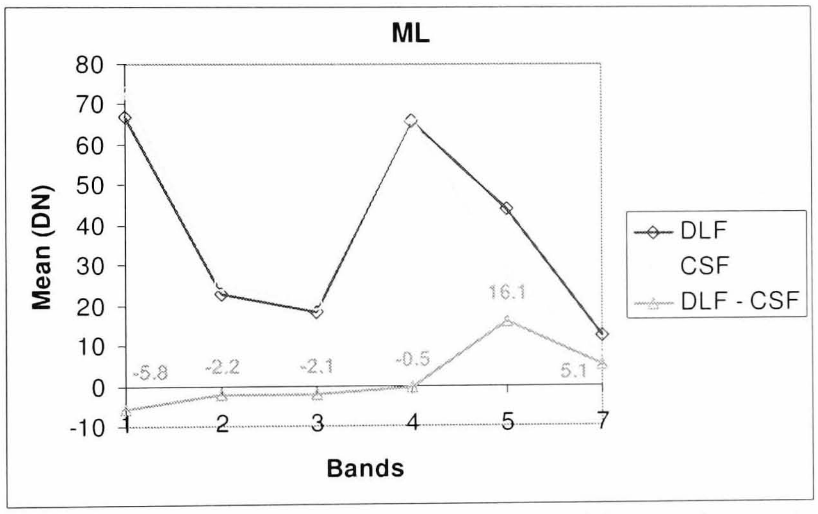


Figure 3.5: *Means of coastal swamp forest and dryland forest classes in ML. DLF and CSF are dryland forest and coastal swamp forest respectively.*

In term of standard deviation, the largest is in band 4 (chlorophyll absorption band) due to the variation in spatial patterns of vegetation species within the forest. The largest difference in standard deviation occurs in band 5, which is due to the variation of moisture conditions of the vegetations within the forests, which is consistent with the mean analysis. The smallest difference occurs in band 4, indicating that the variation of chlorophyll absorption by vegetation within the forests is not significant.

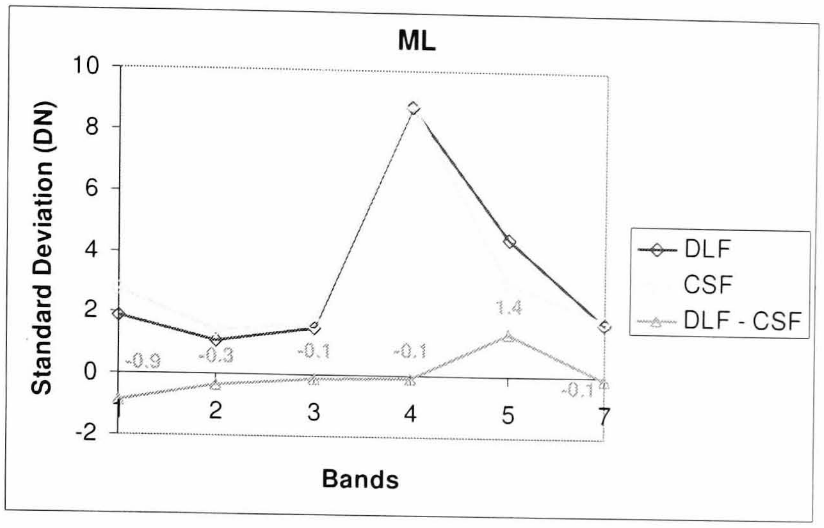


Figure 3.6: *Standard deviations of the coastal swamp forest and dryland forest classes in ML.*

Overall, band 5, with the biggest difference in terms of forest mean and standard deviation, seems to be the most effective band for discriminating between the coastal swamp forest and dryland forest.

3.4.5 Decision Boundary Analysis

In this section, we will investigate ML further in terms of the decision boundaries generated from Equation 3.18 between coastal swamp forest and dryland forest.

The 15 sets of decision boundaries, generated for all band pairs are shown in Figure 3.7; 'M1' and 'M2' are the means for dryland forest and coastal swamp forest respectively, 'Band k Vs. Band l' denotes that the vertical axis is band k while horizontal axis is band l and 'DLF' indicate dryland forest respectively, i.e. to which class the boundary belongs to. For convenience, the points associated with the forests are also plotted.

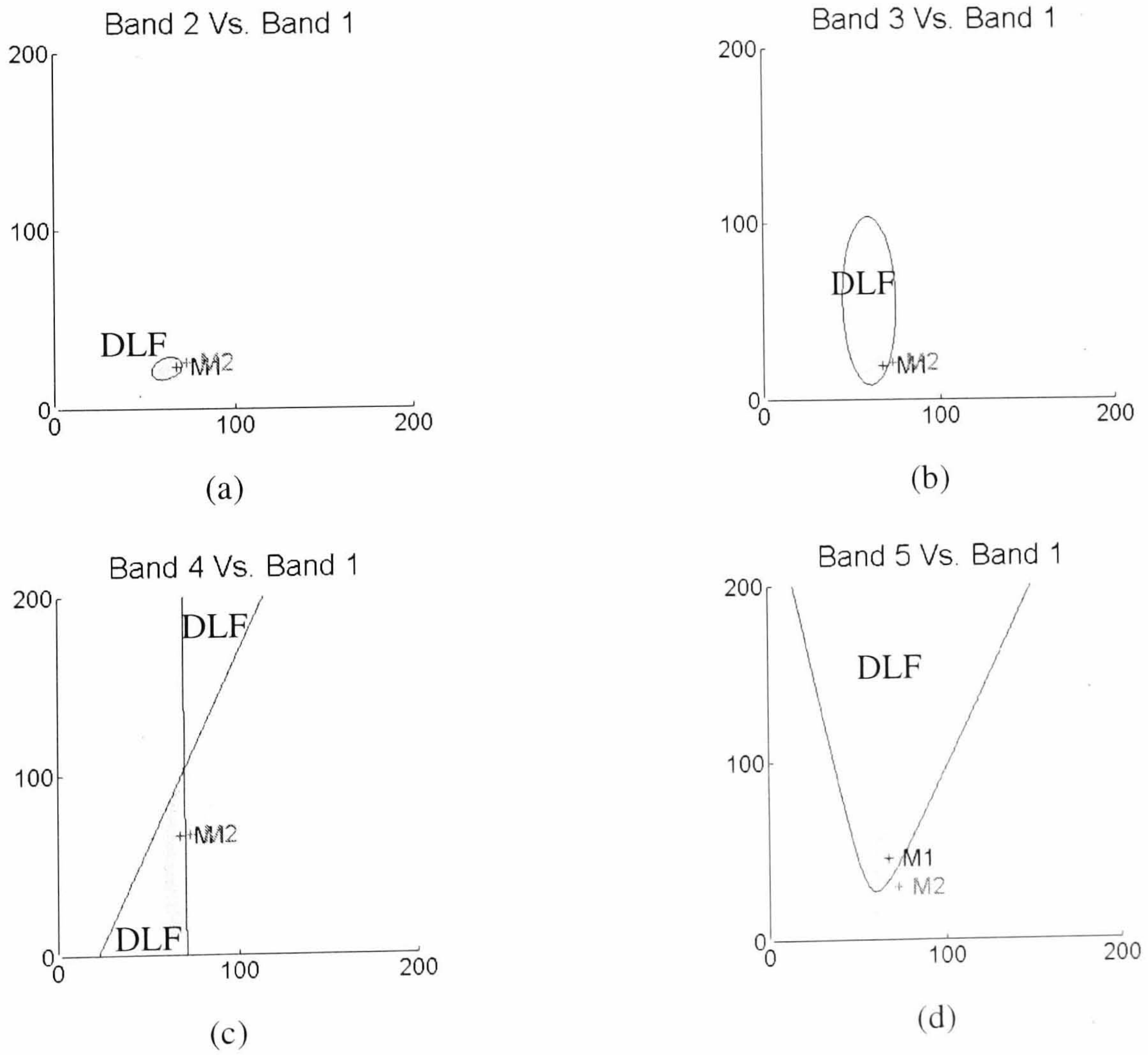
The decision boundaries has the form of conic sections; pairs 2:1, 3:1, 7:1, 3:2 and 7:2 form an elliptic curve, while pairs 5:1, 5:2, 5:3, 7:3 and 7:5 form a parabolic curve and pairs 4:1, 4:2, 4:3, 5:4 and 7:4 form a hyperbolic curve.

It can be seen that some dry land forest points are outside of its boundary; this inconsistency is due to the misclassification occurred (see Table 3.3), i.e. 0.3% and 0.2% of the dry land forest pixels were classified as rubber and oil palm respectively, while 0.05% pixels were classified as coconut and industry. The boundaries of these classes are not shown here because the focus of this discussion is on the forests that are spectrally very similar, besides them involving lengthy computational times.

For pairs 2:1, 3:1, 7:1, 3:2 and 7:2, the decision region for dryland forest is located inside the decision boundary because dryland forest has a smaller variance than coastal swamp forest in these bands (see Figure 3.6). This is consistent with the coastal swamp forest points that are more widely scattered than dryland forest. For pairs 2:1, 3:1 and 3:2, as expected, the orientation of the points indicates that these pairs have quite similar spectral properties, therefore are highly correlated. This causes much pixel redundancy and produces limited information for separating the forests. The decision boundaries for pairs 2:1 and 3:2 seem to be very small due to the quite small variance in both bands.

For pairs 5:1, 5:2, 5:3, 7:3 and 7:5, the points for both forests seem to concentrate at the narrower part of the decision boundary. For pairs 5:1, 5:2 and 5:3 the coastal swamp forest points seem scattered in a circle-like shape, indicating that the pairs have quite low correlations due to the quite distict spectral properties of the bands. A longer vertical shape is shown by the dry land forest points due to the much bigger variance in band 5 compared to band 1, 2 and 3. For pair 7:3, both forests seem to have low correlations, but a higher correlation can be seen in pair 7:5, due to the quite similar spectral properties of the bands. It is clear that, compared to other bands, pairs involving band 5 locate quite a large number of points within the boundary while reasonably large portions of dry land forest points are located within the boundary with less overlapping occurs with the coastal swamp fores, indicating that band 5 is very useful in discriminating between the forests. The main advantage of band 5 is its ability to separate the forest means quite effectively (Figure 3.5).

For pair pairs 4:1, 4:2, 4:3, the points are aligned vertically along the lower part of the boundaries due to the much bigger variances in band 4 compared to bands 1, 2 and 3 (see Figure 3.6). This indicates the usefulness of band 4 in discriminating between the species within the forests. The quite significant overlapping which occurs in x-direction is because band 4 has the smallest difference between the forest means. For 5:4 and 7:4, the points seem to be concentrated horizontally along the lower boundary, due to the much bigger variance in band 4 compared to bands 5 and 7. However, less overlapping can be seen in pair 5:4 compared to 7:4, indicating the usefulness of pair 5:4 in discriminating the forests. In conclusion, the ability of ML to position the forests means (although the difference is very small) that the different side of most of the decision boundaries appears to be one of the key factors that enable ML to discriminate effectively between the forests.

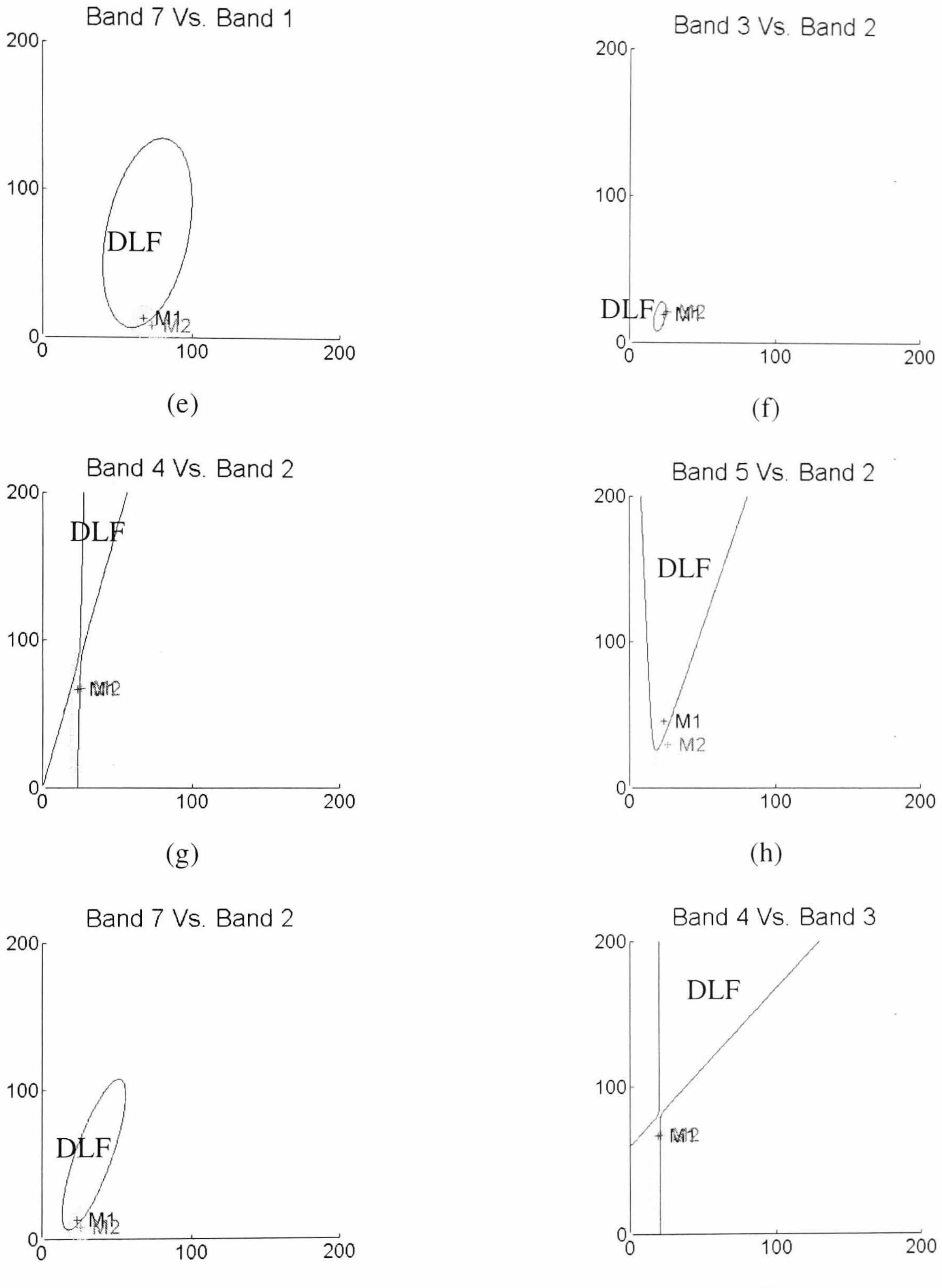


(a)

(b)

(c)

(d)


Band 7 Vs. Band 1
200
100
0
0
100
200
DLF
M1
M2
Band 3 Vs. Band 2
DLF
M1
M2
Band 4 Vs. Band 2
DLF
M1
M2
Band 5 Vs. Band 2
DLF
M1
M2
Band 7 Vs. Band 2
DLF
M1
M2
Band 4 Vs. Band 3
DLF
M1
M2

(e)

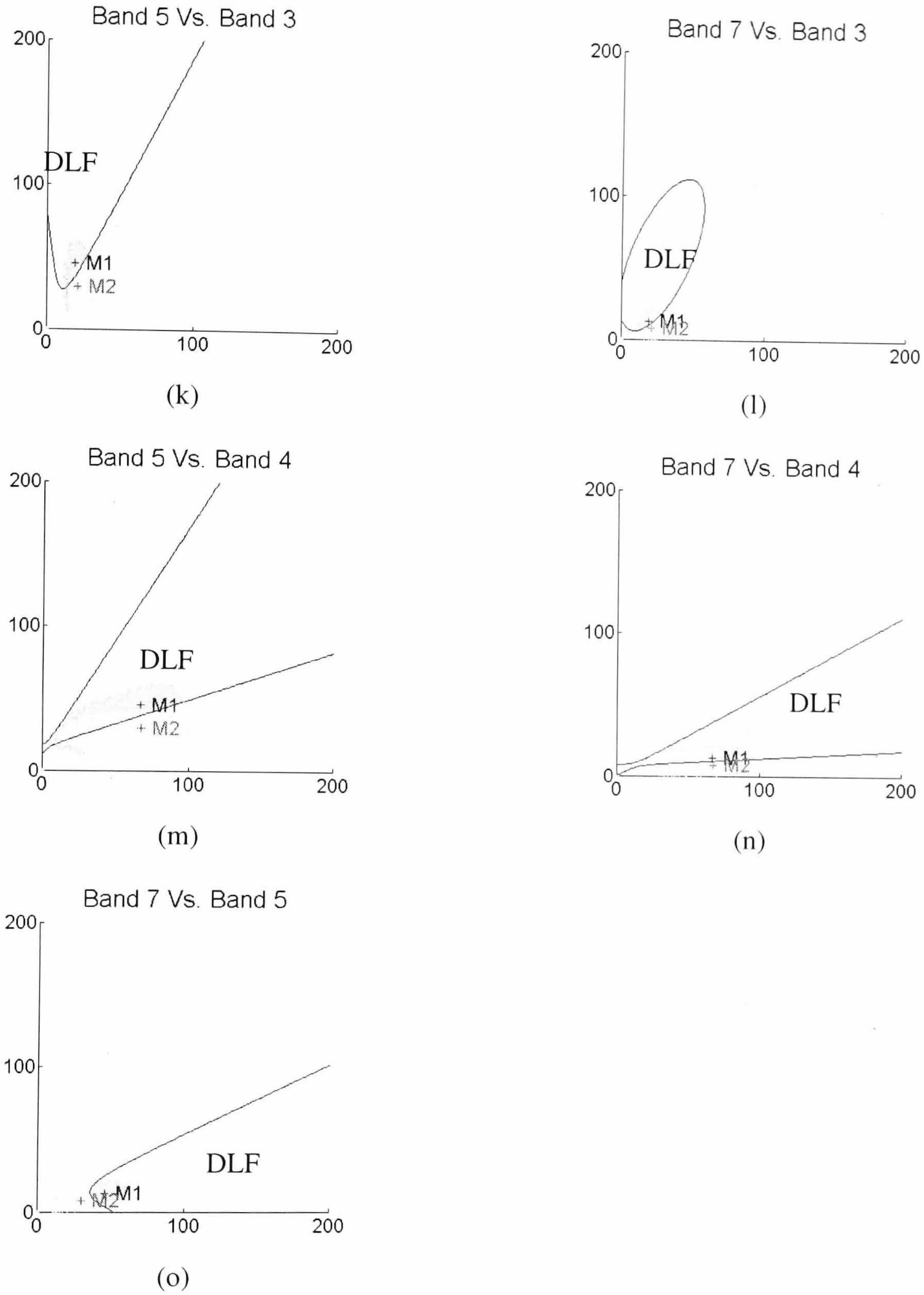
(f)

(g)

(h)

(i)

(j)



(k)

(l)

(m)

(n)

(o)

Figure 3.7: *Decision boundaries between coastal swamp forest with points coloured in yellow and dryland forest with points coloured in cyan for ML classification. 'M1' and 'M2' are the means for dryland forest and coastal swamp forest respectively. 'Band k Vs. Band l' denotes that the vertical axis is band k while horizontal axis is band l. The decision space for dryland forest is indicated by'DLF'.*

Here, we have shown just a two dimensional decision boundary for two classes. It is anticipated that higher dimensional plots (i.e. taking account three or more bands) will cause difficulty in analysing and interpretation. To explain this, a six-dimensional scatter plot of the forests, without a decision boundary, is shown in Figure 3.8.

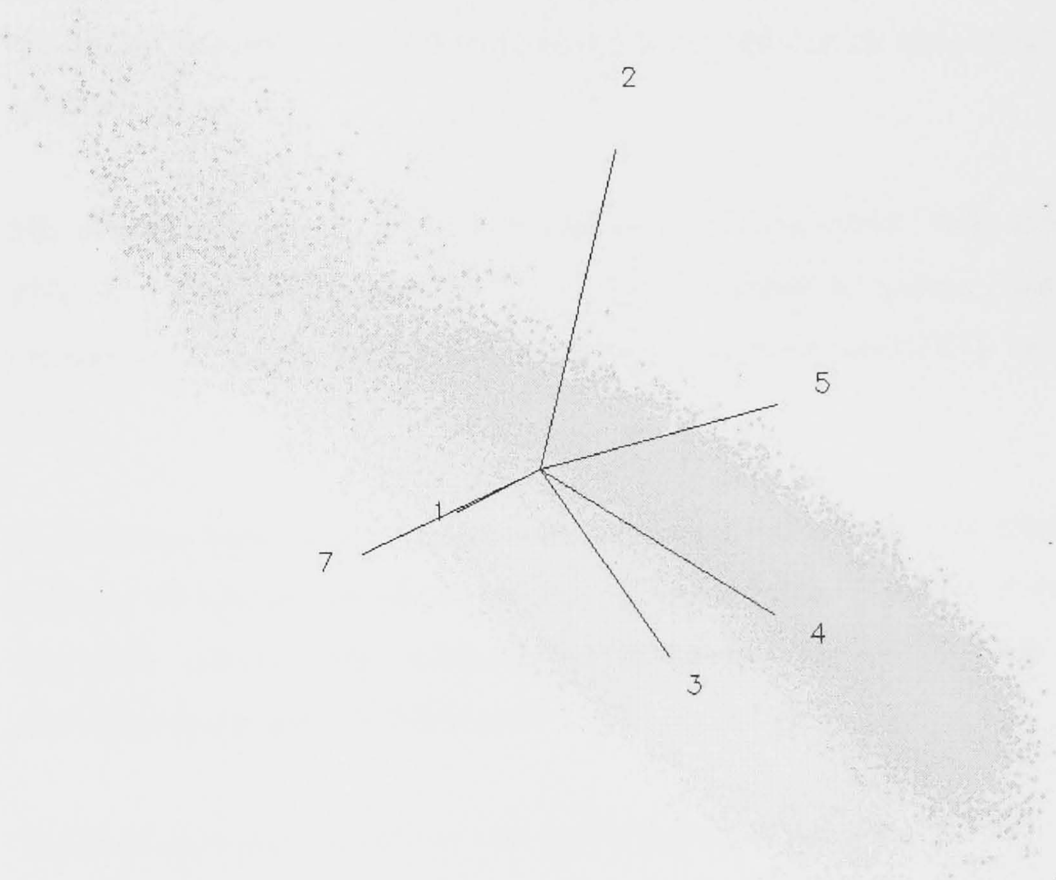


Figure 3.8: *A six-dimensional scatter plot of coastal swamp forest (cyan) and dryland forest (yellow); the numbers correspond to Landsat band.*

3.5 Summary and Conclusions

In this chapter, an analysis of ML classification for Selangor, Malaysia, has been carried out using a Landsat dataset:

1. ML classification is suitable for Malaysian land covers due to its simplicity, objectivity and ability in classifying land covers with a good agreement with the land cover map.

2. ML classified the study area into 9 classes, as chosen earlier, with accuracy of 97%, κ = 0.97 and CSI = 0.9, i.e. overall and producer accuracy were fairly consistence with those indicated by Kaokoulas and Blackburn (2001), i.e. CSI and ICSI.

3. Land covers had a unique band correlation trends that explains the relationships between the spectral measurements from different bands. However, it was found that there was no clear relationship between the correlation trends and the classification accuracy of a land cover.

4. The band correlation of classes with high reflectance, e.g. industry, was quite high for all band pairs because of the strong relationships of variation between the brightness of pixels and the mean brightness in all bands.

5. Band 5 was found to be the most effective band in discriminating between the coastal swamp forest and the dry land forest, due to its having shown the biggest difference in terms of forest mean and standard deviation.

6. The ability of ML to position the forests means that the different side of most of the decision boundaries appeared to be one of the key factors that enabled ML to discriminate effectively between the forests.

Chapter 4

The Effects of Haze on Land Classification

4.1 Introduction

This chapter deals with one of the crucial aims of this thesis, which is to investigate the effects of haze on land classification. In achiving this aim, we need to make use of datasets with known haze conditions; nevertheless, acquiring real hazy datasets with a desired range of haze concentrations over an area is not possible. A more practical way is to model the haze and simulate its effects on clear datasets. To do so, we need to know the effects that haze concentrations have on scene visibility and to translate it onto real remote sensing data. *The primary objectives of this chapter is therefore to simulate hazy datasets, investigate their spectral and statistical properties and examine classification performance on these hazy datasets in terms of classification accuracy.*

To achieve this aim, we need to know the existing approaches and issues encountered; Section 4.2 discusses previous studies of the effects of haze on land classification. Since this chapter deal with haze, Section 4.3 relates haze with visibility while Section 4.4 discusses haze scenario that occurs in Malaysia. An important issue in investigating the effects of haze is to model hazy dataset; in Section 4.5, a model for integrating haze with a clear atmosphere dataset is described. Next, we need to translate the model to practical processes; Section 4.6 discusses simulation of hazy datasets by incorporating simulated haze path radiance and the effects of signal attenuation onto clear datasets. Since the primary issue is to investigate the effects that haze has on classification; Section 4.7 discusses ML classification of the hazy datasets when the training pixels are from hazy datasets, and measures the performance of the classification. We are left wanting to know the effects of haze on classification when the training pixels are from clear dataset; Section 4.8 discusses this issue and compares the results with classification that uses training pixels from the hazy dataset itself. Finally, in Section 4.9, we carry out an extended analysis on classification accuracy when haze scattering component is ommited from hazy dataset .

4.2 Previous Studies

Atmospheric aerosols and molecules scatter and absorb solar radiation, thus affecting the downward and upward radiance. Atmospheric scattering and absorption depend substantially on the wavelength of the radiation (Kaufman 1987; Kaufman and Sendra1988; Kaufman et al. 1997). Scattering is usually much stronger for short wavelengths than for long wavelengths and significantly affects the classification of surface features (Kaufman and Fraser 1984). Studies related to haze effects on remote sensing measurements commonly use real hazy datasets. An alternative way is to simulate hazy datasets based on a suitable model; subsequently, classification can be carried out on these datasets and their performance can be examined systematically.

Fraser et al. (1977) carried out an experiment to analyse the effect of a change in the atmospheric turbidity on classification of Landsat MSS data. ML classification on an original dataset was carried out to segment the land cover types. The pixel brightness (i.e. in radiance) of the dataset was then reduced by using a simulation approach. The modified dataset was subsequently classified by using ML classification, making use of training pixels from the original dataset. 22% of the modified pixels changed compared to the classification of the original dataset. The modified dataset was then classified by using training pixels from the modified dataset itself, yielding only 3% change from the classification of the unmodified dataset. They concluded that using training pixels from the modified data itself weakens the effect of atmospheric turbidity on the accuracy of a ML classification of surface features. The main shortcoming of the method is that the use of only two datasets does not help much in systematically examining the effects of haze, which can occur at various levels.

Kaufman and Sendra (1988) established a mathematical model to be used in the development of algorithms for automatic atmospheric corrections in visible and near-IR satellite imagery. The model defines the radiance reflected from the Earth as a combination of three components: the radiance scattered by the atmospheric constituents directly to the sensor, without being reflected by the Earth surface (path radiance), the radiance reflected by the surface to the sensor (attenuated signal) and the radiance reflected from the surface and then scattered by the atmosphere to the sensor (diffuse radiance). The model introduced by Chandrasekhar (1960) was then employed to establish the relationship between the observed radiance and the surface reflectance, assuming a Lambertian surface and cloudless atmosphere. The model seems simple but is very useful to account for atmospheric effects and also to remove them.

Molenar et al. (1994) combined an atmospheric aerosol model and a radiative transfer (RT) model to simulate different visibility conditions of an image. The optical properties (e.g. single scattering albedo and extinction, scattering and absorption coefficients, etc.) of the atmospheric aerosol model were calculated by using a model developed by Colorado State University. These were then incorporated in the RT model. Molenar et al. (1994) define the observed image radiance as:

$$N_r = N_o T_r + N^* \qquad \text{... (4.1)}$$

where N_O is the radiance of a clear image, T_r is the atmospheric transmittance, which is calculated from the atmospheric aerosol model, and N^* is the path radiance, which is calculated based on an equilibrium radiance model as:

$$N^* = N_s(1 - T_r) \qquad \text{... (4.2)}$$

Here N_s is the sky radiance (radiance arising from radiation that is scattered downward by the atmosphere and then reflected into the instantaneous field of view (IFOV) of the pixel of interest), calculated by using a Monte Carlo model, where Lambertian reflection was assumed. Using this approach, Molenar et al. (1994) successfully simulated different visibility conditions, but its main disadvantage is that use of a Monte Carlo model was time consuming compared to later types of RT model (Kotchenova et al. 2007).

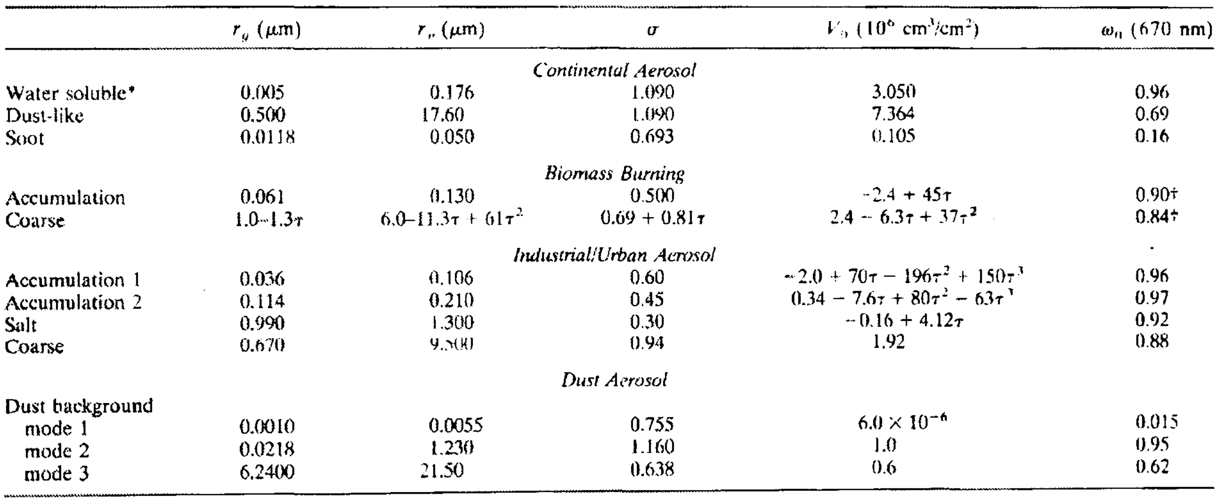
Oakley and Satherley (1998) developed a model for contrast degradation to improve the quality of images taken using forward-looking airborne sensors. Initially, a forward model that uses information about the scene and imaging conditions was developed to predict image characteristics. The assumptions made for this model were:

- The scattering effects caused by aerosols lead to exponential attenuation of the reflected flux.
- The flux observed by the sensor is the sum of the flux due to reflection from the Earth's surface and flux due to light scattered by aerosols.
- The surface is Lambertian.

Subsequently, a statistical model for image degradation was introduced to characterise the contribution of the reflected flux from the surface and the contribution of the scattered flux from the aerosol. It also incorporates random effects that represent noise introduced into the image by the sensor. Variations of the scattered flux from the aerosol were modelled using Gaussian random variables. A satisfactory agreement between the model and the experimental data was achieved. The procedure and assumptions made by Oakley and Satherley (1998) seem very relevant in simulating hazy datasets.

Kaufman and Tanre (1996) determine aerosol optical thickness, one of the key parameters to calculate path radiance, based on the following steps: (1) estimating the surface reflectance of the dark pixels in the red and blue channels using the measurements in the mid-IR, (2) determination of the aerosol type using information on the global aerosol distribution and the ratio between the red and blue channels, (3) selection of the appropriate aerosol model (Remer et al. 1996). Finally, the aerosol optical thickness is calculated based on a lookup table that relates Lambertian surface reflectance to the measured reflectance as a function of the optical thickness and the solar illumination and satellite viewing geometry. The aerosol model in (3) consists of six aerosol categories, viz. continental, maritime, industry/urban, background desert, biomass burning and stratospheric, and has been adopted in a number of radiative transfer codes, e.g. 6S and 6SV1.Table 4.1 shows the key aerosol parameters, i.e. geometric mean radius, volume mean radius, standard deviation of the natural logarithm of the radius, column volume of particles per cross section of atmospheric column and single scattering albedo, for continental, biomass burning, industrial/urban and dust aerosol. In 6S and 6SV1, the vertical profile of aerosol is based on the method of successive orders of scattering (SOS) approximations (Kotchenova et al. 2006). In this method, the atmosphere is divided into a number of layers and the radiative transfer equation is solved numerically for each layer using iteration technique. The total intensity of scattered photons is obtained as the sum of all orders of scattering.

Table 4.1: *Geometric mean radius (r_g), volume mean radius, (r_v), standard deviation of the natural logarithm of the radius (σ), column volume of particles per cross section of atmospheric column (V_o) and single scattering albedo (ω_0) for for continental, biomass burning, industrial/urban and dust aerosol (Remer et al. 1996).*



	r_g (μm)	r_v (μm)	σ	V_0 (10^6 cm^3/cm^2)	ω_0 (670 nm)
			Continental Aerosol		
Water soluble*	0.005	0.176	1.090	3.050	0.96
Dust-like	0.500	17.60	1.090	7.364	0.69
Soot	0.0118	0.050	0.693	0.105	0.16
			Biomass Burning		
Accumulation	0.061	0.130	0.500	$-2.4 + 45\tau$	0.90†
Coarse	$1.0-1.3\tau$	$6.0-11.3\tau + 61\tau^2$	$0.69 + 0.81\tau$	$2.4 - 6.3\tau + 37\tau^2$	0.84†
			Industrial/Urban Aerosol		
Accumulation 1	0.036	0.106	0.60	$-2.0 + 70\tau - 196\tau^2 + 150\tau^3$	0.96
Accumulation 2	0.114	0.210	0.45	$0.34 - 7.6\tau + 80\tau^2 - 63\tau^3$	0.97
Salt	0.990	1.300	0.30	$-0.16 + 4.12\tau$	0.92
Coarse	0.670	9.500	0.94	1.92	0.88
			Dust Aerosol		
Dust background					
mode 1	0.0010	0.0055	0.755	6.0×10^{-6}	0.015
mode 2	0.0218	1.230	1.160	1.0	0.95
mode 3	6.2400	21.50	0.638	0.6	0.62

A comprehensive database of atmospheric models has been designed by Anderson et al. (1986), which tabulated parameters, such as temperature, pressure and density profiles, and mixing ratios of water vapour and gaseous (i.e. ozone, carbon monoxide and methane), for vertical atmospheric profile. The atmospheric models were categorised into six different categories, i.e. Tropical (15° N – annual average), Middle latitude' (45° N) summer (July) and winter (Jan), Sub Arctic (60° N) summer (July) and winter (Jan) and the U.S. Standard Model Atmosphere (1976). These models have been used widely in atmospheric studies, including satellite atmospheric correction, and are available in many radiative transfer models, such as LOWTRAN (Kneizys et al. 1988), MODTRAN (Berk et al. 1998), 6S (Vermote et al. 1997) and 6SV1 (Kotchenova et al. 2006).

Mahmud (2009) studied mesoscale characteristics in an equatorial environment that encompassed the island of Sumatra and Peninsular Malaysia during the haze period in August 2005 using TAPM (The Air Pollution Model) model. The model used meteorological parameters such as wind speed, temperature, and humidity to simulate haze trajectories (i.e. indicating haze transportation) during the most severe hazy days (i.e. 7 to 13 August 2005), where PM_{10} concentration was considered to be the main parameter to indicate haze. Results showed that throughout the haze period, the high loadings of aerosols from the biomass burning in Sumatera into Malaysia was mainly due to the sea-land breeze conditions. The simulated haze trajectories, integrated with MODIS fire counts from 8 to 12 August 2005 (red dots), and contours of PM_{10} concentrations ($\mu g\ m^{-3}$) and backward trajectories (yellow) for Peninsular Malaysia on 10 August 2005 are shown in Figure 4.1.

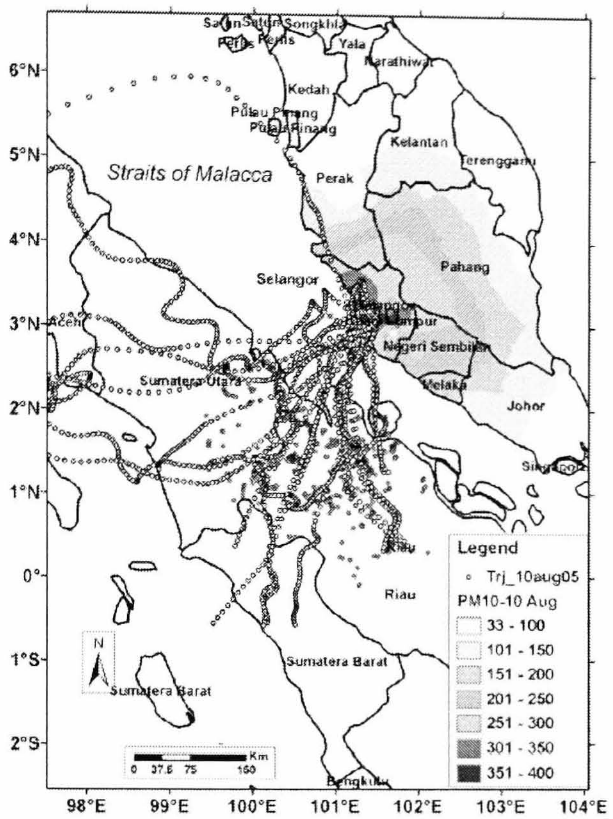
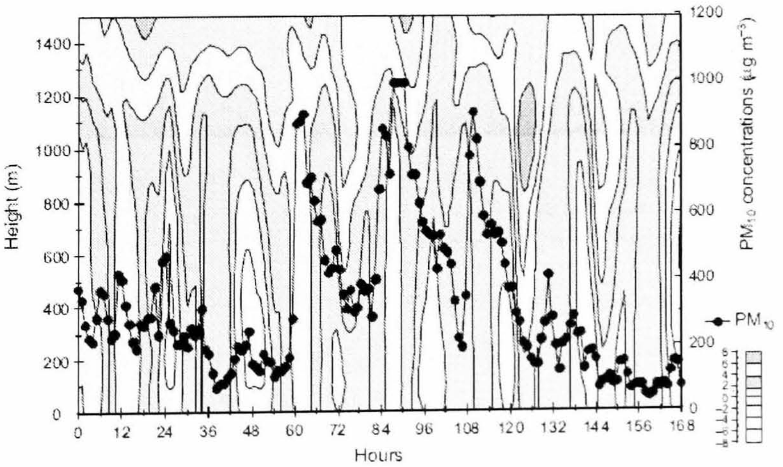
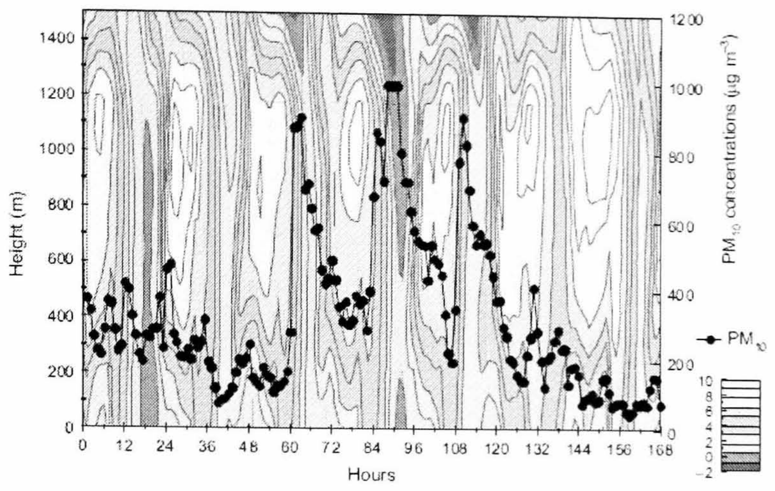


Figure 4.1: *The MODIS fire counts from 8 to 12 August 2005 (red dots), and contours of PM_{10} concentrations ($\mu g\ m^{-3}$) and backward trajectories (yellow) forPeninsular Malaysia on 10 August 2005.*

Mahmud (2009) also showed that PM_{10} concentrations were higher from midday to sunset compared to the rest of the day, due to the zonal (east-west direction) amd meridional (north-south direction) wind components (Figure 4.2).



(a)



(b)

Figure 4.2: *The vertical profiles of the (a) zonal and (b) meridional wind components, and the hourly PM10 concentrations.*

Heil et al. (2007) investigated the influence of meteorological conditions and fuel type burnt on large-scale smoke haze pollution. REMO (REgional MOdel), a three-dimensional regional scale atmospheric chemistry module, was used to investigate the atmospheric transport and removal of aerosols emitted from the Indonesian fires in the second half of 1997. The REMO model incorporated 20 vertical layers of increasing thickness between the Earth's surface, the 10-hPa pressure level using terrain - following hybrid pressure-sigma coordinates, and it was assumed that aerosols were released into the lowest atmosphere. Four land cover classes were considered, i.e. agriculture, grassland and savannah, forest plantations, fragmented forests and peat soil. Among the parameters considered were spatial distribution of monthly mean PM$_{10}$ concentrations, monthly mean total precipitation and wind vectors, monthly mean wet deposition and monthly mean PM$_{10}$ vertical distribution. The study focussed on three main locations, viz. Palangkaraya – located in Kalimantan, Indonesia, Kuching – located in Sarawak, Malaysia and Petaling Jaya – located in Selangor Malaysia. Results from modelling were found to be consistent with those from measurements and other experiments; fires from peat were identified to be a major source of PM$_{10}$. An example of the vertical PM$_{10}$ distribution for Petaling Jaya is shown in Figure 4.3; PM$_{10}$ concentration decreases with height and reaches very low concentrations at around 4500m, with the highest PM$_{10}$ loading occuring approximately at 1 km and below.

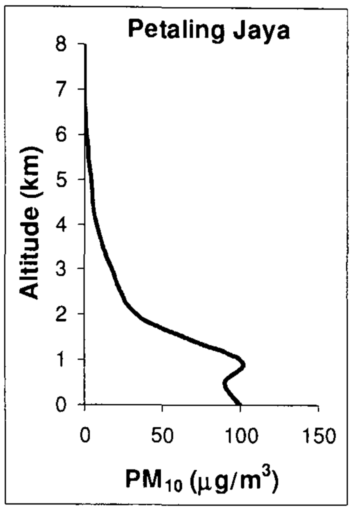


Figure 4.3: *Altitude versus PM10 concentrations for Petaling Jaya.*

In our study, a modified version of the model presented by Kaufman and Sendra (1988) will be used to model hazy datasets (Section 4.3). A similar definition of the observed image radiance (Equation 4.1 and 4.2) will be applied but a more up-to-date radiative transfer model will be used, i.e. 6SV1 (Kotchenova et al. 2006; Vermote et al. 1997) (Section 4.4). Some procedures and assumption made by Oakley and Satherley (1998) in simulating hazy datasets will be adopted (Section 4.4). The use of ML classification accuracy as a performance measure (Fraser et al. 1977) will be applied to hazy datasets, and visibility conditions ranging from 20 km (clear conditions) through to 0 km (pure haze) will be considered (Section 4.7 to 4.8).

4.3 Haze and Visibility

Haze reduces visibility due to the attenuation (i.e. scattering and absorption) of solar radiation by the haze constituents. Most haze consists of aerosols (suspension of fine solid particles or liquid droplets in the atmosphere) and trace gases, ranging in size from a few nanometres to a few micrometers. Studies have shown that haze that is due to biomass burning contains large amounts of hazardous gases, i.e., carbon monoxide (CO), nitrogen dioxide (NO_2) and sulphur dioxide (SO_2), and particulate matter, i.e., PM_{10} (Heil and Goldammer 2001; Radojevic 2003; Mahmud 2009). Compared to gases, aerosol has a more significant impact on visibility. The largest aerosol loading from biomass burning occurs below 5 km in altitude (Chiang et al. 2007).

Atmospheric scattering and absorption depend very much on the wavelength of the radiation and the size of the atmospheric constituents it interacts with. Scattering is usually much stronger for short wavelengths than long wavelengths. Particles with size approximately 0.1 to 10 μm are particularly effective in Mie scattering in the visible wavelength regions (0.4 – 0.7 μm) hence can impair ground level visibilities.

In order to define visibility, the fraction of light intensity reaching the observer can be expressed as:

$$\frac{I_{obs}}{I_o} = \exp\left[-\int_0^{x_{obs}} b_{ext}(x)\,dx\right] \qquad \text{... (4.3)}$$

where x is distance from observer, I_o is the original intensity of an object, I_{obs} is the intensity reaching the observer, and b_{ext} is the extinction coefficient evaluated at 0.55 μm wavelength, assumed to be horizontally uniform. The visibility is defined as the value of x_{obs}. The fraction on the left at which the contrast between an object and its background can no longer be distinguished is often approximated by 0.02 (Horvath 1971; Dzubay et al. 1982; Vallero 2008). Hence:

$$0.02 = \exp\left[-\int_0^{x_{obs}} b_{ext}(x)\,dx\right] \qquad \text{... (4.4)}$$

$$0.02 = \exp\left(-b_{ext} x \,|_0^{x_{obs}}\right) \qquad \text{... (4.5)}$$

$$\ln(0.02) = -b_{ext} x_{obs} \qquad \text{... (4.6)}$$

Hence,

$$\text{visibility} = -\frac{\ln(0.02)}{b_{ext}} \qquad \text{... (4.7)}$$

This is known as the Koschmieder equation and is used to define visibility. Visibility is inversely proportional to the extinction coefficient, which can be related to atmospheric aerosol concentrations (Godish 1991).

b_{ext} represents the extinction coefficient for 1 km thickness, which is relevant to the whole vertical haze loading (Heil et al. 2007), and is therefore equivalent to the corresponding aerosol optical thickness. b_{ext} depends on the presence of gases and molecules that scatter and absorb light in the atmosphere and is given by (Dzubay et al. 1982; Vallero 2008):

$$b_{ext} = b_{ray} + b_{mie} + b_{ns} + b_{abs} \qquad \text{... (4.8)}$$

where b_{ray} is Rayleigh scattering by gaseous molecules, b_{mie} is Mie scattering by particles, b_{ns} is non-selective scattering caused by bigger particles and b_{abs} is absorption by gaseous molecules. These various extinction components are functions of wavelength.

In most cases, particle scattering controls visibility reduction (Vallero 2008). Figure 4.4 illustrates the scattering and absorption efficiency per unit volume as a function of particle diameter, and shows the predominance of scattering over absorption at 550 nm wavelength. It also signifies that most scattering is caused by particles with diameters 0.1 – 1.0 μm, which are within the PM_{10} category (Vallero 2008). Visibility reduction associated with forest fires is due to scattering by particles (i.e. PM_{10}) and to a lesser extent, absorption of light by trace gases with diameters of approximately 5 x 10^{-4} μm (i.e. NO_2, SO_2, and CO) (Heil and Goldammer 2001; Mahmud 2009). The definition for clear sky visibility varies, but is normally considered to be 20 to 23 km (Longshore et al. 1976; Bird and Hulstrom 1981; Richter 2008).

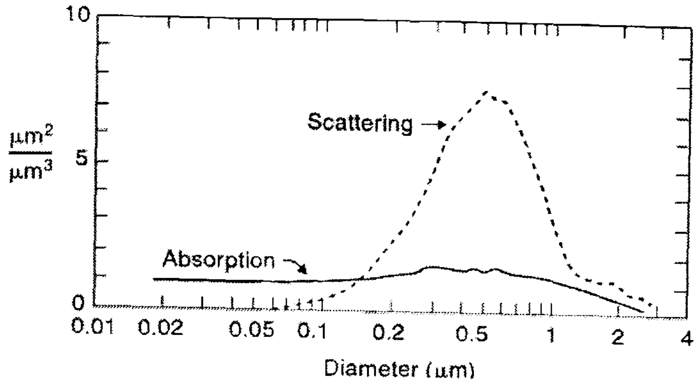


Figure 4.4: *Scattering and absorption cross section per unit volume as a function of particle diameter (Vallero 2008).*

4.4 Haze in Malaysia

4.4.1 Haze Monitoring

In Malaysia, haze monitoring is carried out by the Malaysian Meteorological Department and Department of Environment Malaysia in terms of visibility and air quality index (API) respectively.

Visibility

Visibility measurement is carried out by the Malaysian Meteorological Department on a daily basis through a network of 149 monitoring stations. For public convenience, haze severity is categorised into five levels; visibilities more than 10 km represent 'clear', 5 to 10 km visibilities represent 'moderate', 2 to 5 km visibilities represent 'hazy', 0.5 to 2 km visibilities represent 'very hazy' and visibilities less than 0.5 km represent 'extremely hazy' (Table 4.2).

Table 4.2: *Visibility levels used by the Malaysian Meteorological Department.*

Severity	Horizontal Visibility (km)
Clear	> 10
Moderate	5 – 10
Hazy	2 – 5
Very hazy	0.5 – 2
Extremely hazy	< 0.5

Air Pollution Index

The Department of Environment Malaysia operates a network of 51 stations, where 36 stations are in West Malaysia (or Peninsular Malaysia) (Figure 4.5) and 15 in East Malaysia. Due to the potential harm to human health, five main pollutants are measured, viz. SO_2, NO_2, CO, O_3 and PM_{10} (Department of Environment 1997). Based on their locations and the types of pollutant measured, the stations are categorised into Residential (20 stations), Industrial (12 stations), Traffic (1 station), Background (1 station) and PM_{10} (2 stations). The difference between these categories is the types of pollutant measured (Table 4.3).

Table 4.3: *Station categories and the type of pollutants measured (Department of Environment 2010).*

Category	SO_2	NO_2	CO	O_3	PM_{10}
Industrial	X	X	-	-	X
Residential	X	X	X	X	X
Traffic	X	X	-	X	X
Background	X	X	X	X	X
PM_{10}	-	-	-	-	X

In the API system, the air quality levels are categorised into: good (0 – 50), moderate (51 – 100), unhealthy (101 – 200), very unhealthy (201 – 300), hazardous (300 – 500) and emergency (> 500) (Table 4.4). The API value reported for a given time period represents the highest API value among all pollutants during that particular time period; the predominant pollutant during haze episodes is PM_{10} (Department of Environment 1997; Heil and Goldammer 2001; Mahmud 2009).

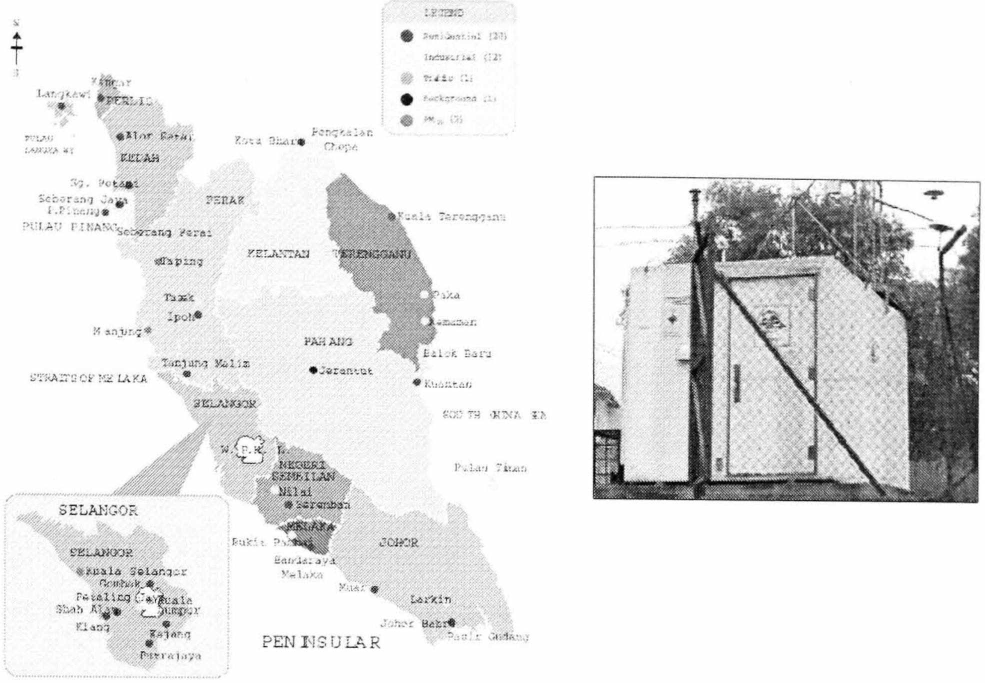


Figure 4.5: *Location of air quality monitoring stations in West Malaysia (left) with an enlarged version of Selangor state (sub-section in the lower left) and a typical monitoring station (right) (Department of Environment 2004).*

Table 4.4: *API status, level of pollution and health measures (Department of Environment 1997).*

API	Status	Level of Pollution	Health Measure
0 – 50	Good	Low, no ill effects on health.	No restriction of activities to all groups.
51 – 100	Moderate	Moderate, no ill effects on health.	No restriction of activities to all groups.
101 – 200	Unhealthy	Mild aggravation of symptoms and decreased exercise tolerance in persons with heart or lung disease.	Restriction of outdoor activities for high-risk persons. General population should reduce vigorous outdoor activity.
201 – 300	Very Unhealthy	Significant aggravation of symptoms and decreased exercise tolerance in persons with heart or lung disease.	Elderly and persons with known heart or lung disease should stay indoors and reduce physical activity. General population should reduce vigorous outdoor activity. Those with any health problems to consult doctor
300 – 500	Hazardous	Severe aggravation of symptoms and endangers health.	Elderly and persons with existing heart or lung disease should stay indoors and reduce physical activity. General population should reduce vigorous outdoor activity.
> 500	Emergency	Severe aggravation of symptoms and endangers health.	General population advised to follow the orders of National Security Council and always follow announcements through the mass media.

The Recommended Malaysian Air Quality Guidelines forms the basis for calculating the API and consists of two key aspects: the averaging time and the Malaysian guidelines (Table 4.5). The averaging time differs for different air pollutants and represents the period of time over which the measurements are made and recorded in running averages. For reporting purposes, the same averaging times are used: PM_{10} and SO_2 (24-hour running averages), CO (8-hour running averages), and O_3 and NO_2 (1-hour running averages) (Department of Environment 1997). The Malaysian guidelines represent the safe level for each pollutant and were derived based on human health data and recommendations from the World Health Organisation (WHO). For example, a PM_{10} concentration of 150 $\mu g/m^3$ corresponds to 100 API (2.1), and is the upper limit for the safe level; PM_{10} concentrations exceeding this are likely to cause adverse health effects (Department of Environment 1997). Conversion of the PM_{10} concentration from $\mu g/m^3$ to API can be done using the equations shown inTable 4.6.

Table 4.5: *Air quality measurement guidelines (Department of Environment 1997).*

Pollutant	Averaging Time	Malaysian Guidelines	
		(ppm)	(μgm^{-3})
O_3	1 hour	0.10	200
	8 hours	0.06	120
CO	1 hour	30	35
	8 hours	9	10
NO_2	1 hour	0.17	320
	24 hours	0.04	-
SO_2	1 hour	0.13	350
	24 hours	0.04	105
PM_{10}	24 hour	N/A	150
	1 year		50

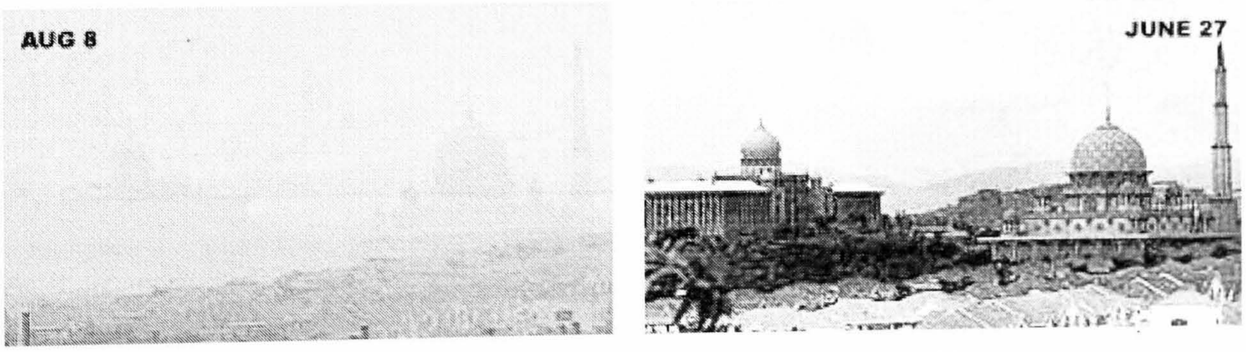
Table 4.6: *Equations for API calculation based on PM_{10} 24-hour running averages (Department of Environment 1997).*

PM_{10} concentration, C ($\mu g/m^3$)	Equation used for conversion to API
$C \leq 50$	$API = C$
$50 < C \leq 350$	$API = 50 + [(C-50) \times 0.5]$
$350 < C \leq 420$	$API = 200 + [(C-350) \times 1.43]$
$420 < C \leq 500$	$API = 300 + [(C-420) \times 1.25]$
$C \geq 500$	$API = 400 + (C-500)$

4.4.2 Climatological Behaviour of Haze in Malaysia

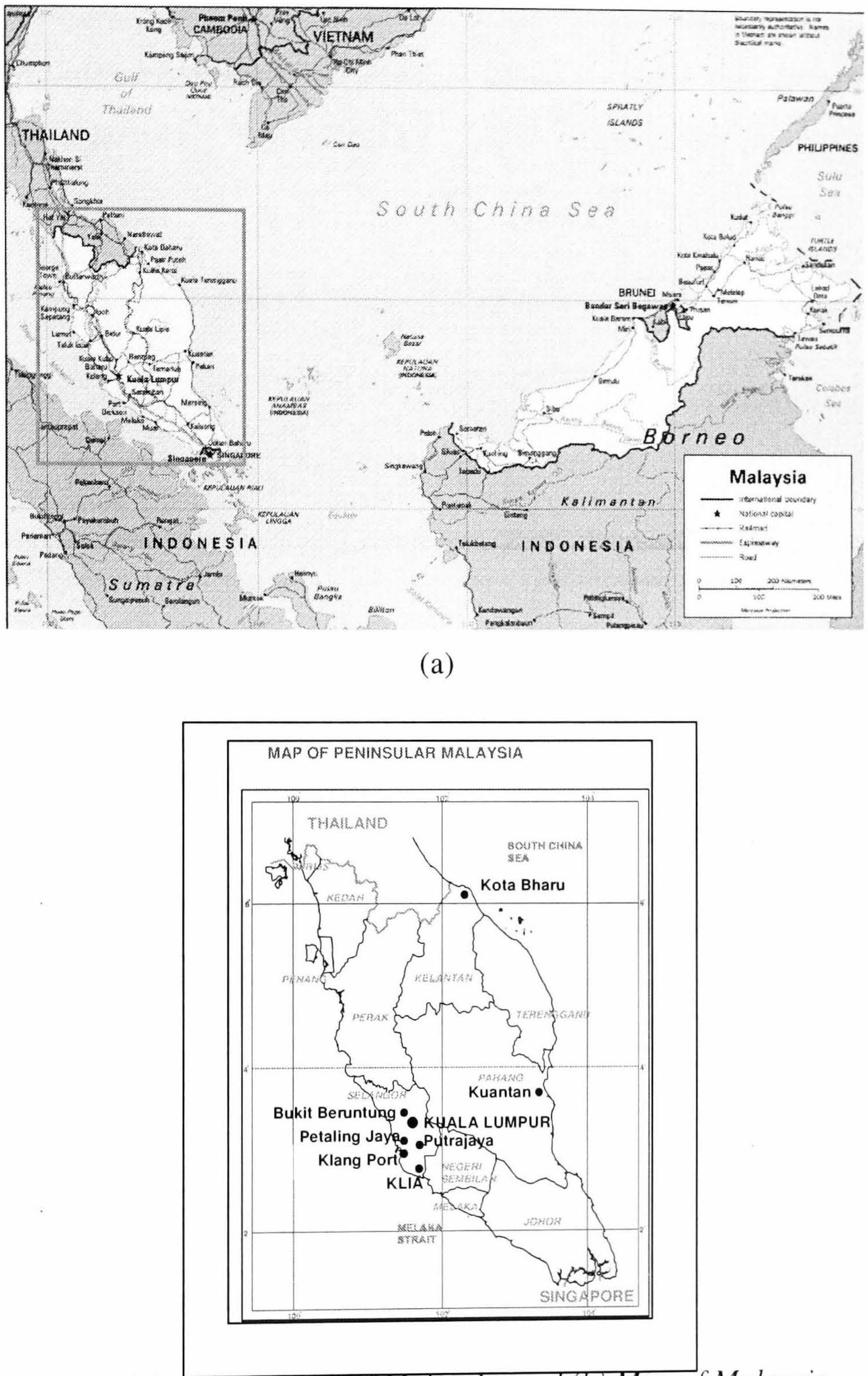
Malaysia has a typical tropical monsoon climate characterized by uniformly high mean temperature (approximately 27°C), with a relatively high mean annual rainfall (exceeding 2000 mm per year) and humidity (70% - 90%) throughout the year. The wind over the country is generally mild and variable. However, there are some periodic changes in the wind flow patterns that describe the two monsoon seasons namely the north-east monsoon, known as the wet season (November to March) and the south-west monsoon, known as the dry season (June to September). The remaining months i.e. April to May and October to November are known as the transitional periods. Because the wind comes from the south-west and there is much less rain during the south-west monsoon and the second transitional period, smoke from the forest fires in Sumatra remains suspended in the atmosphere for a long time and drifts to Malaysia, causing haze.

During the 2005 haze episode, the haze caused a drop in visibility in most places in Malaysia. Figure 4.6 shows photos of clear and hazy conditions in Putrajaya, the federal administrative centre of Malaysia, located about 30 km from Kuala Lumpur (Figure 4.7) . Due to the hazardous properties of the haze constituents, a sudden increase in respiratory and eye-related illnesses cases was reported. The drop in visibility conditions also badly affected economy-related activities including tourism, transportation, fisheries and production sectors, which caused a big loss to Malaysia.



(a) (b)

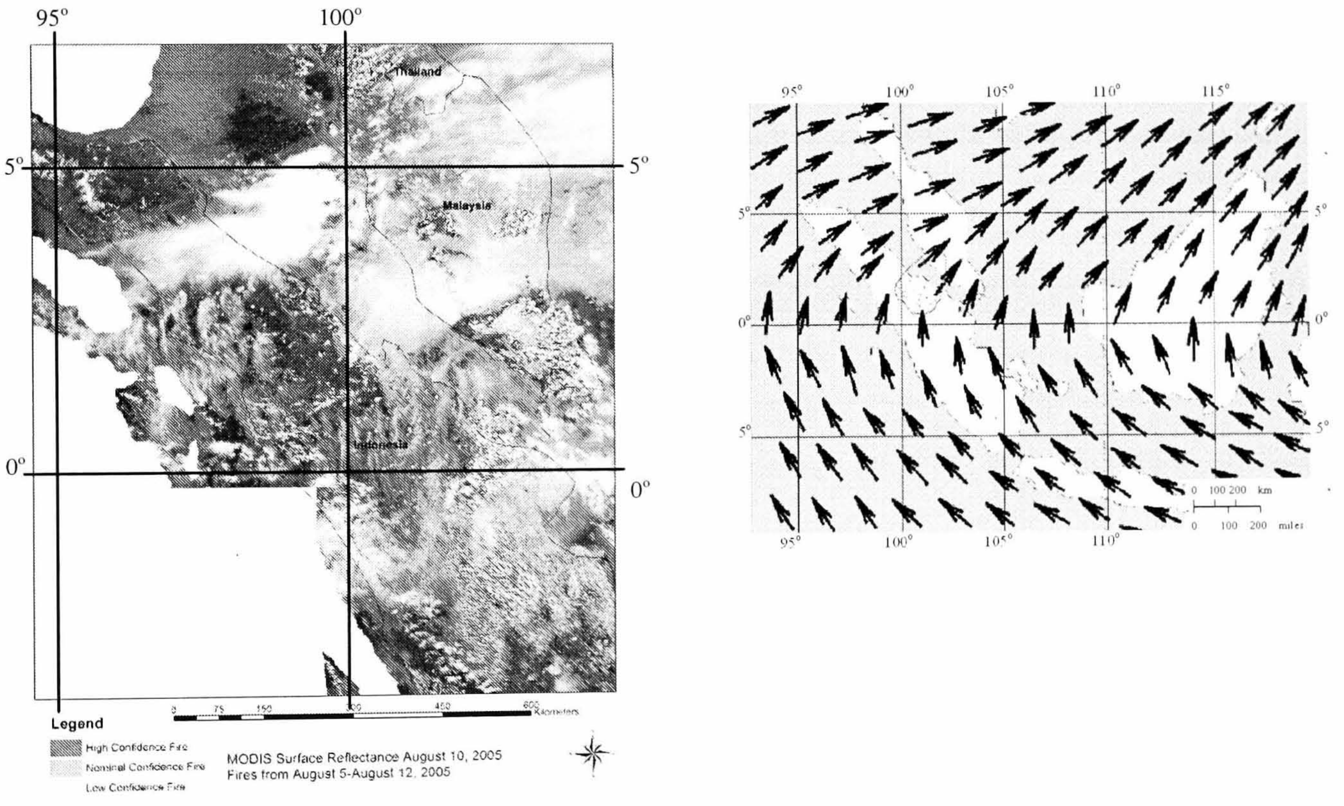
Figure 4.6: *The Prime Minister's Department (left building) and the Putra Mosque (right building) in Putrajaya, the federal administrative centre of Malaysia during (a) hazy (8 August 2005) and (b) clear (27 June 2005) (The Star Online 2005).*



(a)

Figure 4.7: *(a) Location of Malaysia and (b) Map of Malaysia.*

Figure 4.8 shows (a) the haze situation over Malaysia and Indonesia on August 10, 2005 and (b) the corresponding wind pattern. Also shown in Figure 4.8(a) are the forest fire hot spots in Sumatra from August 5 to 12 as observed by MODIS; more than 660 hot spots were detected during this period (Mahmud 2009). It is clear that the haze that blanketed Malaysia was due to Indonesian forest fires. It can be seen that the western parts are much more hazy than the eastern parts of Malaysia because they are nearer to the burning areas. On 10 August 2005, it was reported that the visibility at Petaling Jaya (0.6 km visibility) and Klang Port (0.7 km visibility), Selangor (located in the western of Malaysia) dropped to less than 1 km; while the conditions in Kuantan, Pahang (2.4 km visibility) and Kota Bharu, Kelantan (9 km visibility) (located in the eastern of Malaysia) were better.

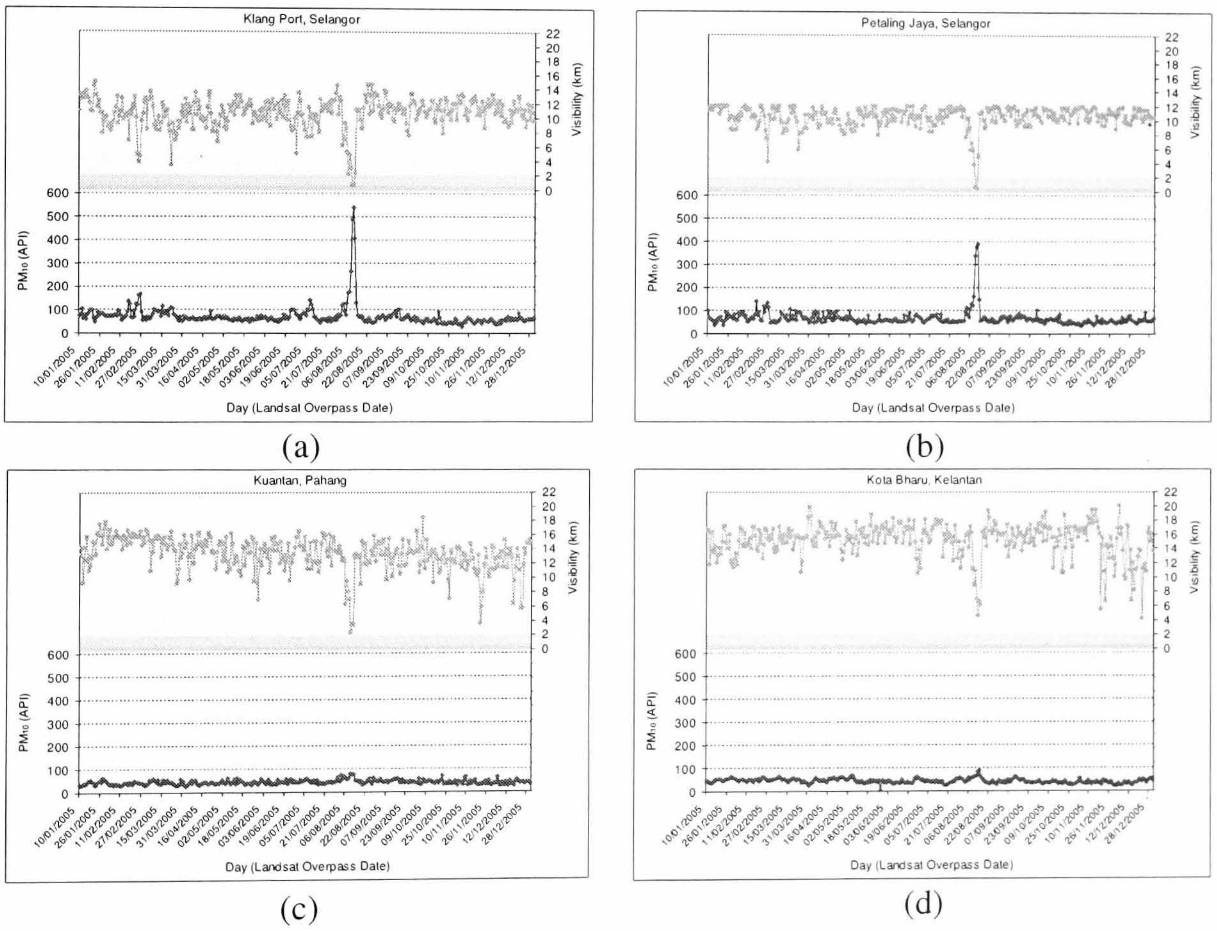


(a)

(b)

Figure 4.8: *(a) The haze condition on August 10, 2005, with active fires from August 5 to 12, 2005, observed by MODIS; masked in white are mostly ocean areas (Henipavirus Ecology Collaborative Research Group 2010) and (b) the Southeast Monsoon wind pattern on August 10, 2005 (Ahmad and Hashim 2002).*

Figure 4.9 shows visibility and PM_{10} intensity against Landsat overpass date in 2005 for Klang Port, Petaling Jaya, Kuantan and Kota Bharu (see Figure 4.7 for location). The sudden increase in PM_{10} and drop in visibility in August 2005, particularly in Klang Port and Petaling Jaya, is associated with the occurrence of haze in that year. It can be seen that extreme haze occurred between 6 and 22 August 2005. Klang Port and Petaling Jaya, which are located on the west of Malaysia (with average visibility and PM_{10} concentration approximately 11 km and 70 API respectively) experienced lower visibility and higher PM_{10} intensity than Kuantan and Kota Bharu (with average visibility and intensity approximately 14 km and 40 API respectively) which are located on the east. Since extremely hazy and very hazy conditions are quite rare in Malaysia, we are more concerned on a more frequently occurring conditions, i.e. moderate; in Chapter 5 the haze removal will be tested onto an image with moderate haze.



(a)

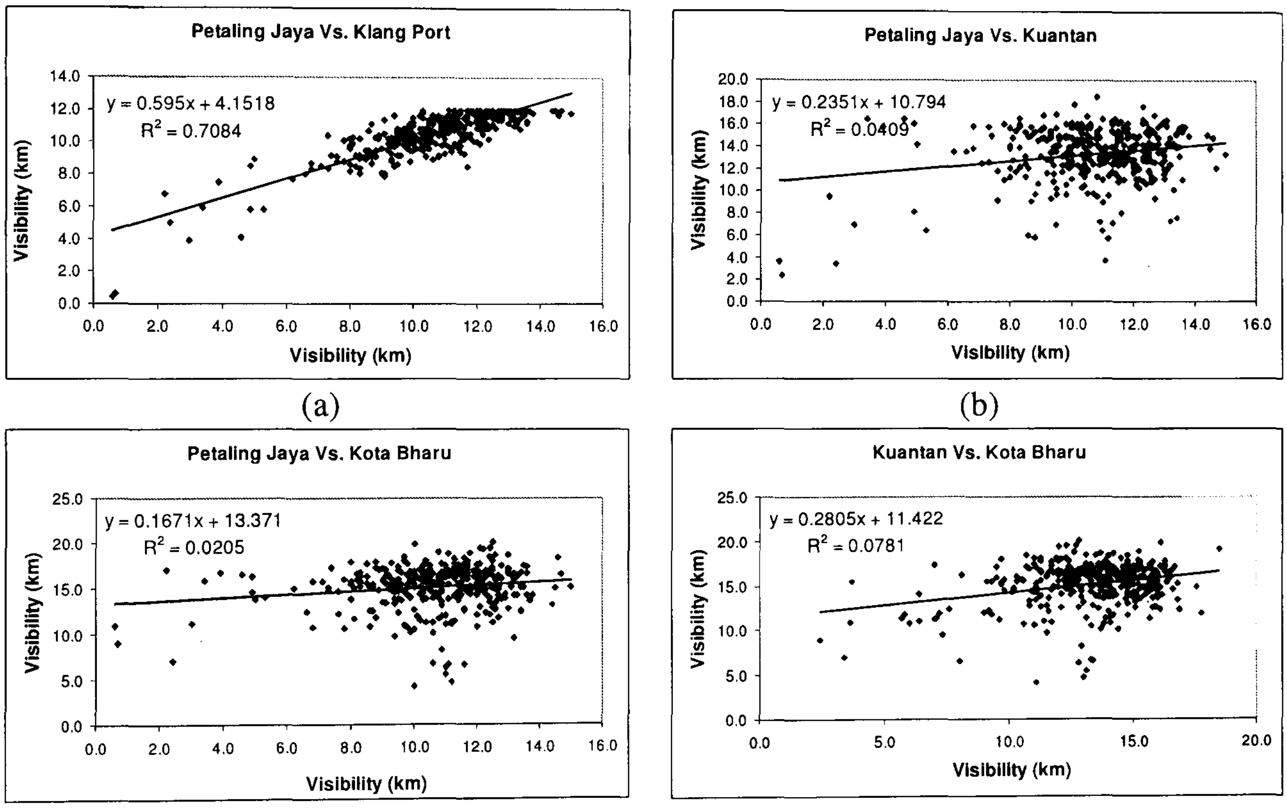
(b)

(c)

(d)

Figure 4.9: *Visibility and PM_{10} intensity for (a) Klang Port, (b) Petaling Jaya, (c) Kuantan and (d) Kota Bharu stations. White, yellow, green, violet and red colours indicate clear (above 10 km), moderate (5 – 10 km), hazy (2 – 5 km), very hazy (0.5 – 2 km) and extremely hazy (less than 0.5 km) conditions respectively.*

Figure 4.10 shows scatterplots of visibility for Petaling Jaya vs. Klang port, Petaling Jaya vs. Kuantan, Petaling Jaya vs. Kota Bharu and Kuantan vs. Kota Bharu, together with linear fits to these plots. It is clear that the visibility correlation between nearby stations, i.e. Petaling Jaya and Klang Port (0.708) is much higher than non-neigbouring stations, i.e. Petaling Jaya and Kuantan (0.04), Petaling Jaya and Kota Bharu (0.02) and Kuantan and Kota Bharu (0.08). In this thesis, the testing of the developed haze removal method will be carried out over Bukit Beruntung area, by using PM_{10} measurements from Petaling Jaya station.



(a) (b)

(c) (d)

Figure 4.10: *Visibility correlation for (a) Petaling Jaya vs. Klang port, (b) Petaling Jaya vs. Kuantan, (c) Petaling Jaya vs. Kota Bharu and (d) Kuantan vs. Kota Bharu.*

4.5 Model for Determining Satellite Observed Radiance under Hazy Conditions

The observed radiance, L , that reaches the sensor for a cloudless and haze-free atmosphere can be expressed as (Kaufman and Sendra 1988):

$$L = L_S + L_D + L_O \qquad \text{... (4.9)}$$

(see Figure 4.11) where L_S is the radiance reflected by the target and directly transmitted through the atmosphere towards the satellite (this gives most information about the target on the Earth's surface), L_D is the radiance reflected from the surface and then scattered by the atmosphere to the sensor (this diffuses radiation between different pixels and thus reduces the spatial variation of the upward radiance), and L_O is the radiance scattered into the sensor's field of view by the atmosphere itself (caused by the atmospheric constituents that exist during clear sky conditions) without reaching the surface - this is independent of surface reflectance and increases the image brightness.

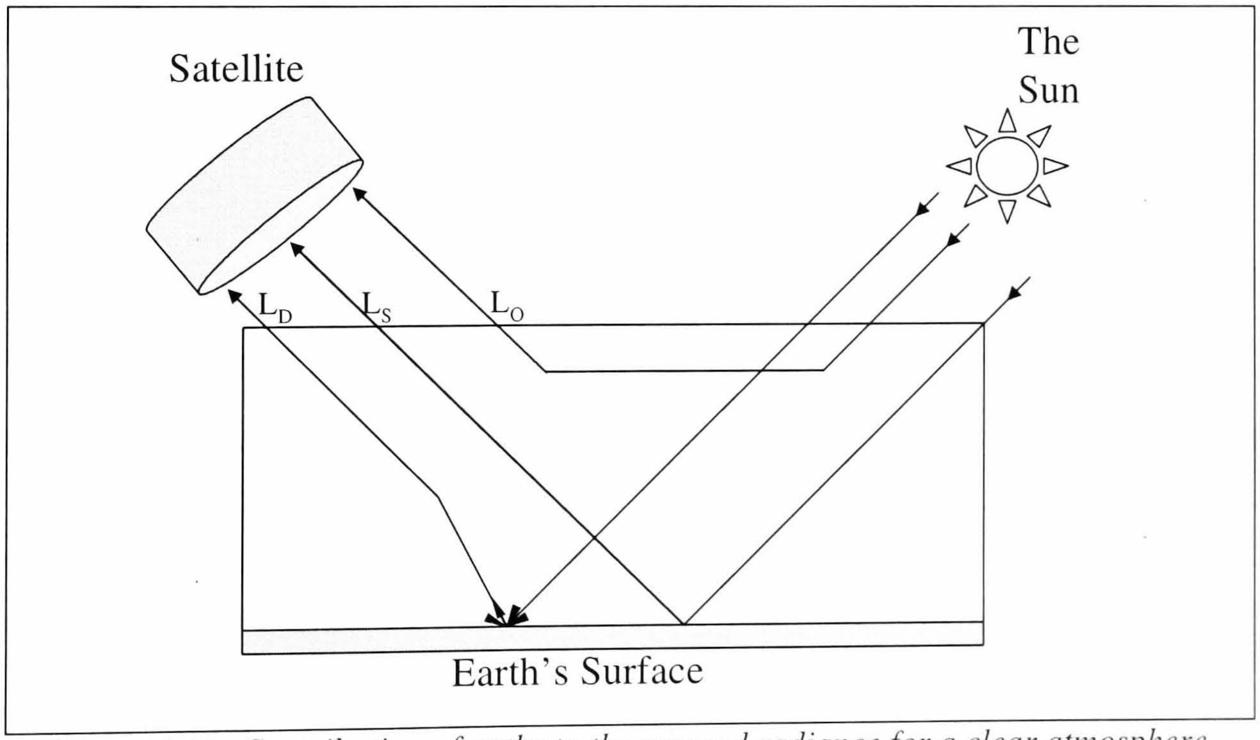


Figure 4.11: *Contribution of paths to the upward radiance for a clear atmosphere (Kaufman and Sendra 1988).*

To account for haze, Equation (4.9) is modified as follows (Figure 4.12):

$$L = L_S + L_D + L_O + L_H \qquad \ldots (4.10)$$

where L_H is the radiance caused by the haze layer. It is independent of the surface reflectance and increases image brightness for dark targets, but decreases it for bright targets. Similarly, equation (4.10) can also be expressed as:

$$L = L_S + L_D + (L_O + L_H) = L_S + L_D + L_P \qquad \ldots (4.11)$$

where $L_P = L_O + L_H$ is known as path radiance. Figure 4.13 and Figure 4.14 show L_S, L_D and L_P as a function of visibility; a more detailed discussion is given later.

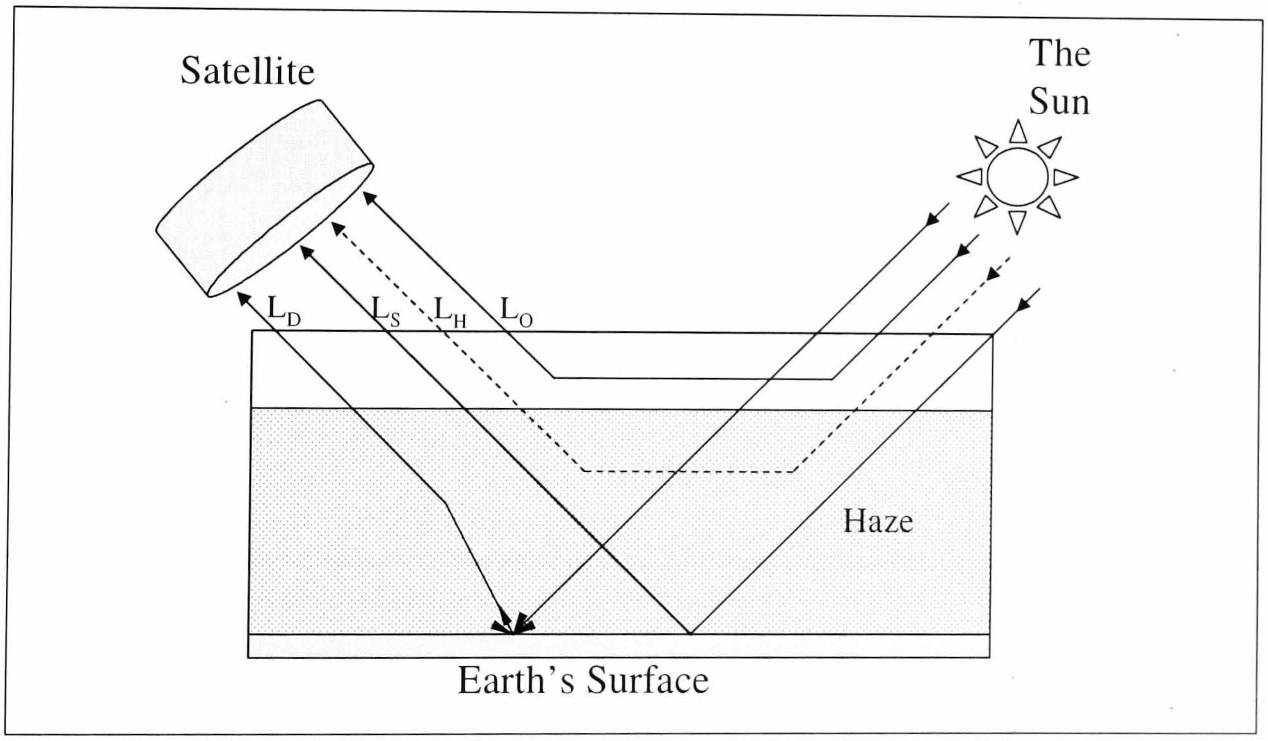


Figure 4.12: *Modification of Figure 4.11 for hazy conditions.*

As will be shown in Section 4.6, the contribution of L_D is insignificant since it is much weaker than the other components (Kaufman and Fraser 1983) and therefore can be neglected, hence:

$$L \approx L_S + (L_O + L_H) = L_S + L_P \qquad \text{... (4.12)}$$

We introduce a model for simulation of hazy data in which Equation (4.12) is written as:

$$L(V) \approx \left(1-\beta^{(1)}(V)\right)L_S(\infty) + \left[L_O + \beta^{(2)}(V)L_H(0)\right] \qquad \text{... (4.13)}$$

where $L_S(V) = \left(1-\beta^{(1)}(V)\right)L_S(\infty)$ and $L_P(V) = L_O + \beta^{(2)}(V)L_H(0)$. V is the visibility, $L(V)$ is the target radiance at visibility V km, so $L_S(\infty)$ is the radiance of pure target (i.e. the atmospheric components are assumed insignificant), $L_H(0)$ is for pure haze and the weightings $\beta^{(1)}(V)$ and $\beta^{(2)}(V)$ are given by:

$$\beta^{(1)}(V) = 1 - \frac{L_S(V)}{L_S(\infty)} \qquad \text{... (4.14)}$$

$$\beta^{(2)}(V) = \frac{L_P(V) - L_O}{L_H(0)} \qquad \text{... (4.15)}$$

In Equation 4.13, during a clear day, $V = \infty$, $\beta^{(1)}(\infty) = \beta^{(2)}(\infty) = 0$, therefore $L(\infty) = L_S(\infty) + L_O$. For very thick haze, $V = 0$, $\beta^{(1)}(0) = \beta^{(2)}(0) = 1$, so $L(0) = L_O + L_H(0)$. In other words, during clear and very hazy conditions, the radiance observed by the satellite sensor is actually the radiance of true signal and pure haze respectively added with L_O. Between the two extremes, $L_S(V)$ is influenced by $\left(1-\beta^{(1)}(V)\right)$ due to atmospheric absorption, while $L_P(V)$ is influenced by $\beta^{(2)}(V)$ due to atmospheric scattering.

We can estimate $L_S(\infty)$ from a clear Landsat dataset, while determination of $L_H(0)$, $\beta^{(1)}(V)$ and $\beta^{(2)}(V)$ are discussed in the following section.

4.6 Simulating a Hazy Dataset

4.6.1 Background on the 6SV1 Code

6SV1 is the vector version of the 6S (Second Simulation of the Satellite Signal in the Solar Spectrum) (Kotchenova et al. 2006; Kotchenova and Vermote 2006; Vermote et al. 1997), though it also works in scalar mode. The vector version is introduced to account for radiation polarisation, due to Rayleigh scattering in a mixed molecular-aerosol atmosphere, which is to be used when performing atmospheric correction (Kotchenova et al. 2006). In our study, the 6SV1 is used in simulating haze effects, therefore the radiation polarisation effect is assumed negligible. Hence, our interest is in the scalar mode of 6SV1, which is similar to 6S.

6S makes use of the Successive Order of Scattering (SOS) algorithm to calculate Rayleigh scattering, aerosol scattering and coupling of scattering-absorption. In the SOS, the atmosphere is divided into a number of layers, and the radiative transfer equation is solved for each layer with an iterative approach (Vermote et al. 1997; Kotchenova et al. 2006).

In terms of the aerosol model, a refined computation of the radiative properties of basic components (e.g. soot, oceanic, dust-like and water soluble) and additional aerosol components (e.g. stratospheric, desertic and biomass burning) is included. It also has a spectroscopic database for important gases in the 0.25 – 4.0 μm spectrum region, and is able to simulate the TOA signal for both Lambertian and non-Lambertian targets.

For a Lambertian uniform target, the apparent radiance ρ^* is calculated using (Vermote et al. 1997):

$$\rho^*(\theta_s,\theta_v,\phi_s-\phi_v)=\rho_{a+r}(\theta_s,\theta_v,\phi_s-\phi_v)+\frac{\rho_t}{1-\rho_t S}T^{\downarrow}(\theta_s)T^{\uparrow}(\theta_v) \quad \text{... (4.16)}$$

where $\rho_{a+r}(\theta_s,\theta_v,\phi_s-\phi_v)$ is the intrinsic atmospheric reflectance associated with aerosol (τ_a) and Rayleigh scattering (τ_r), ρ_t is the surface reflectance, S is the atmospheric spherical albedo, $T^{\downarrow}(\theta_s)$ is the total downward transmission and $T^{\uparrow}(\theta_v)$ is the total upward transmission. θ_s and θ_v are the solar and satellite zenith angle respectively and ϕ_s and ϕ_v are solar and satellite azimuth angle respectively.

$$T^{\downarrow}(\theta_s)=e^{-\tau/\mu_s}+t_d(\theta_s) \quad \text{... (4.17)}$$

$$T^{\uparrow}(\theta_v)=e^{-\tau/\mu_v}+t_d(\theta_v) \quad \text{... (4.18)}$$

where $\tau=\tau_a+\tau_r$ is the atmospheric optical thickness associated with aerosol (τ_a) and Rayleigh scattering (τ_r), μ_s and μ_v are $\cos(\theta_s)$ and $\cos(\theta_v)$ respectively and t_d is the diffuse transmittance due to molecules and aerosols. Substituting (4.18) into (4.16), we have:

$$\rho^*(\theta_s,\theta_v,\phi_s-\phi_v)=\rho_{a+r}(\theta_s,\theta_v,\phi_s-\phi_v)+\frac{T^{\downarrow}(\theta_s)}{1-\rho_t S}\left(\rho_t e^{-\tau/\mu_v}+\rho_t t_d(\theta_v)\right) \quad \text{... (4.19)}$$

For a non-uniform surface:

$$\rho^*(\theta_s,\theta_v,\phi_s-\phi_v)=\rho_{a+r}(\theta_s,\theta_v,\phi_s-\phi_v)+\frac{T^{\downarrow}(\theta_s)}{1-\rho_e S}\left(\rho_t e^{-\tau/\mu_v}+\rho_e t_d(\theta_v)\right) \quad \text{... (4.20)}$$

where ρ_e is the environmental reflectance and can be expressed as:

$$\rho_e=\frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty}\rho(r,\phi)\frac{dF(r)}{dr}dr\,d\phi \quad \text{... (4.21)}$$

The intrinsic atmospheric reflectance, $\rho_{a+r}(\theta_s,\theta_v,\phi_s-\phi_v)$, is computed using:

$$\rho_{a+r}(\theta_s,\theta_v,\phi_s-\phi_v)=\rho'_{a+r}(\theta_s,\theta_v,\phi_s-\phi_v)+\left(1-e^{-\tau/\mu_s}\right)\left(1-e^{-\tau/\mu_v}\right)\Delta(\tau) \qquad \text{... (4.22)}$$

where $\rho'_{a+r}(\theta_s,\theta_v,\phi_s-\phi_v)$ is the single-scattering contribution associated with aerosol and molecule scattering and $\left(1-e^{-\tau/\mu_s}\right)\left(1-e^{-\tau/\mu_v}\right)\Delta(\tau)$ accounts for higher orders of scattering. The atmospheric spherical albedo S is calculated using:

$$S=\frac{1}{4+3\tau}\left[3\tau-4E_3(\tau)+6E_4(\tau)\right] \qquad \text{... (4.23)}$$

where $E_3(\tau)$ and $E_4(\tau)$ are exponential integrals depending on τ.

4.6.2 Radiance Calculation Using the 6SV1

To exploit Equation (4.13), we need to determine $L_H(0)$, $\beta^{(1)}(V)$ and $\beta^{(2)}(V)$. To do so, we use the 6SV1 code to calculate satellite apparent radiance, path radiance, signal radiance and diffuse radiance, for bands 1, 2, 3, 4, 5, and 7 of the Landsat satellite. The atmospheric and aerosol model used were Tropical Model and Biomass Burning respectively. The latter accounts for haze that originates from forest fires. Initially, visibility was varied from 0.5 to 100 km. Figure 4.13 shows satellite apparent radiance, path radiance, signal radiance and diffuse radiance as a function of visibility.

It is obvious that for bands 1, 2 and 3, the path radiance is higher than the diffuse radiance at all visibilities. For bands 4, 5 and 7, the path and diffuse radiance are about the same but the signal radiance is comparatively much higher. This shows that at shorter wavelengths the haze effects are significant and dominated by the path radiance; while at longer wavelengths, the haze effects are almost negligible due to the much higher signal radiance.

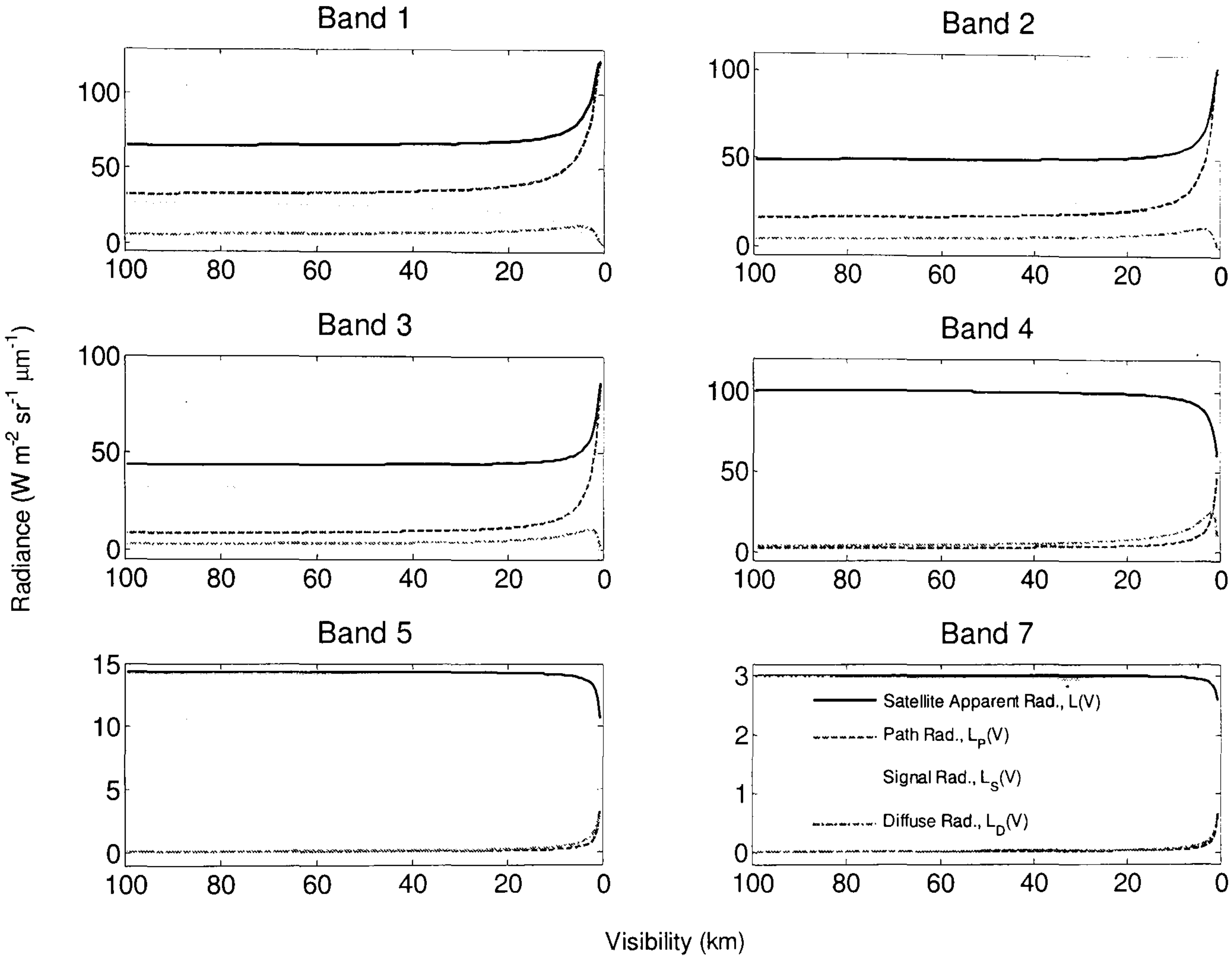


Figure 4.13: *Satellite apparent radiance, path radiance, signal radiance and diffuse radiance as a function of visibility as visibility runs from 0.5 to 100 km.*

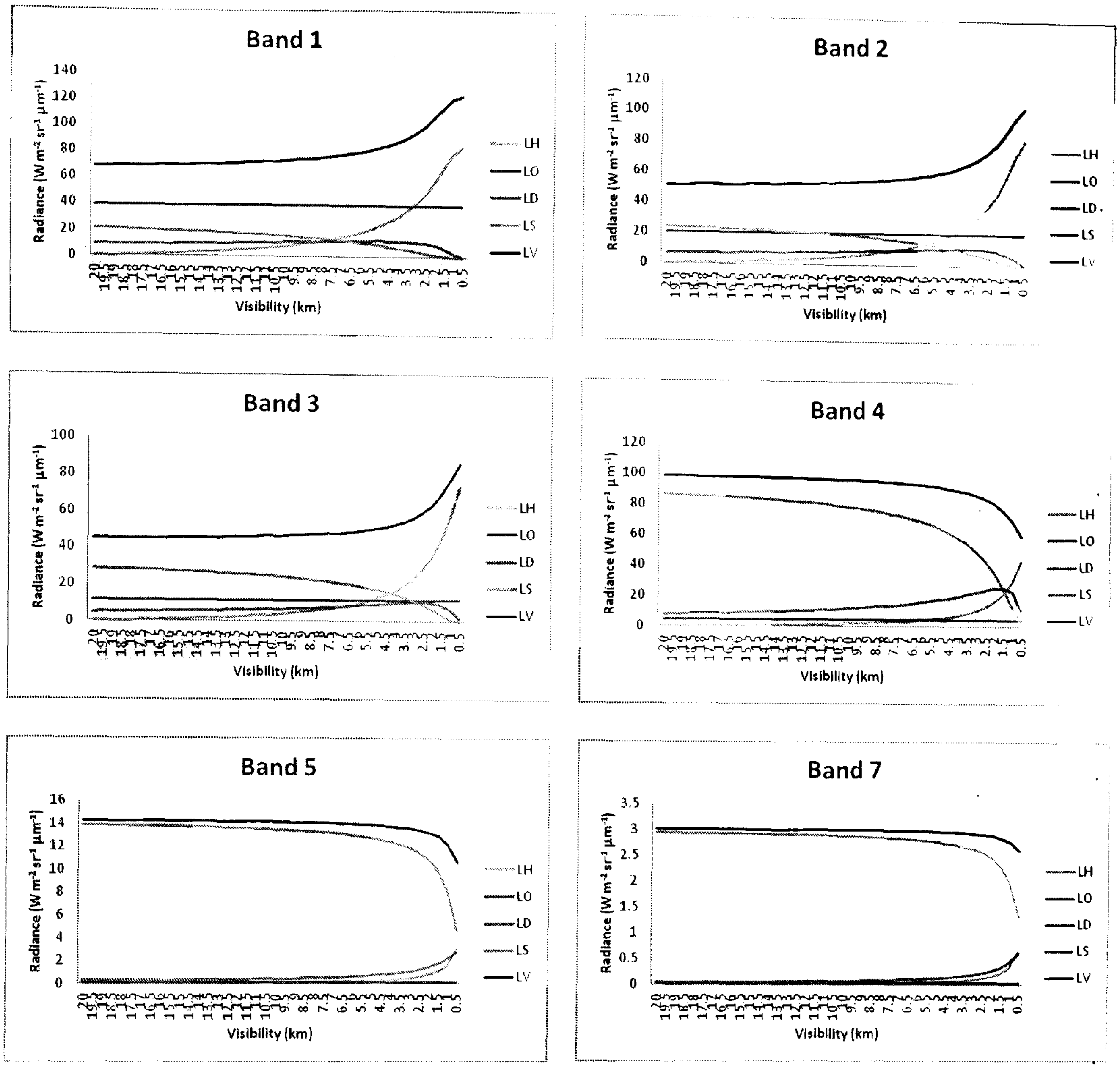


Figure 4.14: *Same as Figure 4.13 but for visibility running from 0.5 to 20 km visibility.*

It is also observed that the major impact on the radiances occurs for visibilities less than 20 km (Figure 4.14). Hence, we make the approximations $L_S(\infty) \approx L_S(20)$ and $L_P(\infty) = L_O \approx L_P(20)$. Also, in 6SV1, calculations cannot be made for visibilities less than 0.5 km, so we assume $L_S(0) \approx L_S(0.5)$ and $L_P(0) \approx L_P(0.5) = L_O + L_H(0.5)$. Thus, Equation 4.13 can be written as:

$$L(V) \approx \left(1-\beta^{(1)}(V)\right)L_S(20)+\left[L_O+\beta^{(2)}(V)L_H(0.5)\right] \qquad \ldots(4.24)$$

Determination of the Weightings, $\beta^{(1)}(V)$ and $\beta^{(2)}(V)$

Calculation of $\beta^{(1)}(V)$ and $\beta^{(2)}(V)$ is carried out using:

$$\beta^{(1)}(V)=1-\frac{L_S(V)}{L_S(20)} \qquad \text{... (4.25)}$$

$$\beta^{(2)}(V)=\frac{L_P(V)-L_O}{L_H(0.5)} \qquad \text{... (4.26)}$$

Plots of $\beta^{(1)}(V)$ and $\beta^{(2)}(V)$ against visibility for bands 1, 2, 3, 4, 5 and 7 are shown in Figure 4.15. The difference between $\beta^{(1)}(V)$ and $\beta^{(2)}(V)$ is due to the difference between the true signal attenuation and the pure haze weighting, which is more obvious in shorter (i.e. lower-numbered bands) than longer wavelengths (i.e. higher-numbered bands). $\beta^{(1)}(V)$ is higher than $\beta^{(2)}(V)$ for shorter wavelengths but this is not the case for longer wavelengths. For shorter wavelengths, $\beta^{(1)}(V)$ and $\beta^{(2)}(V)$ show a small difference at longer visibilities. The difference increases as visibility decreases to about 4 km, but then decreases towards 0 km. For longer wavelengths, the difference is only obvious at very short visibilities. At very long visibilities, the difference is not significant, hence use of a single weighting can be considered for these visibilities. Nevertheless, to account for the entire visibilities, the use of different weightings is appropriate due to the inconsistency between the signal attenuation and haze weighting, particularly at moderate and longer visibilities.

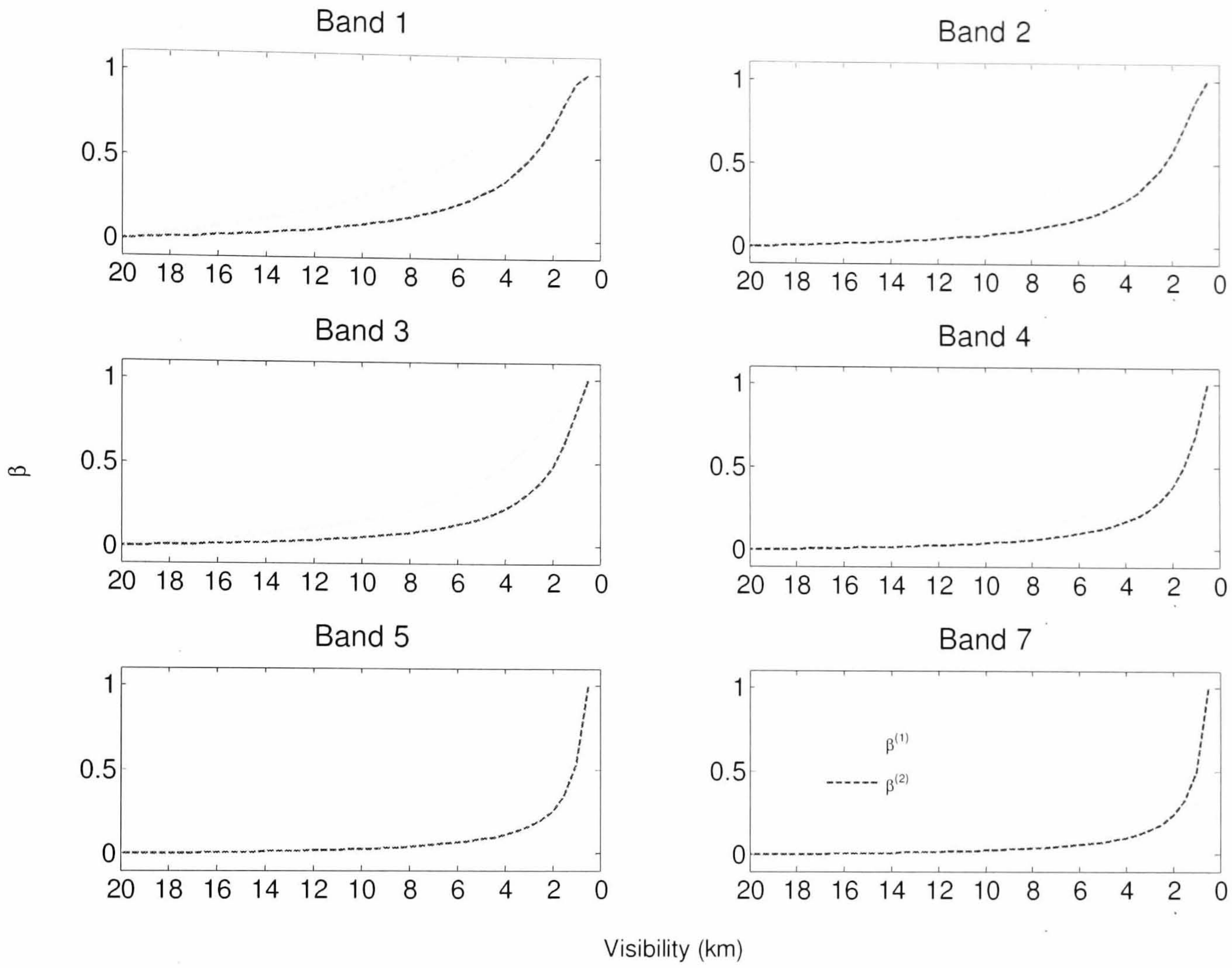


Figure 4.15: *Plots of $\beta^{(1)}(V)$ and $\beta^{(2)}(V)$ against visibility.*

4.6.3 Simulation of the Haze Radiance Component, $\beta^{(2)}(V)L_H(0)$

The interaction between solar radiation and the haze constituents affects the spectral measurements made from a satellite remote sensing system and degrades the quality of the images, for example by reducing the contrast between objects and ultimately making them inseparable. Statistically, for a single spectral measurement, haze modifies the mean and standard deviation, but, in a multispectral system the covariance structure of the multispectral measurements is also affected. These effects need to be simulated when using Equation 4.13 to model haze observed in multispectral Landsat data. We assume that haze can be treated as a random noise that can be modelled as a multivariate

Gaussian random variable with mean $\boldsymbol{\mu}$ (haze radiance) and covariance matrix C (covariance structure of haze observed by Landsat).

Representation of Haze as A Multivariate Gaussian Random Variable

In simulating haze in remote sensing dataset, we need to take into account its spatial distribution and spectral correlation. In practice, these parameters are difficult to measure due to dynamic behaviour of haze. Here, we assume haze to be spatially uncorrelated, so that it can sensibly simulate haze effects. If the haze was to be spatially correlated, it will appear as patches with high spatial correlation, which are unlikely to represent a real haze condition.

Hence, haze can be modelled with an N-dimensional Gaussian probability density function which has the form:

$$P(\mathbf{X}) = (2\pi)^{-\frac{N}{2}} (|C|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^{t} C^{-1} (\mathbf{X}-\boldsymbol{\mu})\right) \qquad \ldots (4.27)$$

where $\mathbf{X}$ is an N-dimensional random variable representing the N Landsat bands, i.e. the observed haze, $\boldsymbol{\mu}$ is the vector of means and C is the N x N covariance matrix .

In order to use this model, we need estimation of $\boldsymbol{\mu}$ and C. We set $\boldsymbol{\mu} = \beta^{(2)}(V)\mathbf{L}_H(0.5)$ where $\mathbf{L}_H(0.5)$ is obtained from the atmospheric haze radiance for 0.5 km visibility, calculated using 6SV1. Since haze is usually considered as thin cloud in satellite spectral measurements (Ji 2008; Moro and Halounova 2007; Lu 2007; Zhang et al. 2002), the covariance, C, is measured using observations of cloud in Landsat bands 1, 2, 3, 4, 5 and 7 from 11 February 1999, as given in Table 4.7. In this table, there is quite a strong correlation between bands, especially amongst band 1, 2 and 3 as well as between band 5 and 7, due to their similar spectral measurement properties. The covariance will play an important role in simulating the effects of haze on land cover classification.

In our study, 11 sets of haze layers, representing 11 different visibilities, were then generated, where each layer consists of 758 x 792 pixels, which is the same size as the clear dataset.

Table 4.7: *Cloud covariances (along and above the diagonal) and correlations (below the diagonal) calculated from Landsat bands 1, 2, 3, 4, 5 and 7.*

Band	1	2	3	4	5	7
1	432.74	771.38	706.31	440.36	84.73	29.59
2	0.80	2147.19	1872.28	1255.49	225.26	90.50
3	0.83	0.98	1692.62	1103.61	207.53	84.07
4	0.71	0.91	0.90	879.83	154.98	59.97
5	0.68	0.81	0.85	0.88	35.63	14.64
7	0.53	0.72	0.76	0.75	0.91	7.31

4.6.4 Simulation of A Hazy Dataset

The haze component is assumed to be independent of the signal component, so the hazy dataset can be synthesised by adding a weighted pure haze, $\beta^{(2)}(V)L_H(0.5)$ to the weighted true signal $(1-\beta^{(1)}(V))L_S(20)$. Here, the true signal component is estimated from Landsat-5 TM dataset (from 11 February 1999 with 20 km visibility, based on the average of 6 stations within 5 to 60 km from the centre of the scene, i.e. Klang Port, Petaling Jaya, Sepang, Serdang, Tanjung Karang and Banting.

Based on Equation (4.24) and for simplicity, we define $T_i = L_S(20)$ and $H_i = L_H(0.5)$. Consequently, due to the vector-based structure of a dataset, the hazy dataset, $L_i(V)$ can be written as:

$$L_i(V) = (1-\beta_i^{(1)}(V))T_i + L_O + \beta_i^{(2)}(V)H_i \qquad \text{... (4.28)}$$

An example of hazy dataset simulation for 4 km visibility is shown in Figure 4.16. The model components, the simulated hazy band 1 and the corresponding histogram are shown in the left, middle and right column respectively.

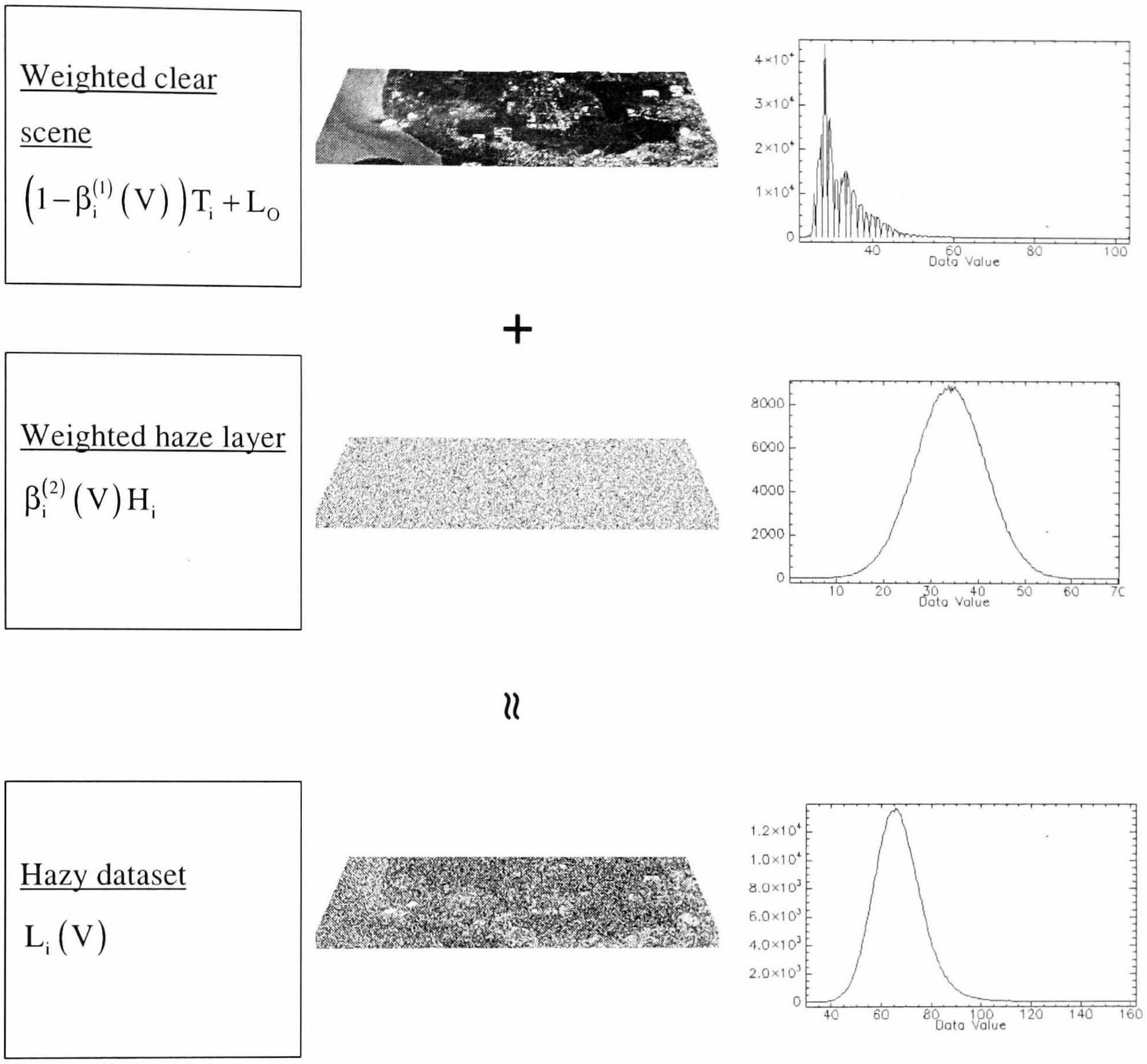
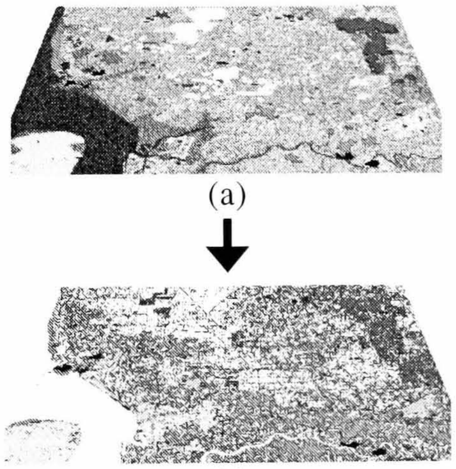


Figure 4.16: *Process of integrating a hazy layer with a clear data from band 1 to produce a 4 km (V = 4) visibility hazy dataset; the corresponding histograms and model terms are shown to the right and left respectively.*

Hazy scenes were generated for 11 visibilities (20, 18, 16, 14, 12, 10, 8, 6, 4, 2 and 0 km) and six Landsat bands (band 1, 2, 3, 4, 5 and 7).

4.7 ML Classification on the Simulated Hazy Dataset

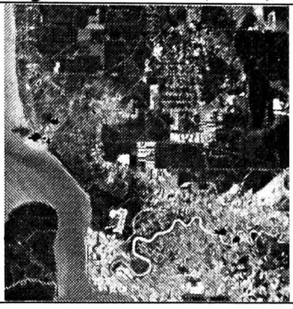
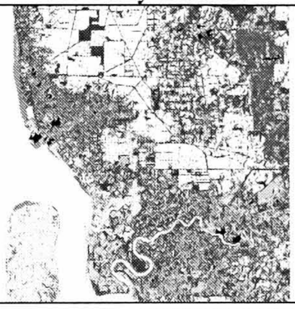
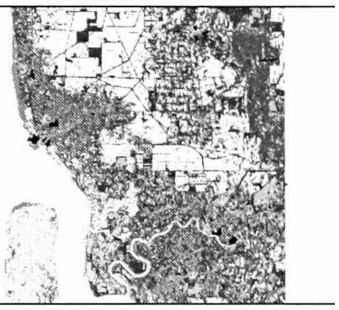
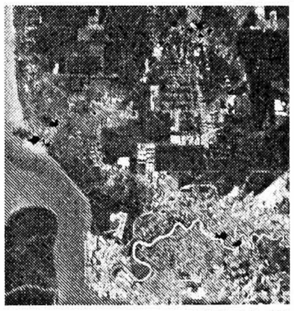
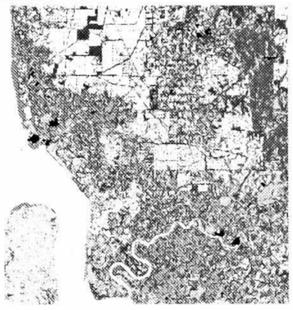
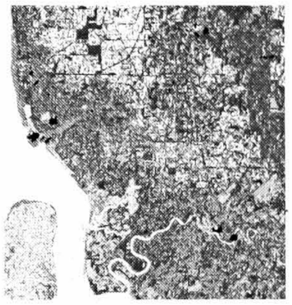
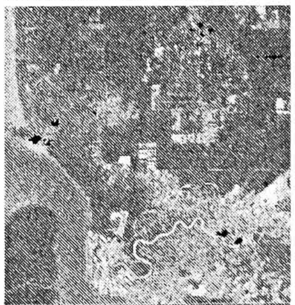
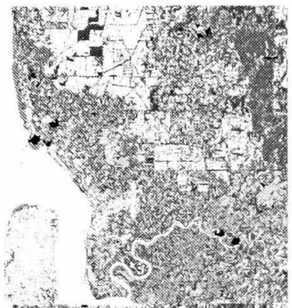
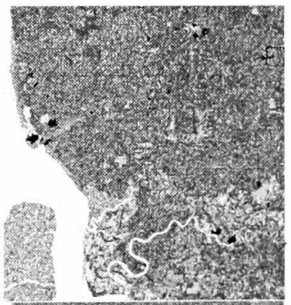
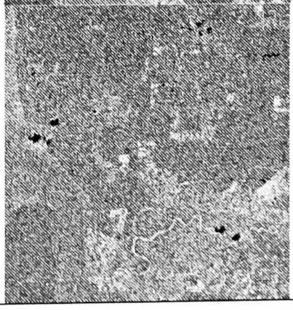
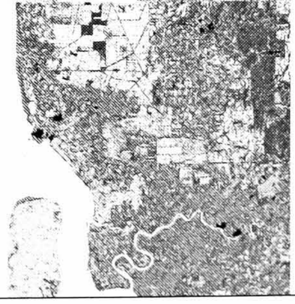
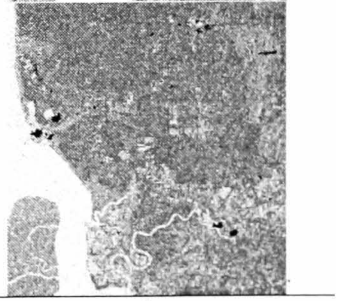
ML classification was carried out using all 6 bands to produce 11 classes, viz. coastal swamp forest, dry land forest, oil palm, rubber, cleared land, sediment plumes, water, coconut, bare land, urban and industry (see Section 3.4). To carry out ML classification on the hazy scenes, we need training pixels within the hazy scene. For this purpose, the ROIs for different land classes (different colours) that were applied on the clear scene were used as a template. Figure 4.17 shows (a) patches of ROIs for different land classes (indicated by different colours) overlaid on bands 4, 5 and 3 (assigned to red, green and blue) of a 4 km visibility hazy scene used for selecting training pixels from the hazy scene and (b) the ML classification. The sampling procedure and colours associated with these classes are described in Section 3.4.


(a)

(b)

Figure 4.17: *(a) Patches of different colours are ROIs for different land classes used for selecting training pixels from a 4 km visibility hazy scene and (b) the ML classification.*

Figure 4.18 shows the 4 km hazy datasets before and after ML classification for visibilities 20 km, 10 km, 6 km, 4 km, 2 km and 0 km. These visibilities are chosen to visually show the transition from clear to very hazy conditions. It is obvious that the effects of haze become more severe on bands 3, 2, and 1 (assigned to the red, green and blue channels respectively) as visibility decreases (images on the left). These bands are displayed since they are more affected by the haze than the bands with longer wavelengths. Therefore, classification is much more influenced by the effects of haze in shorter than longer wavelengths. The middle images show the corresponding ML classification using training pixels from the hazy dataset itself. As expected, the ML classification performance degrades as visibility drops. Some of the classes are clearly inseparable at 2 km visibility. The images on the right show ML classification using training pixels from the clear dataset, which will be discussed further in Section 4.6.

	Before Classification	After ML Classification	
	(i) Band (channel) 3 (red), 2 (green) and 1 (blue)	(ii) Training pixels taken from the hazy dataset	(iii) Training pixels taken from the clear dataset
(a) 20 km visibility			
(b) 10 km visibility			
(c) 6 km visibility			
(d) 4 km visibility			

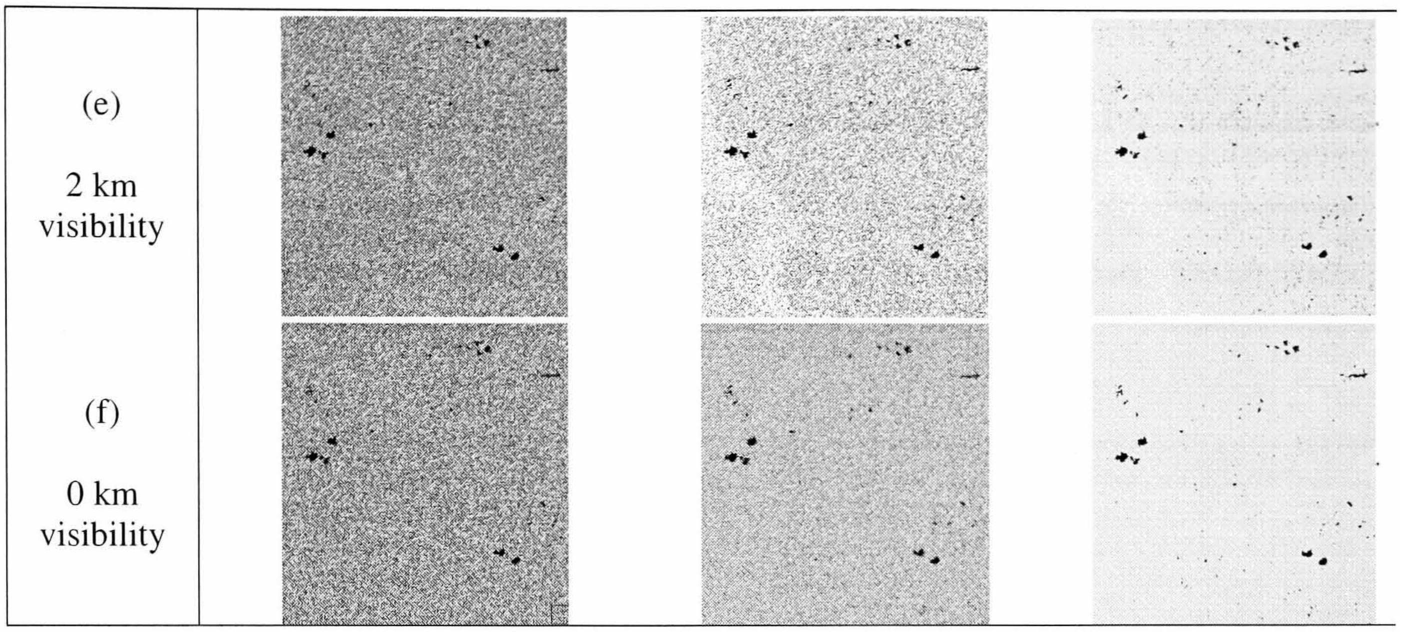


Figure 4.18: *Bands 3, 2 and 1 assigned to red, green and blue channels respectively (left), the ML classification using training pixels from hazy datasets (middle) and ML classification using training pixels from clear datasets for (a) 20 km (clear), (b) 10 km, (c) 6 km, (d) 4 km, (e) 2 km and (f) 0 km visibility. Note that images a(ii) and a(iii) are the same and are displayed for convenience. Black patches are cloud and its shadow (masked black).*

Statistical Analysis of Classes for the Hazy Datasets

In order to extract the statistics of the classes generated by the ML classification, the classification produced from the clear scene was used as a template to demarcate the pixels in each class and then to compute the class means and correlation between bands in hazy data. To illustrate this, Figure 4.19 shows plots of mean radiances versus bands for all classes. As expected, the means are more affected at shorter than longer visibilities. At 16 and 12 km visibility (Figure 4.19(a and b)), the difference between the original class radiance (red curve) and hazy class radiance (black curve) is very small for most classes. Similarly, there is little difference between the standard deviation of the original class radiance (red vertical bars) and the hazy class radiance (black vertical bars). At 10 and 6 km visibility, the differences are increasing but are still small (Figure 4.19(c and d)). At 4 km visibility (Figure 4.19(e)), the haze clearly increases the radiance of bands 1, 2 and 3 for most classes except for bare land and industry, which decrease. The increase in radiance tends to occur for dark classes (e.g. forests and vegetation) because the apparent radiance is dominated by the haze radiance (i.e. radiance scattered directly to the satellite's field of view). A decrease in radiance tends to occur for bright classes because the haze scatters some of the solar radiance out of the satellite's field of view before reaching the ground, and attenuates the reflected radiation on the way back. These effects become more apparent as haze severity increases (Figure 4.19(f)).

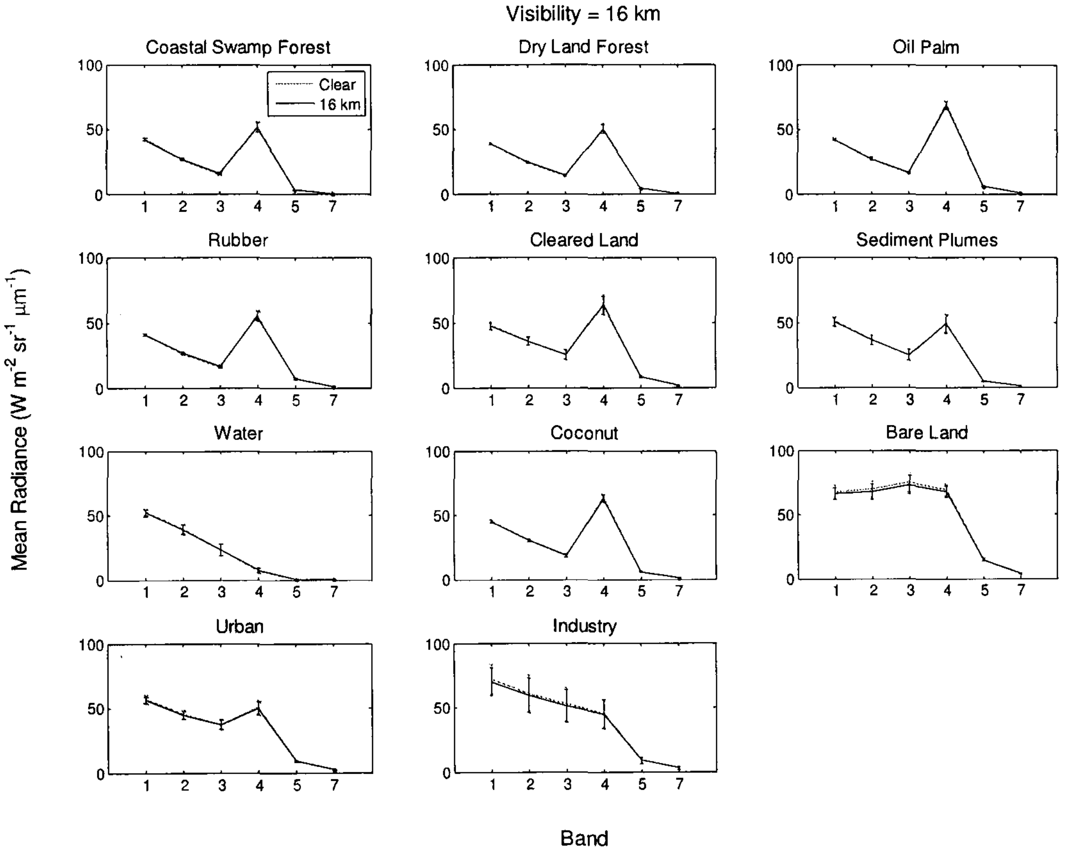
This is consistent with Equation 4.28 that shows that haze increases $L_i(V)$ through the scattering effects on $\beta_i^{(2)}(V)H_i$ but at the same time decreases $L_i(V)$ through the absorption effects on $\left(1-\beta_i^{(1)}(V)\right)T_i$. Hence, the absorption effect is proportional to T_i; therefore the brighter the surface, the higher the absorption, consequently the more $L_i(V)$ decreases. However, it should be noted that this is true in absolute terms but not relative; $\beta_i^{(1)}(V)$ does not depend on T_i.It can also be seen that most classes exhibit an increase in standard deviation as visibility reduces. In other words, the haze increases the variability in the intensity of the class pixels and consequently leads to an increase in the pixels' standard deviation. This is expected from Equation 4.28; since:

$$\begin{aligned}\mathrm{Var}\left[L_i(V)\right] &= \mathrm{Var}\left[\left(1-\beta_i^{(1)}(V)\right)T_i + L_O + \beta_i^{(2)}(V)H_i\right] = \mathrm{Var}\left[\left(1-\beta_i^{(1)}(V)\right)T_i + \beta_i^{(2)}(V)H_i\right]. \\ &= \mathrm{Var}\left[\left(1-\beta_i^{(1)}(V)\right)T_i\right] + \mathrm{Var}\left[\beta_i^{(2)}(V)H_i\right] + 2C\left[\left[\left(1-\beta_i^{(1)}(V)\right)T_i\right],\left[\beta_i^{(2)}(V)H_i\right]\right]\end{aligned}$$

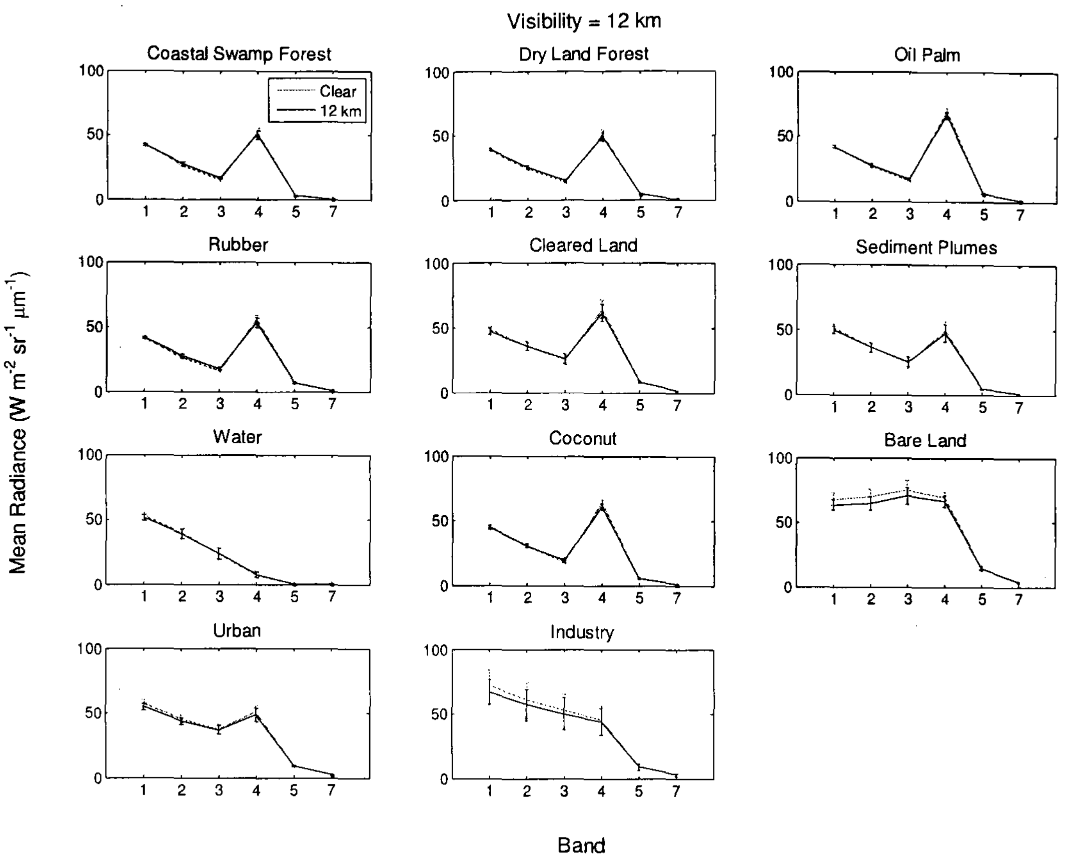
T_i is independent of H_i, so the third term equal to zero. However we cannot discard the T_i term since the target is not constant, so there still exist some variance.

$$\mathrm{Var}\left[L_i(V)\right] = \left[1-\beta_i^{(1)}(V)\right]^2 \mathrm{Var}(T_i) + \left[\beta_i^{(2)}(V)\right]^2 \mathrm{Var}(H_i)$$

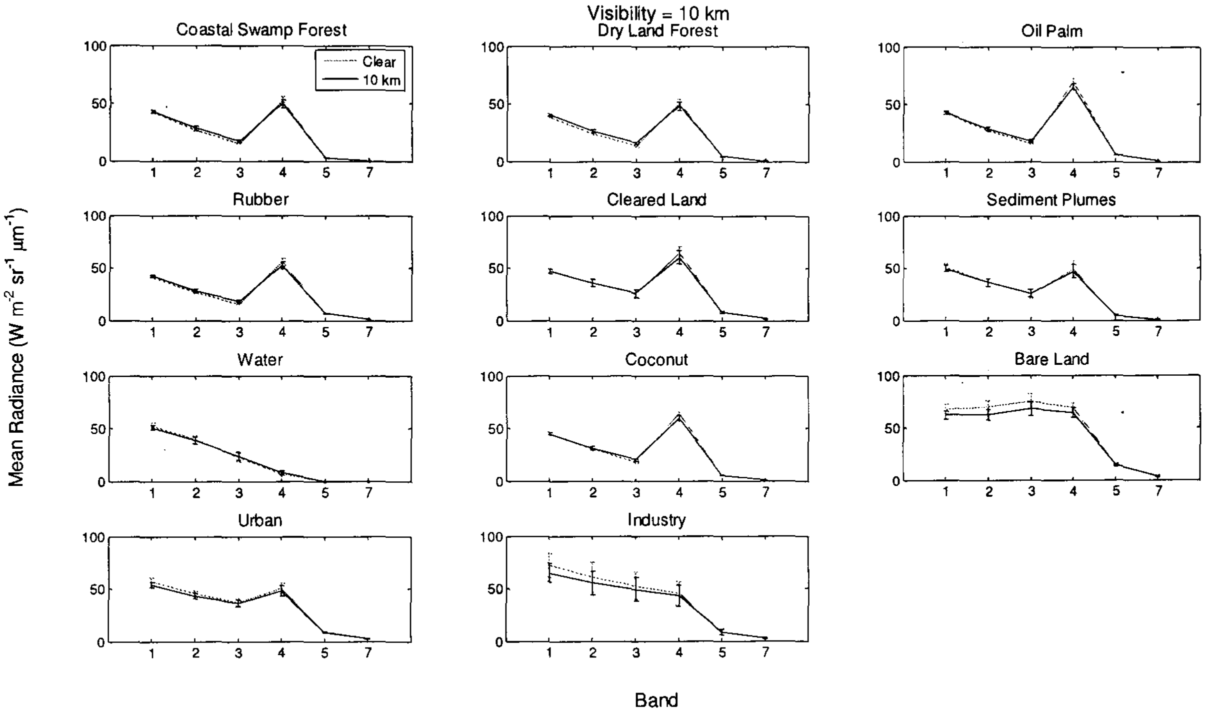
When haze gets more severe, V decreases but $\beta_i^{(1)}(V)$ and $\beta_i^{(2)}(V)$ both increase. Hence, the contribution from the target variance decreases but that from the haze increases. The balance between the two depends on target brightness because bright targets (such as industry and bare land) have larger variance. This is more noticeable in the dark classes (such as vegetation and water) due to the greater difference between the haze and dark class spectral measurement.



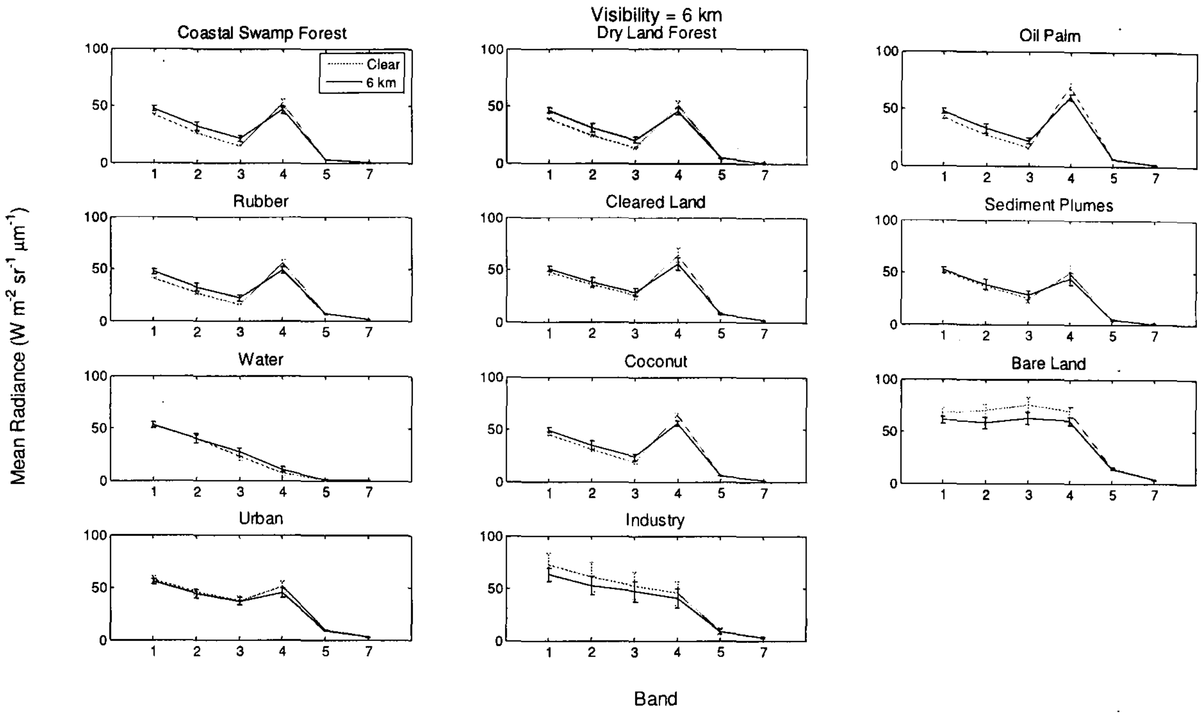
(a)


Visibility = 12 km
Coastal Swamp Forest
Clear
12 km
Dry Land Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Water
Coconut
Bare Land
Urban
Industry
Mean Radiance (W m^{-2} sr^{-1} µm^{-1})
Band
0
50
100
1 2 3 4 5 7

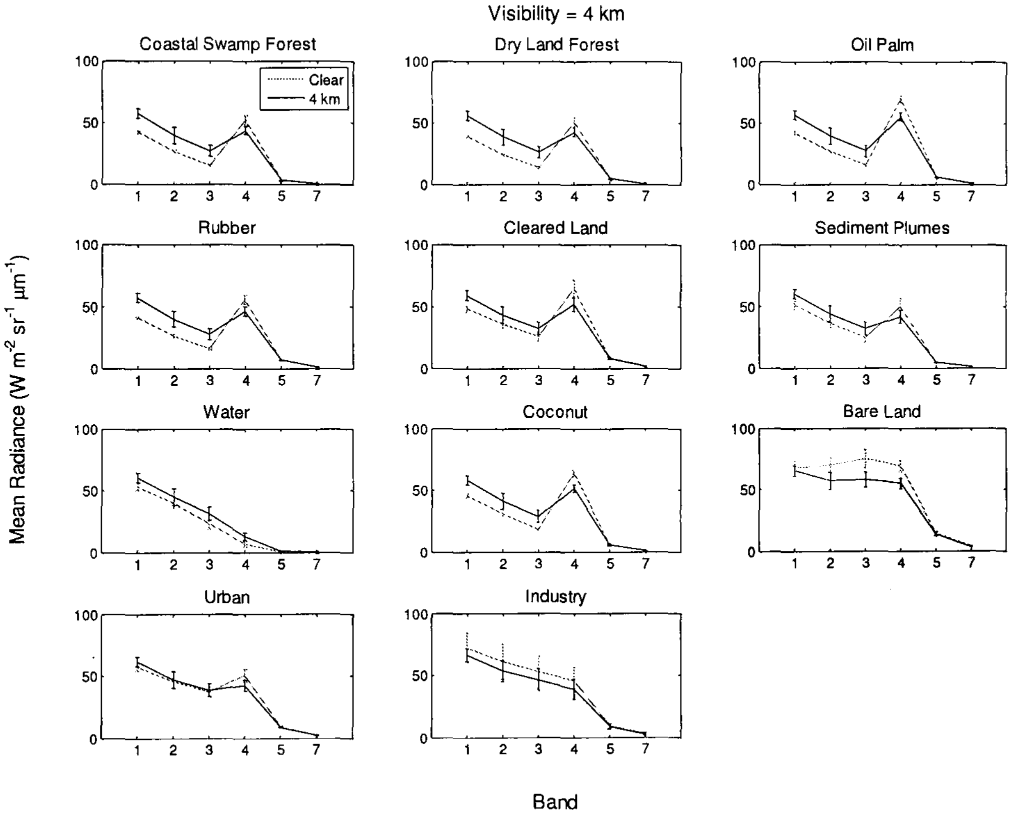
(b)


Visibility = 10 km
Coastal Swamp Forest
Clear
10 km
Dry Land Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Water
Coconut
Bare Land
Urban
Industry
Mean Radiance (W m^{-2} sr^{-1} µm^{-1})
Band
0
50
100
1 2 3 4 5 7

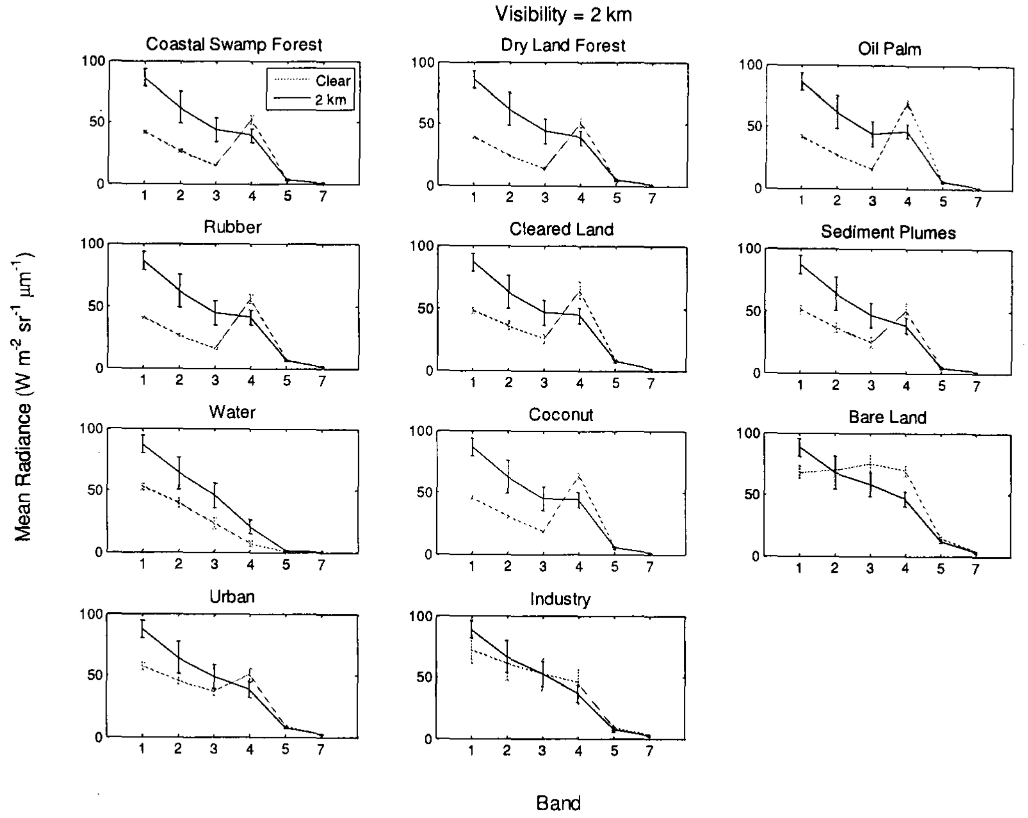
(c)


Visibility = 6 km
Coastal Swamp Forest
Dry Land Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Water
Coconut
Bare Land
Urban
Industry
Clear
6 km
100
50
0
1
2
3
4
5
7
Mean Radiance (W m^{-2} sr^{-1} µm^{-1})
Band

(d)


Visibility = 4 km
Coastal Swamp Forest
Dry Land Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Water
Coconut
Bare Land
Urban
Industry
Clear
4 km
100
50
0
1
2
3
4
5
7
Mean Radiance (W m^{-2} sr^{-1} µm^{-1})
Band

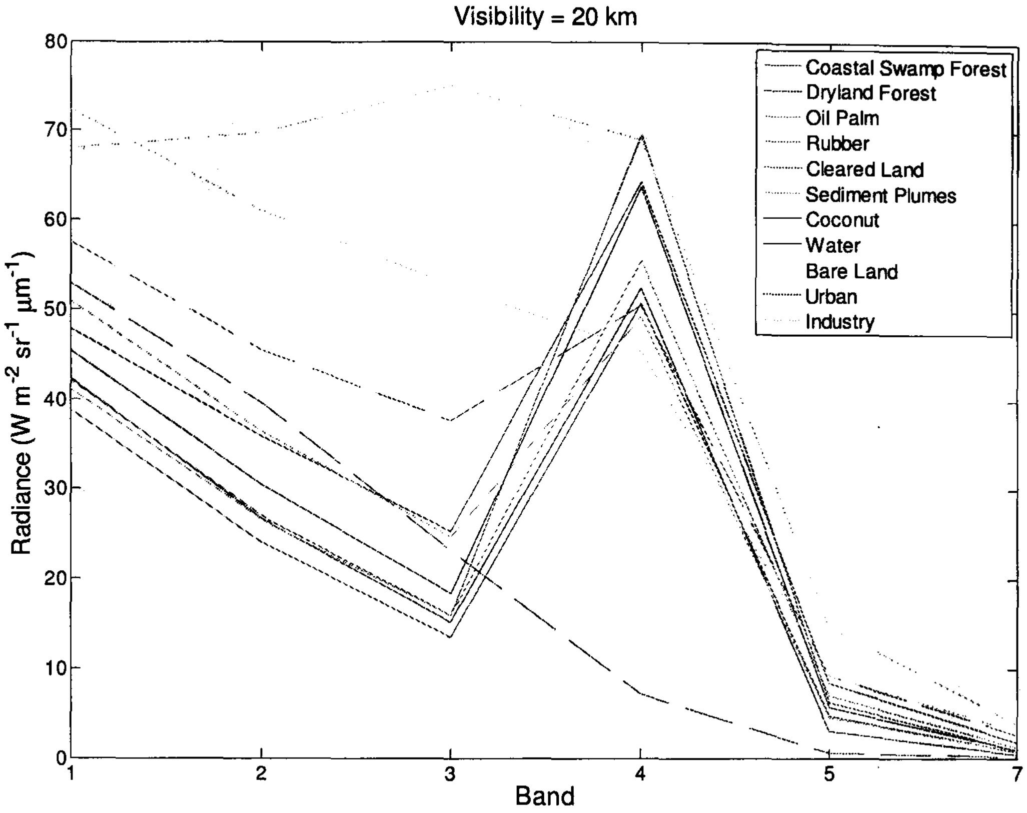
(e)



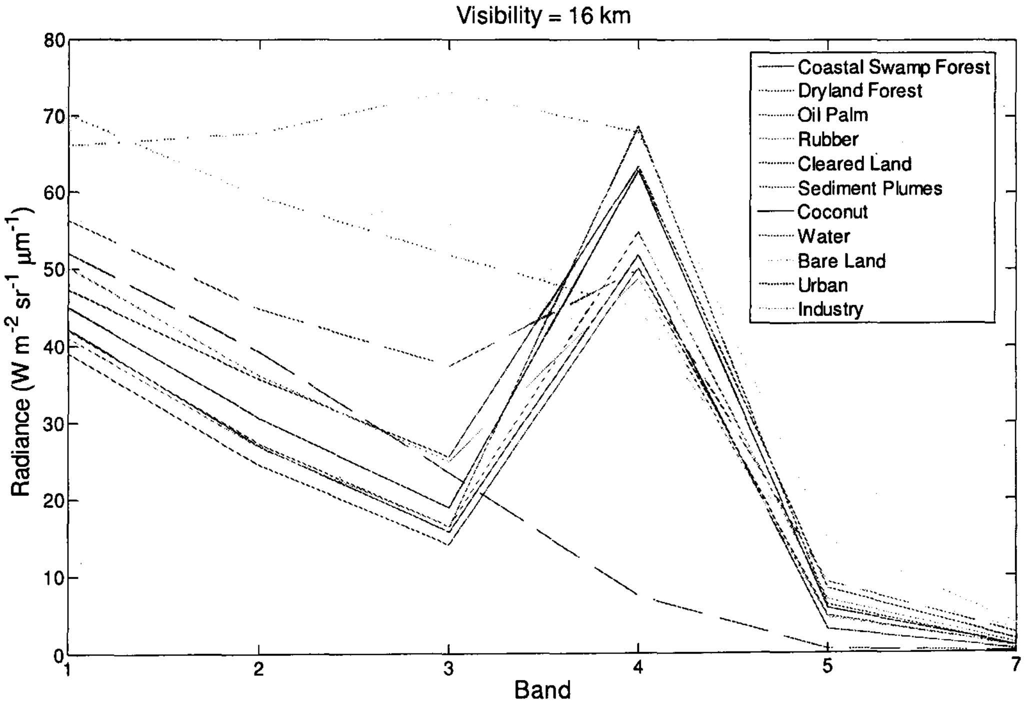
(f)

Figure 4.19: *Mean radiances versus bands of individual classes for a scene with haze (black) and without haze (red) at visibility (a) 16, (b) 12, (c) 10, (d) 6, (e) 4 and (f) 2 km. Vertical bars indicate standard deviations.*

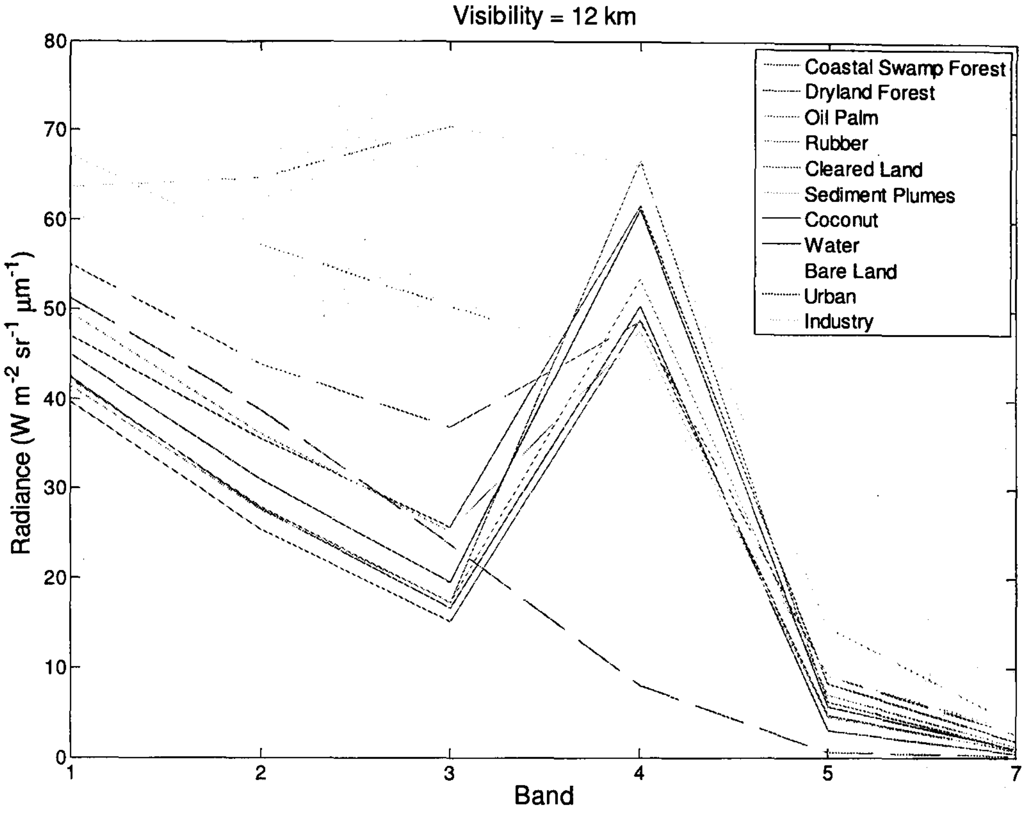
Figure 4.20 shows plots of the means of all classes versus bands for visibilities 20 km, 16 km, 12 km, 6 km, 4 km, 2 km and 0 km. At 20 km visibility (Figure 4.20(a)), all classes exhibit their true spectral signature curves, but experience little modification by haze at 16 and 12 km visibility (Figure 4.20(c and d)). At 4 and 2 km visibility (Figure 4.20(e and f)), these curves are severely modified by the haze and become inseparable as the visibility reduces to 0 km (Figure 4.20(g)). At 0 km visibility, all curves become very close, approximating the pure haze spectral signature. In other words, during no or light haze, the spectral signature of the classes are evident because the true signal radiance predominates, but as the haze gets severe, the spectral signature of the classes vanishes and is replaced by that of pure haze.


Visibility = 20 km
Radiance (W m^{-2} sr^{-1} µm^{-1})
Band
0
10
20
30
40
50
60
70
80
1
2
3
4
5
7
Coastal Swamp Forest
Dryland Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Coconut
Water
Bare Land
Urban
Industry

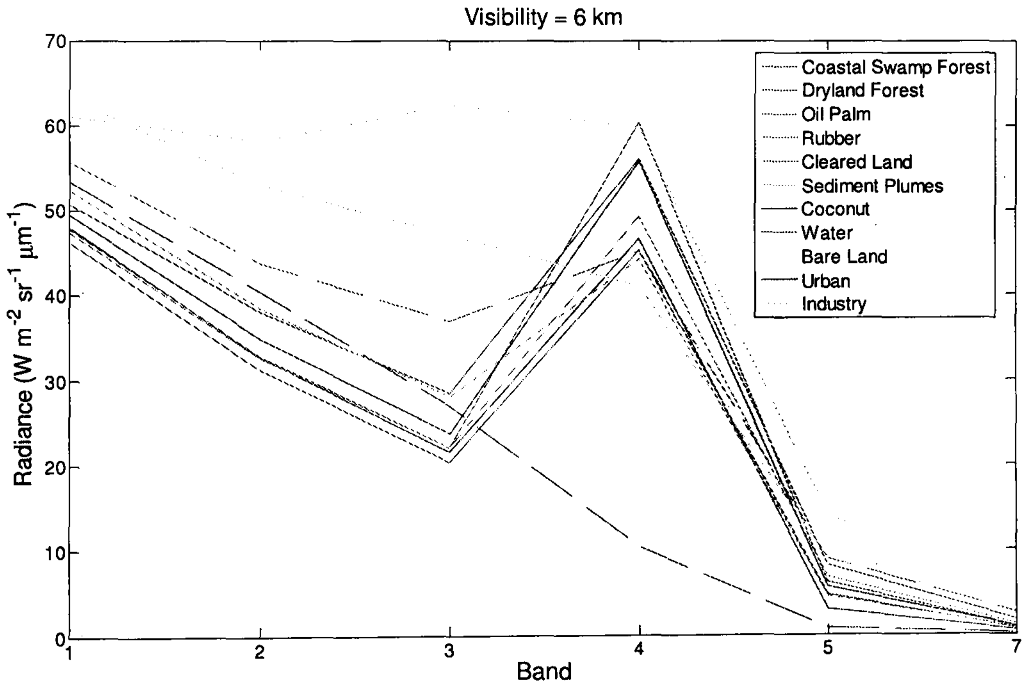
(a)


Visibility = 16 km
Radiance (W m^{-2} sr^{-1} µm^{-1})
Band
0
10
20
30
40
50
60
70
80
1
2
3
4
5
7
Coastal Swamp Forest
Dryland Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Coconut
Water
Bare Land
Urban
Industry

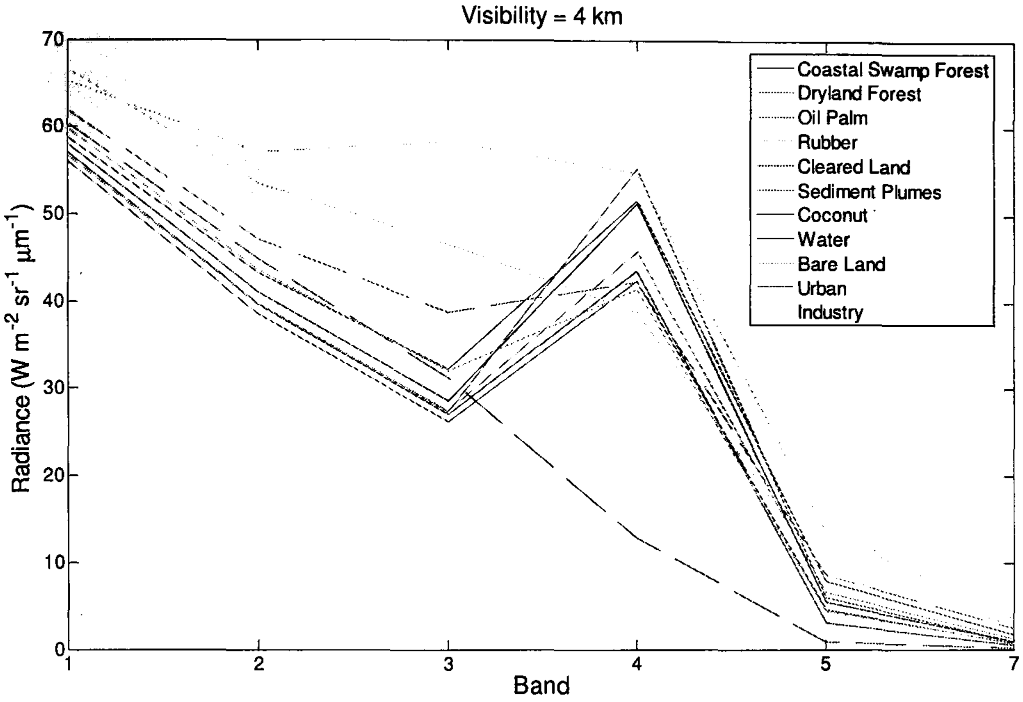
(b)


Visibility = 12 km
Radiance (W m^{-2} sr^{-1} µm^{-1})
Band
0
10
20
30
40
50
60
70
80
1
2
3
4
5
7
Coastal Swamp Forest
Dryland Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Coconut
Water
Bare Land
Urban
Industry

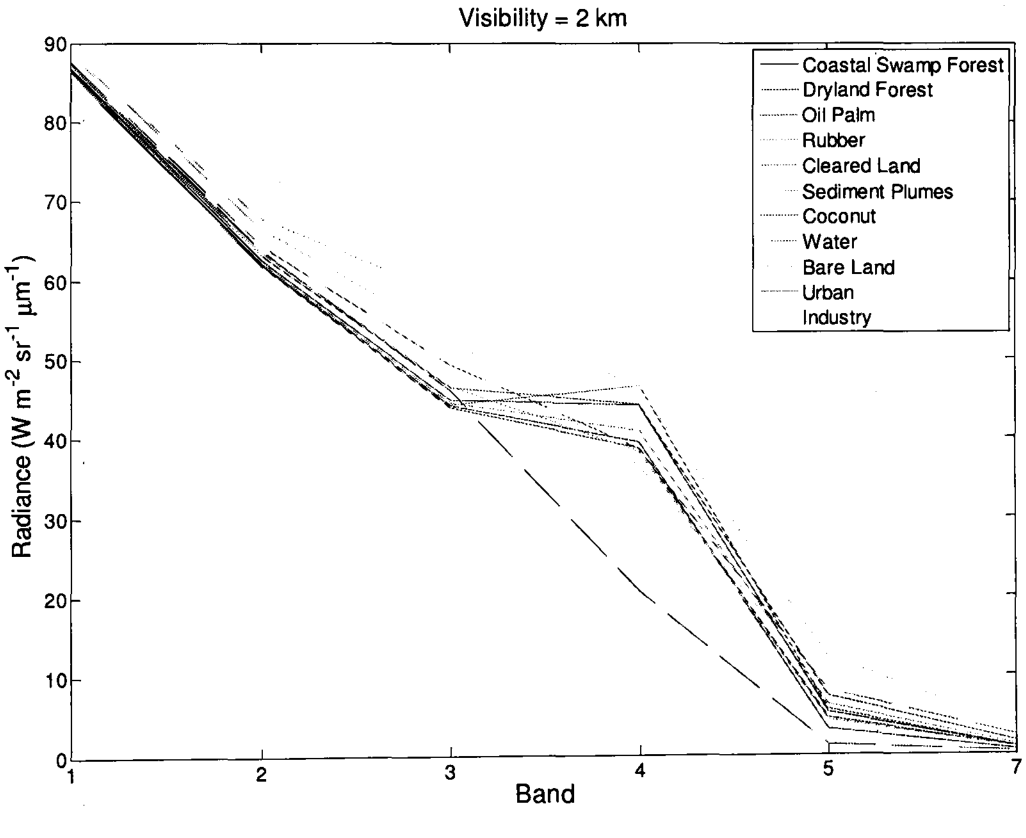
(c)


Visibility = 6 km
Radiance (W m^{-2} sr^{-1} µm^{-1})
Band
0
10
20
30
40
50
60
70
1
2
3
4
5
7
Coastal Swamp Forest
Dryland Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Coconut
Water
Bare Land
Urban
Industry

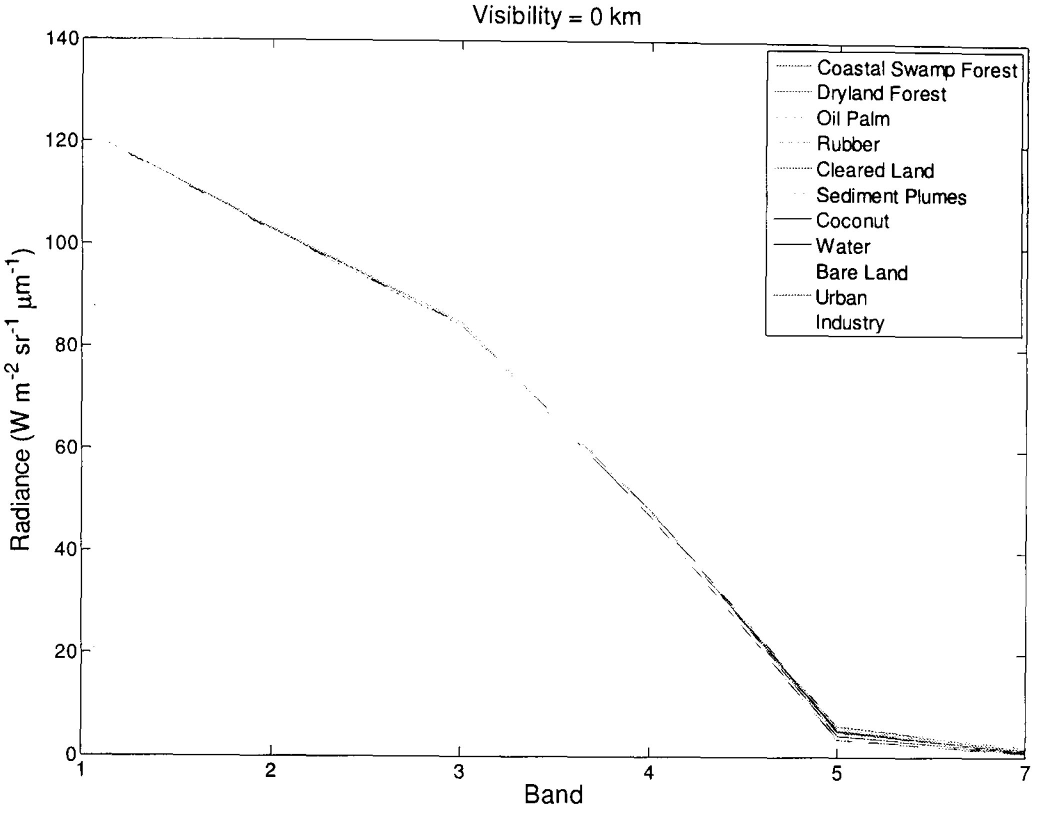
(d)


Visibility = 4 km
Radiance (W m^{-2} sr^{-1} µm^{-1})
Band
70
60
50
40
30
20
10
0
1
2
3
4
5
7
Coastal Swamp Forest
Dryland Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Coconut
Water
Bare Land
Urban
Industry

(e)


Visibility = 2 km
Radiance (W m^{-2} sr^{-1} µm^{-1})
Band
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
7
Coastal Swamp Forest
Dryland Forest
Oil Palm
Rubber
Cleared Land
Sediment Plumes
Coconut
Water
Bare Land
Urban
Industry

(f)



(g)

Figure 4.20: *Mean spectral signatures of the 12 classes at visibilities (a) 20, (b) 16, (c) 12, (d) 6, (e) 4, (f) 2 km and (g) 0 km.*

For each class, correlations between different band pairs were computed for visibilities running from 20 km to 0 km from the simulated hazy datasets by using ENVI and then checked using Equation 4.29 with MATLAB; both show a very good agreement. The correlation between band k and band l of a simulated radiance scene with a particular visibility V can be expressed by the following equation:

$$\rho_S(k,l) = \frac{C\left[L(V,k,l)\right]}{\sqrt{\text{Var}\left[L(V,k)\right]\,\text{Var}\left[L(V,l)\right]}} \qquad \text{... (4.29)}$$

where

$C\left[L(V,k,l)\right]=\left(1-\beta_k^{(1)}(V)\right)\left(1-\beta_l^{(1)}(V)\right)C\left[T(k,l)\right]+\beta_k^{(2)}(V)\beta_l^{(2)}(V)\,C\left[H(k,l)\right]$ and

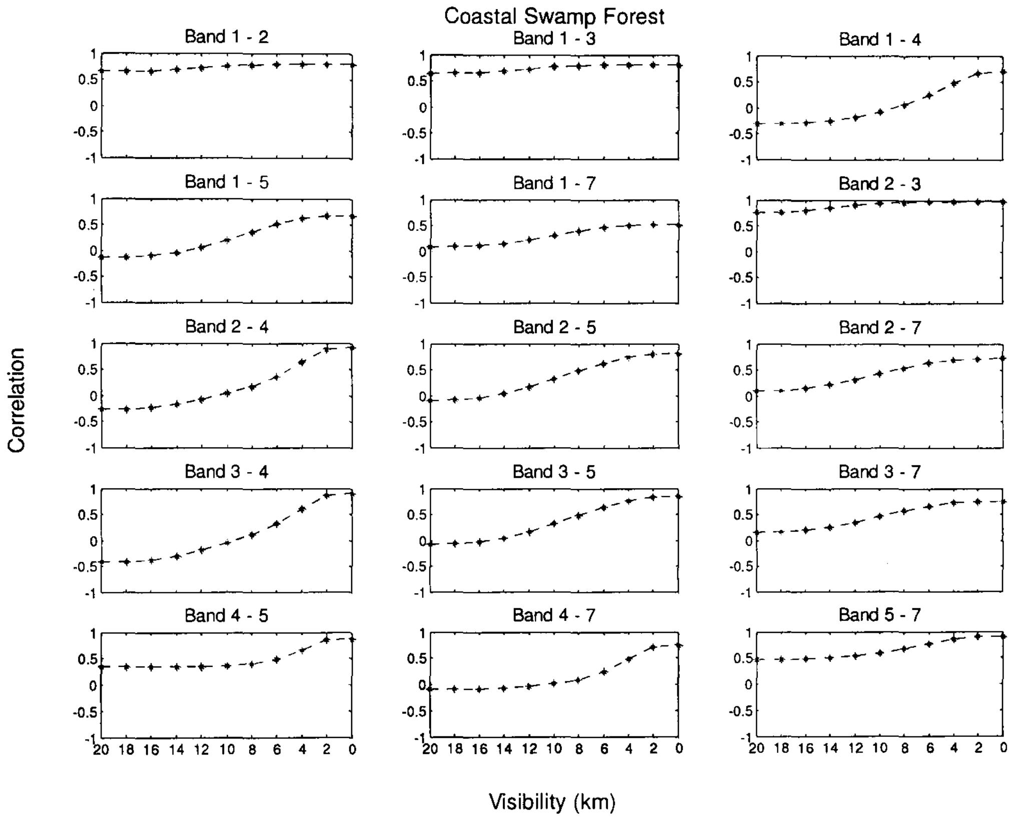
$\mathrm{Var}\left[L(V,k)\right]=\left(1-\beta_k^{(1)}(V)\right)^2\mathrm{Var}\left[T(k)\right]+\beta_k^{(2)2}(V)\,\mathrm{Var}\left[H(k)\right]$,

$C\left[T(k,l)\right]$, $\mathrm{Var}\left[T(k)\right]$ and $\mathrm{Var}\left[T(l)\right]$ are measured from the clear dataset while $C\left[H(k,l)\right]$, $\mathrm{Var}\left[H(k)\right]$ and $\mathrm{Var}\left[H(l)\right]$ are measured from the pure haze dataset. Here, we assume $\beta_k^{(1)}(V)$ and $\beta_k^{(2)}(V)$ are constant throughout the image.

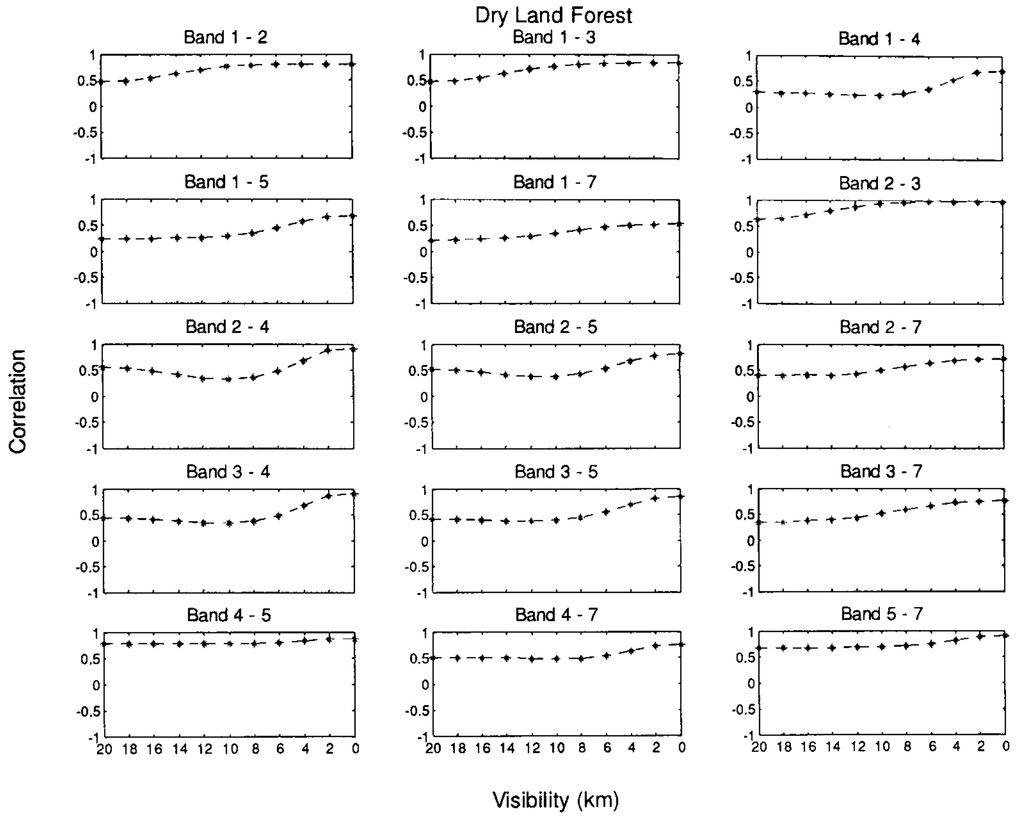
Plots of correlation against visibility for coastal swamp forest, dryland forest, oil palm, urban, bare land and water are shown in Figure 4.21(a-f). Correlations at 20 km visibility represent the classes' original correlation during clear sky condition (i.e. no haze); the correlation of pairs 1-2, 1-3 and 2-3 is higher than other pairs due to their adjacent wavelengths. On the other hand correlations at 0 km visibility represent those of pure haze; e.g. pairs 2-3, 2-4, 3-4, and 5-7 have the highest correlations at 0 km visibility, while pairs 1-5 and 1-7 the lowest (see Table 4.7).

For coastal swamp forest, i.e. a very dark class, the correlation in most pairs starts to increase steadily at longer visibilities (i.e. 18 to 12 km), gets rapid at moderate visibilities (12 to 4 km) but steady again at shorter visibilities (i.e. less than 4 km). This shows the haze significantly modifies the correlations at shorter compared to longer visibilities, with a rapid increase in modification occurs at moderate visibilities. In such case, as haze becomes more severe, $C\left[L(V,k,l)\right]$ in Equation (4.29) gets bigger and so does $\rho_S(k,l)$. However, such trend is not so obvious for dryland forest and water because they already posseses quite high correlations at longer visibilities due to the original spectral properties of the classes. For oil palm, the rapid modification occurs at quite short visibilities (i.e. 6 to 2 km) due to the less dark properties of the class; i.e. its spectral properties are influenced by the ground reflectance from the spaces between the oil palm trees.

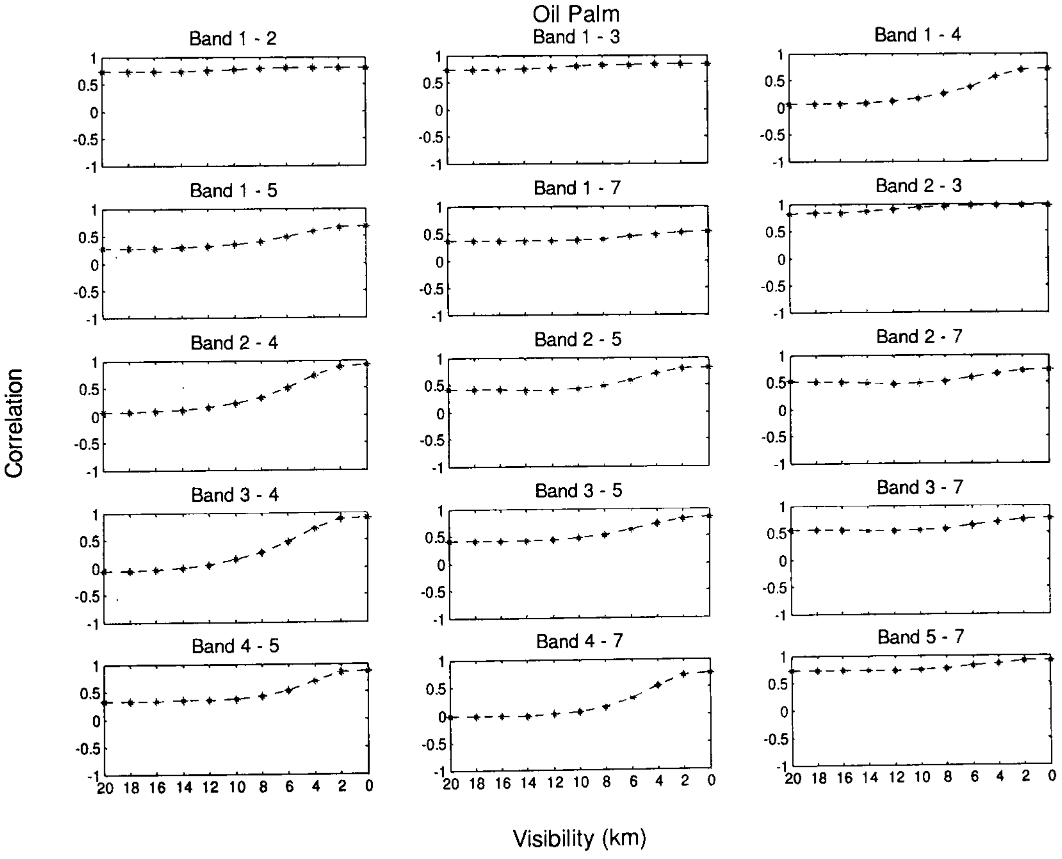
For bare land, i.e. a very bright class, the correlation is quite high and constant from 20 to 6 km visibility, but a rapid change in correlation occurs at visibilities less than 6 km. This shows that compared to dark classes, for bright classes, most modification by haze occur at very short visibilities, signifying a much severe haze is required to modify the correlation of bright classes. For urban, a rapid increase in correlation occur at slightly longer visibilities compared to bare land, signifying the stronger effects of haze due to the less bright properties of the class.



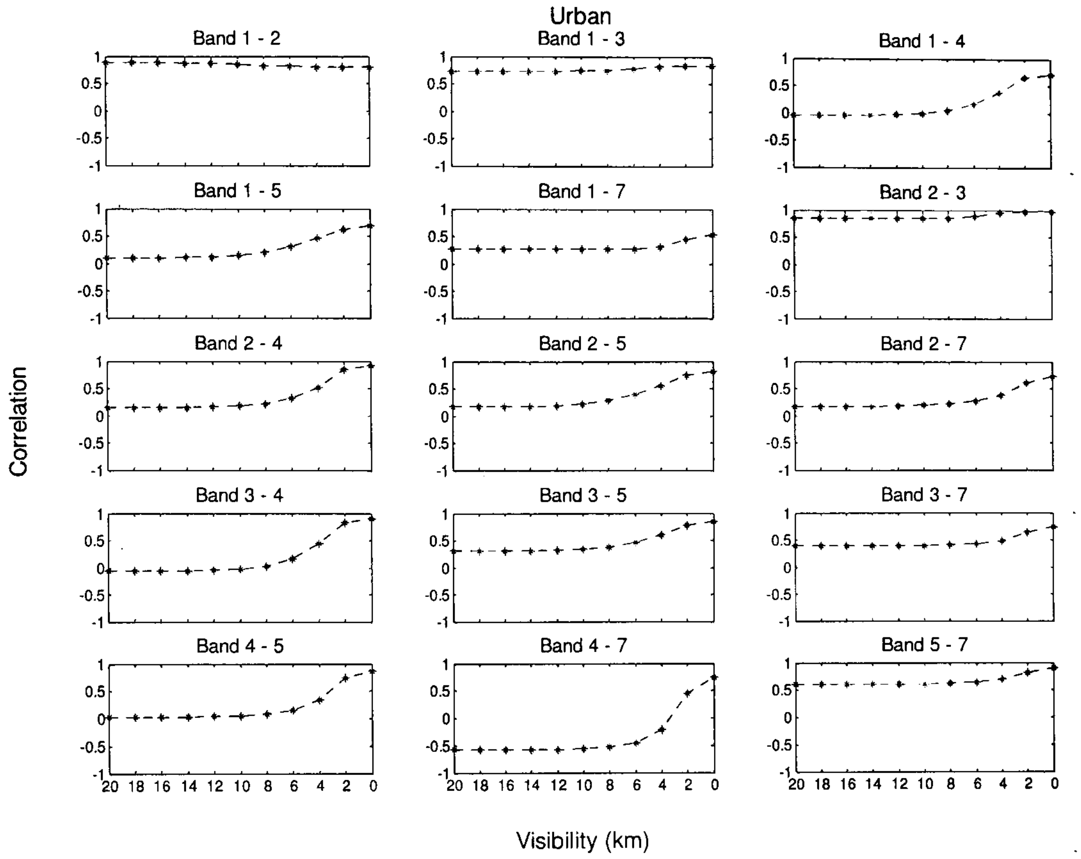
(a)


Dry Land Forest
Band 1 - 2
Band 1 - 3
Band 1 - 4
Band 1 - 5
Band 1 - 7
Band 2 - 3
Band 2 - 4
Band 2 - 5
Band 2 - 7
Band 3 - 4
Band 3 - 5
Band 3 - 7
Band 4 - 5
Band 4 - 7
Band 5 - 7
Correlation
1
0.5
0
-0.5
-1
20 18 16 14 12 10 8 6 4 2 0
Visibility (km)

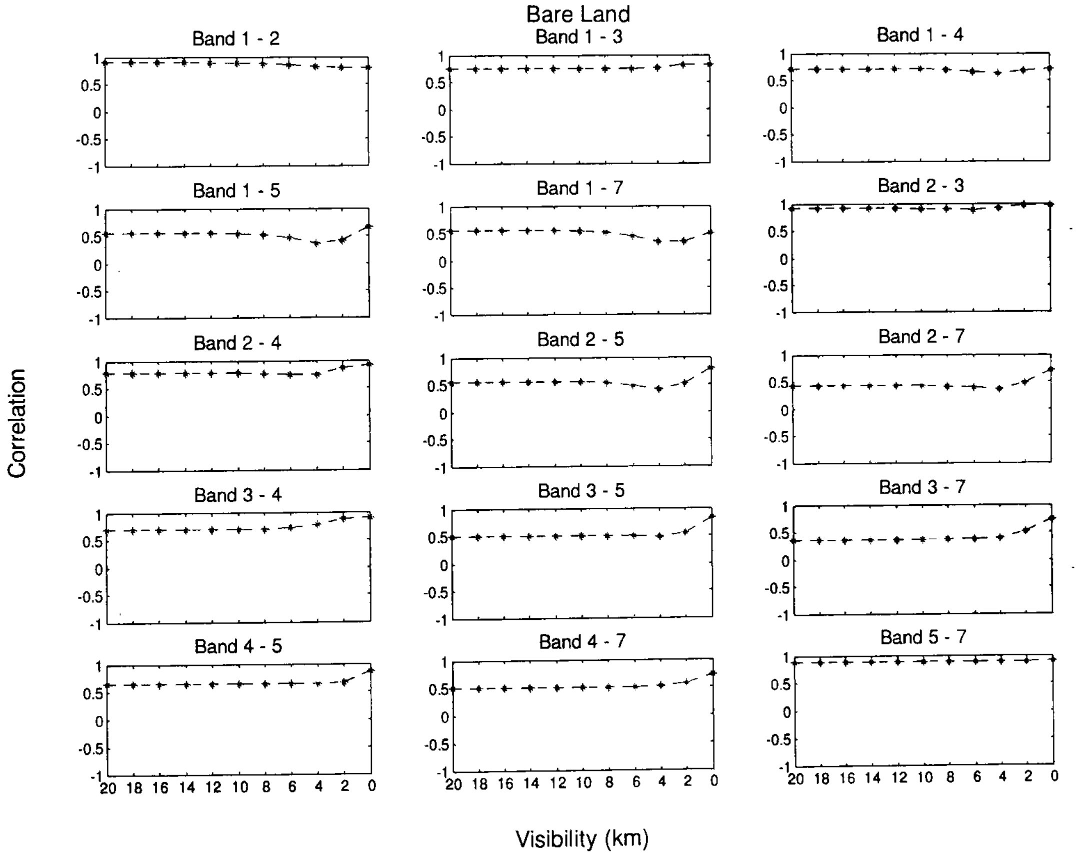
(b)


Oil Palm
Band 1 - 2
Band 1 - 3
Band 1 - 4
Band 1 - 5
Band 1 - 7
Band 2 - 3
Band 2 - 4
Band 2 - 5
Band 2 - 7
Band 3 - 4
Band 3 - 5
Band 3 - 7
Band 4 - 5
Band 4 - 7
Band 5 - 7
Correlation
1
0.5
0
-0.5
-1
20 18 16 14 12 10 8 6 4 2 0
Visibility (km)

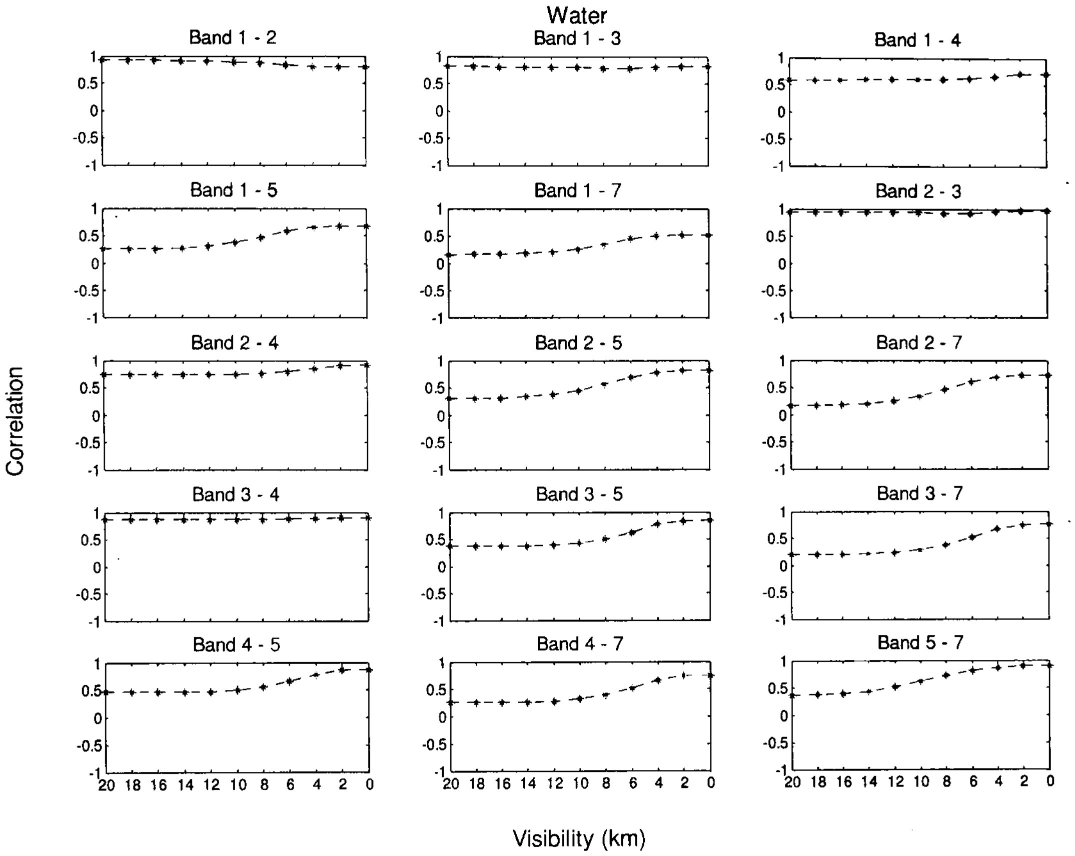
(c)


Urban
Band 1 - 2
Band 1 - 3
Band 1 - 4
Band 1 - 5
Band 1 - 7
Band 2 - 3
Band 2 - 4
Band 2 - 5
Band 2 - 7
Band 3 - 4
Band 3 - 5
Band 3 - 7
Band 4 - 5
Band 4 - 7
Band 5 - 7
Correlation
1
0.5
0
-0.5
-1
20 18 16 14 12 10 8 6 4 2 0
Visibility (km)

(d)


Bare Land
Band 1 - 2
Band 1 - 3
Band 1 - 4
Band 1 - 5
Band 1 - 7
Band 2 - 3
Band 2 - 4
Band 2 - 5
Band 2 - 7
Band 3 - 4
Band 3 - 5
Band 3 - 7
Band 4 - 5
Band 4 - 7
Band 5 - 7
Correlation
1
0.5
0
-0.5
-1
20 18 16 14 12 10 8 6 4 2 0
Visibility (km)

(e)

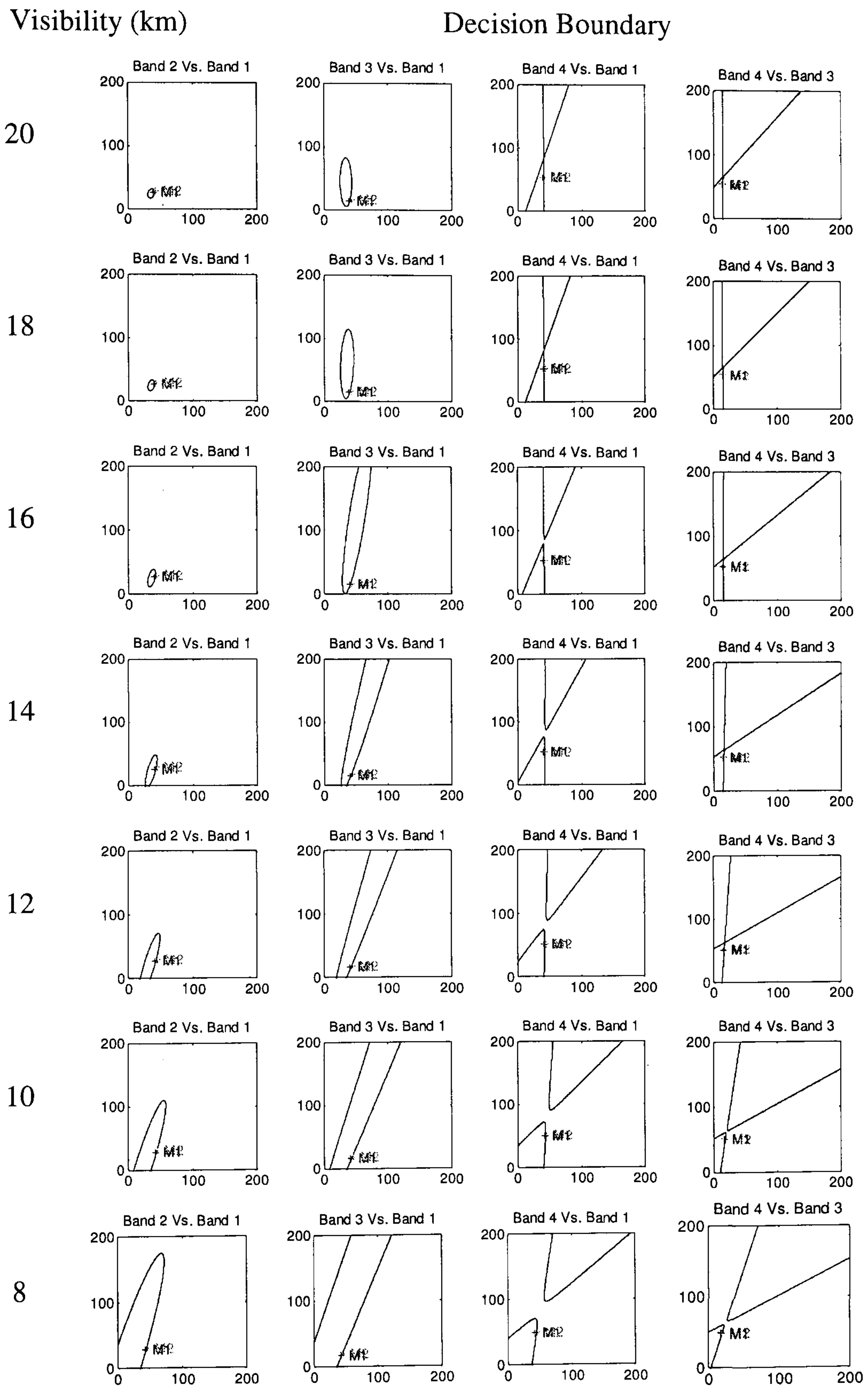


(f)

Figure 4.21: *Correlation between bands with reducing visibility for (a) coastal swamp forest and (b) dryland forest, (c) oil palm, (d) urban, (e) bare land and (f) water.*

Figure 4.22 shows decision boundaries (see Equation 3.14), of dryland forest and coastal swamp forest, with means indicated by 'M1' and 'M2' respectively, for band pairs 2:1, 3:1, 4:1 and 4:3 from visibilities 20 to 0 km. Although having very similar spectral properties, we showed in Chapter 3 that ML able to classify these classes with high accuracies. Here, we intend to analyse the decision boundary of the classes as haze gets severe. The pairs are arranged to form combinations of measurements between adjacent (i.e. bands 2 (green) and 1 (blue)) and non-adjacent (bands 3 (red) and 1) of visible bands as well as between near infrared (band 4) and visible bands (bands 1 and 3), to see the haze effects as visibility reduces. For pair 2:1, the decision boundary begins with a small elliptic curve at 20 km visibility which grows bigger through to 14 km visibility, changes to parabolic curves at 12 km which increase in size through to 2 km visibility and finally change to a hyperbolic curve at 0 km visibility. Pair 3:1 begins with a bigger elliptic curve; it experiences a similar but more rapid change in size compared to pair 2:1. Pairs 4:1 and 4:3 begin with hyperbolic curves; the separation between the curves gets bigger as visibility decrease and finally the curves change to parabolic shape at 0 km visibility. The distance between the means gets smaller as visibility declines and eventually the means overlap at 0 km visibility.

For pairs 2:1 and 3:1, the inner elliptic curves represent dryland forest due to having a smaller variance in bands 1, 2 and 3. The growing of the elliptic curves from 20 to 14 km visibilities (i.e. haze is not severe) is due to the increase of the dryland forest covariance of the pairs. The increase of the major axis of the elliptic curve is due to the increase in correlation between the bands (see Figure 4.21(b)). As haze gets more severe, the shape of the curve changes (i.e. from elliptic to parabolic and then to hyperbolic curve), indicating that haze has a significant effects on the decision boundary which involves these bands. The shifting of M1 and M2 is associated with the increase of the forest means in these bands as haze gets severe. For pairs 4:1 and 4:3, from 20 to 14 km visibilities, the change in shape and size is not that apparent because haze has less effects in band 4 which has longer wavelengths. These changes get clearer from 8 to 0 km visibilities (i.e. short visibilities), indicating that the effects of haze only become apparent when haze becomes severe. The analysis show that the changes to the decision boundary involving bands 1, 2 and 3, due to haze, are easier compared to when band 4 is involved; this suggests that the effects of haze on bands with shorter wavelengths are more significant than for longer wavelengths.


Visibility (km)
Decision Boundary
20
Band 2 Vs. Band 1
Band 3 Vs. Band 1
Band 4 Vs. Band 1
Band 4 Vs. Band 3
18
Band 2 Vs. Band 1
Band 3 Vs. Band 1
Band 4 Vs. Band 1
Band 4 Vs. Band 3
16
Band 2 Vs. Band 1
Band 3 Vs. Band 1
Band 4 Vs. Band 1
Band 4 Vs. Band 3
14
Band 2 Vs. Band 1
Band 3 Vs. Band 1
Band 4 Vs. Band 1
Band 4 Vs. Band 3
12
Band 2 Vs. Band 1
Band 3 Vs. Band 1
Band 4 Vs. Band 1
Band 4 Vs. Band 3
10
Band 2 Vs. Band 1
Band 3 Vs. Band 1
Band 4 Vs. Band 1
Band 4 Vs. Band 3
8
Band 2 Vs. Band 1
Band 3 Vs. Band 1
Band 4 Vs. Band 1
Band 4 Vs. Band 3
M1
M2
0
100
200

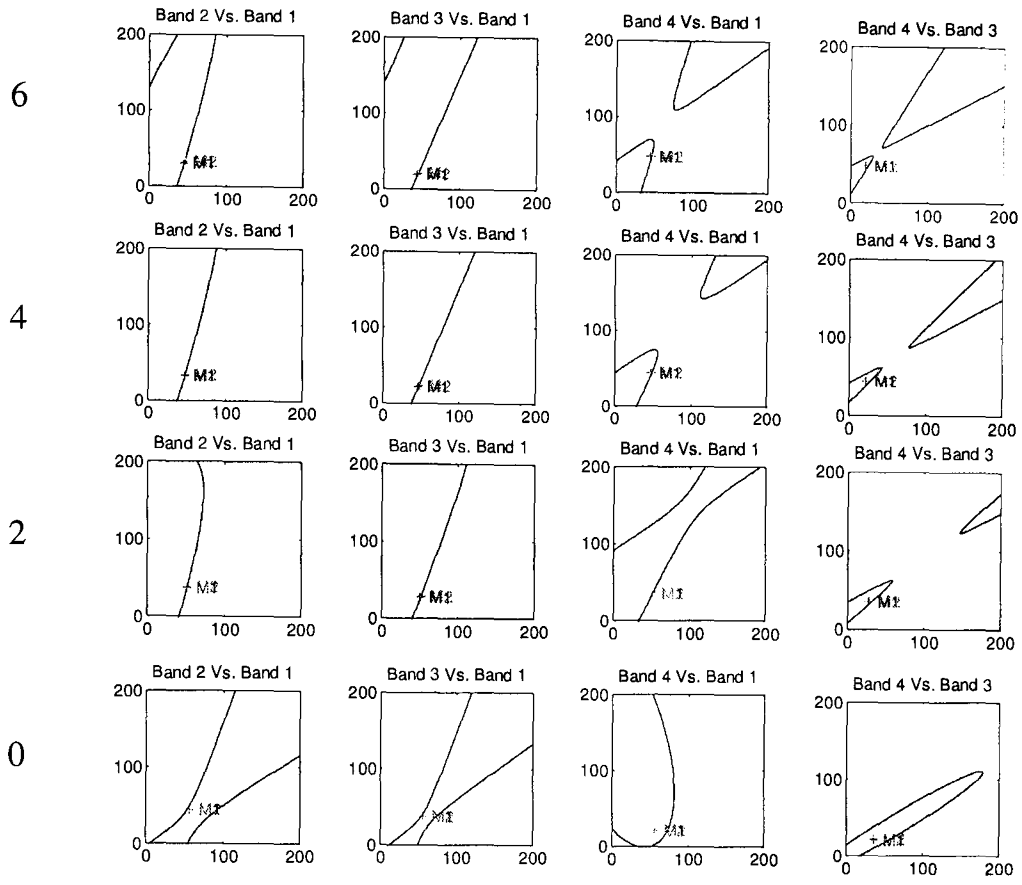


Figure 4.22: *Decision boundaries of coastal swamp forest and dryland forest for band pairs 2:1, 3:1, 4:1 and 4:3 from visibilities 20 to 0 km. 'M1' and 'M2' are the means for dryland forest and coastal swamp forest respectively.*

Accuracy Assessment and Accuracy Analysis of ML Classification

Haze modifies the means and band correlations of a class, but these govern ML classification (Equation 3.1). In this section we therefore investigate how haze affects the classification accuracy. The assessment is carried out using the confusion matrix (see Section 3.4). Figure 4.24 shows producer accuracy plots for all 11 cover types. All classes show a decrease in classification accuracy as visibility reduces. Less reflective classes, such as forest, oil palm, rubber and water, experience a gradual decline at longer visibilities but then a more rapid decline at shorter visibilities. Haze starts to severely affect these classes at visibilities less than 4 km. Cleared land and sediment plumes exhibit a nearly linear decline.

Some classes, i.e. rubber, water, coconut, bare land, urban and industry, exhibit a non-zero accuracy at 0 km visibility; this is because some pixels are still correctly classified to these classes because not severely influenced by very thick haze compared to other classes. For industry, an unexpected increasing trend is observed from 2 km to 0 km visibility. This is primarily because of similarity between the statistics (i.e. mean and covariance structure) of haze and industry. Figure 4.23 shows the conditions of the industry pixels (grey) for 20 km, 2 km and 0 km visibility. At 2 km visibility (Figure 4.23(b)), a large portion of industry pixels are misclassified as urban (red), but at 0 km visibility (Figure 4.23(c)), some of them are again correctly classified as industry (shown as scattered grey pixels), thus causing an increase in producer accuracy. This is because the hazy condition at 0 km visibility tends to increase the number of industry pixels that are correctly classified.

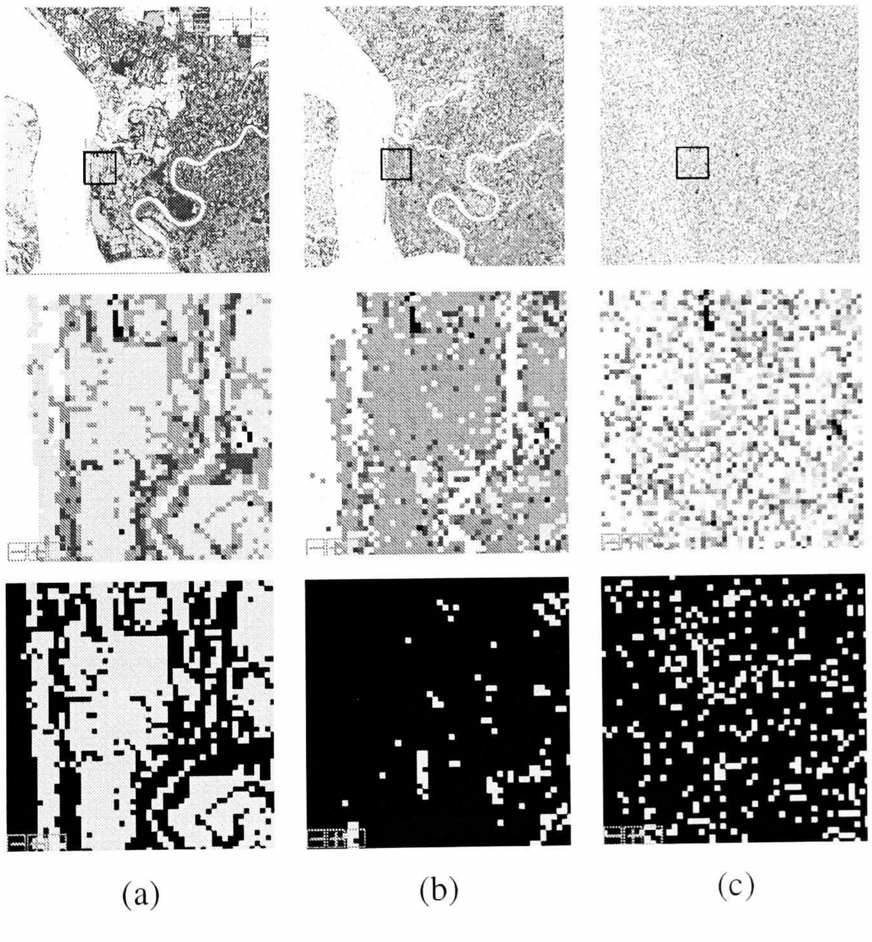


Figure 4.23: *A portion of ML classification for (a) 20 km, (b) 2 km and (c) 0 km visibility datasets (top), the corresponding enlarged versions (second row) and enlarged versions with non-industry pixels masked in black (c).*

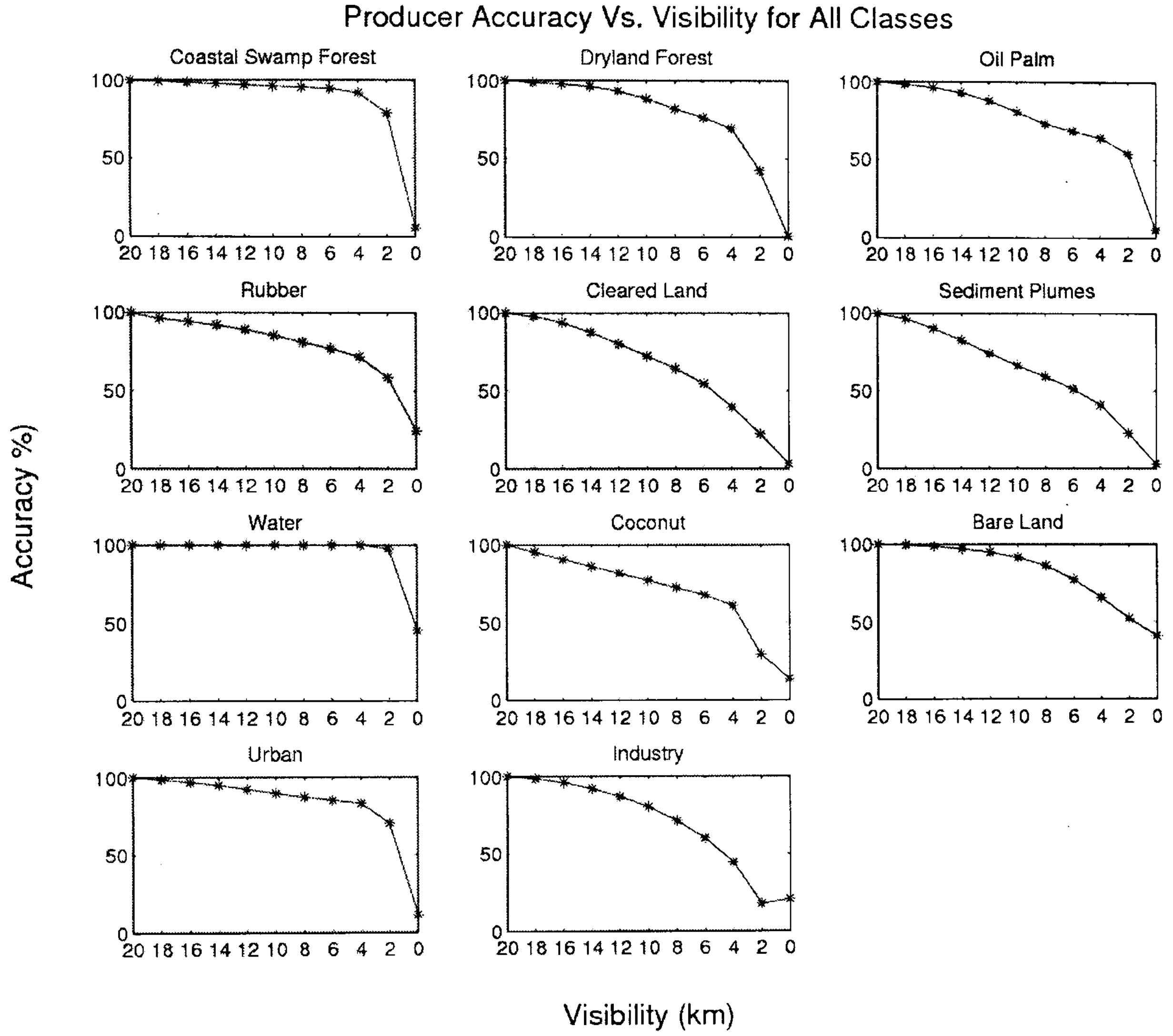
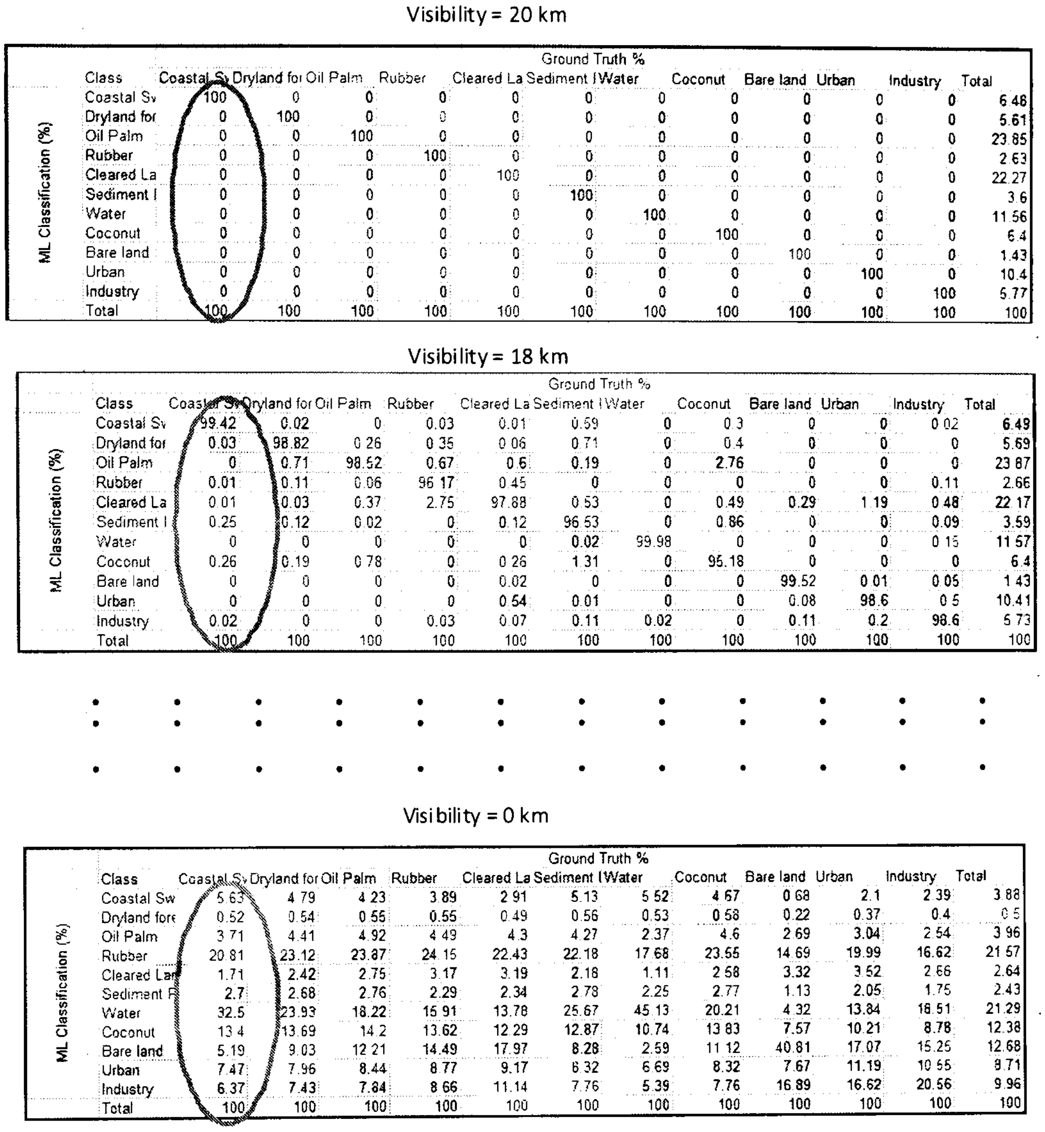


Figure 4.24: *Producer accuracy for each class with reducing visibility.*

Manual comparison by simultaneously displaying the different visibility confusion matrices is not possible. A more convenient way is by plotting the elements from a particular column of the confusion matrix for each visibility (Figure 4.25). By doing so, the distribution of ground truth pixels assigned to the different classes as visibility changes can be analysed. Figure 4.26 shows the percentage of pixels for (a) coastal swamp forest, (b) dryland forest, (c) oil palm, (d) rubber, (e) cleared land, (f) sediment plumes, (g) water, (h) coconut, (i) bare land, (j) urban and (k) industry, against ground truth classes. For each plot, 100% represents all the pixels from a given ground truth class.



Visibility = 20 km

ML Classification (%)					Ground Truth %							
Class	Coastal Sv	Dryland for	Oil Palm	Rubber	Cleared La	Sediment I	Water	Coconut	Bare land	Urban	Industry	Total
Coastal Sv	100	0	0	0	0	0	0	0	0	0	0	6.48
Dryland for	0	100	0	0	0	0	0	0	0	0	0	5.61
Oil Palm	0	0	100	0	0	0	0	0	0	0	0	23.85
Rubber	0	0	0	100	0	0	0	0	0	0	0	2.63
Cleared La	0	0	0	0	100	0	0	0	0	0	0	22.27
Sediment I	0	0	0	0	0	100	0	0	0	0	0	3.6
Water	0	0	0	0	0	0	100	0	0	0	0	11.56
Coconut	0	0	0	0	0	0	0	100	0	0	0	6.4
Bare land	0	0	0	0	0	0	0	0	100	0	0	1.43
Urban	0	0	0	0	0	0	0	0	0	100	0	10.4
Industry	0	0	0	0	0	0	0	0	0	0	100	5.77
Total	100	100	100	100	100	100	100	100	100	100	100	100

Visibility = 18 km

ML Classification (%)					Ground Truth %							
Class	Coastal Sv	Dryland for	Oil Palm	Rubber	Cleared La	Sediment I	Water	Coconut	Bare land	Urban	Industry	Total
Coastal Sv	99.42	0.02	0	0.03	0.01	0.59	0	0.3	0	0	0.02	6.49
Dryland for	0.03	98.82	0.26	0.35	0.06	0.71	0	0.4	0	0	0	5.69
Oil Palm	0	0.71	98.52	0.67	0.6	0.19	0	2.76	0	0	0	23.87
Rubber	0.01	0.11	0.06	96.17	0.45	0	0	0	0	0	0.11	2.66
Cleared La	0.01	0.03	0.37	2.75	97.88	0.53	0	0.49	0.29	1.19	0.48	22.17
Sediment I	0.25	0.12	0.02	0	0.12	96.53	0	0.86	0	0	0.09	3.59
Water	0	0	0	0	0	0.02	99.98	0	0	0	0.15	11.57
Coconut	0.26	0.19	0.78	0	0.26	1.31	0	95.18	0	0	0	6.4
Bare land	0	0	0	0	0.02	0	0	0	99.52	0.01	0.05	1.43
Urban	0	0	0	0	0.54	0.01	0	0	0.08	98.6	0.5	10.41
Industry	0.02	0	0	0.03	0.07	0.11	0.02	0	0.11	0.2	98.6	5.73
Total	100	100	100	100	100	100	100	100	100	100	100	100

Visibility = 0 km

ML Classification (%)					Ground Truth %							
Class	Coastal Sv	Dryland for	Oil Palm	Rubber	Cleared La	Sediment I	Water	Coconut	Bare land	Urban	Industry	Total
Coastal Sw	5.63	4.79	4.23	3.89	2.91	5.13	5.52	4.67	0.68	2.1	2.39	3.88
Dryland fore	0.52	0.54	0.55	0.55	0.49	0.56	0.53	0.58	0.22	0.37	0.4	0.5
Oil Palm	3.71	4.41	4.92	4.49	4.3	4.27	2.37	4.6	2.69	3.04	2.54	3.96
Rubber	20.81	23.12	23.87	24.15	22.43	22.18	17.68	23.55	14.69	19.99	16.62	21.57
Cleared Lar	1.71	2.42	2.75	3.17	3.19	2.18	1.11	2.58	3.32	3.52	2.66	2.64
Sediment P	2.7	2.68	2.76	2.29	2.34	2.78	2.25	2.77	1.13	2.05	1.75	2.43
Water	32.5	23.93	18.22	15.91	13.78	25.67	45.13	20.21	4.32	13.84	18.51	21.29
Coconut	13.4	13.69	14.2	13.62	12.29	12.87	10.74	13.83	7.57	10.21	8.78	12.38
Bare land	5.19	9.03	12.21	14.49	17.97	8.28	2.59	11.12	40.81	17.07	15.25	12.68
Urban	7.47	7.96	8.44	8.77	9.17	6.32	6.69	8.32	7.67	11.19	10.55	8.71
Industry	6.37	7.43	7.84	8.66	11.14	7.76	5.39	7.76	16.89	16.62	20.56	9.96
Total	100	100	100	100	100	100	100	100	100	100	100	100

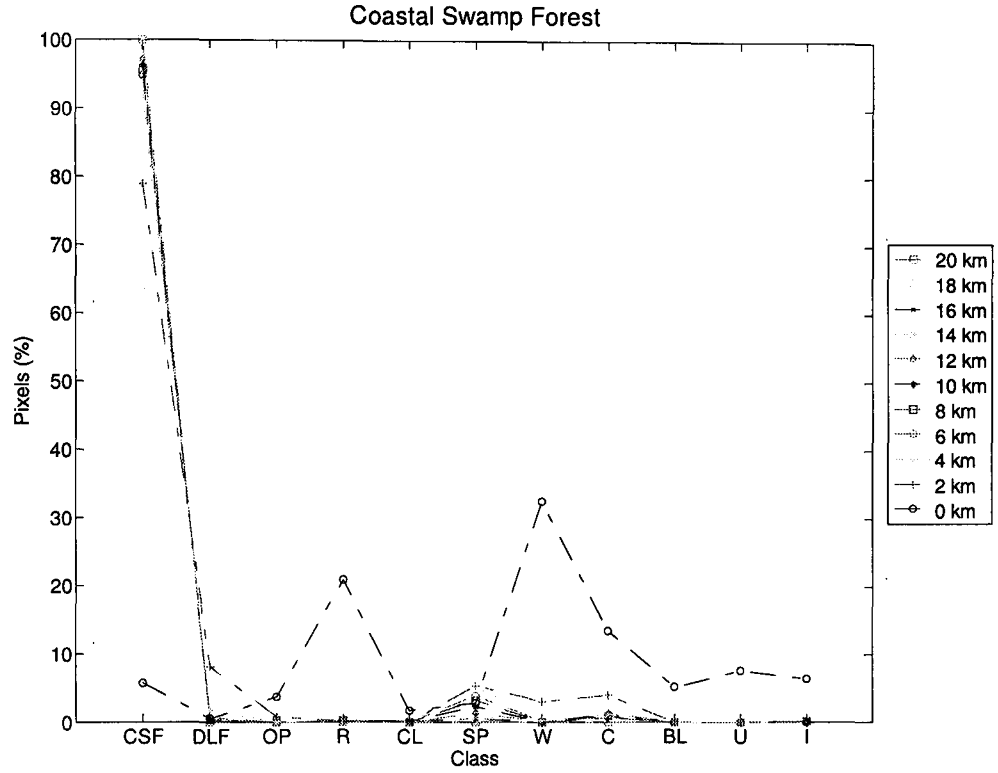
Figure 4.25: *Extraction of the element from a column of the confusion matrices.*

In Figure 4.26(a), the highest points (referring to the percentages of correctly classified coastal swamp forest pixels at different visibilities) concentrate between 90% and 100%, for 20 km to 4 km-visibility curves, indicating that most coastal swamp forest pixels are correctly classified at good to quite poor visibilities (see Figure 4.18). A similar case is observed for water (Figure 4.26(g)). Hence, haze has little effect on these classes even when it is quite severe. For other classes (i.e. dryland forest, oil palm, rubber, coconut, bare land and urban) that are more affected by the haze, the peaks are less concentrated (Figure 4.26(b), (c), (d), (h), (i) and (j)). The classes most affected are cleared land, sediment plumes and industry (Figure 4.26(e), (f) and (k)), in which the peak is only about 40% for 4 km visibility. An upward trend in the plots represents the pixels being misclassified to other classes as the visibility reduces. This happens because, when haze exists, the pixels tend to migrate to incorrect classes, as summarised in Table 4.8. Due to the very distinct spectral properties of water, almost no migration of water pixels occurs at all visibilities except 0 km. For most classes, the pixels tend to migrate to a single class. Coastal swamp forest, water, coconut, bare land, urban and industry pixels are likely to migrate to sediment plumes, rubber, oil palm, industry, cleared land and urban classes respectively. Dryland forest, oil palm and rubber pixels tend to migrate to the coconut class. The cleared land and sediment plumes pixels tend to migrate to multiple classes, which are oil palm, rubber, coconut and urban for the former, and forests and coconut for the latter.

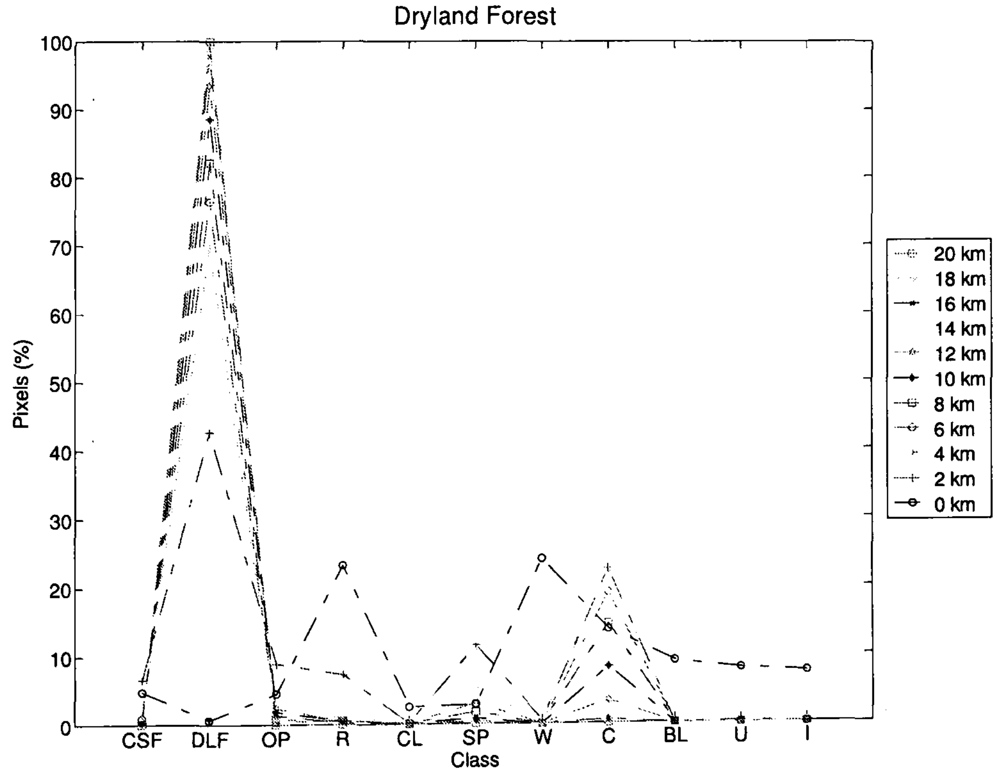
Table 4.8: *The main incorrect classes to which the pixels migrate as visibility reduces. The grey shaded boxes are not relevant for this analysis.*

Ground Truth Pixels	Incorrect ML Class which the pixels fall into										
	Coastal Swamp Forest	Dryland Forest	Oil Palm	Rubber	Cleared Land	Sediment Plumes	Water	Coconut	Bare land	Urban	Industry
Coastal Swamp Forest	▨					√					
Dryland Forest		▨						√			
Oil Palm			▨					√			
Rubber				▨				√			
Cleared Land			√	√	▨			√		√	
Sediment Plumes	√	√				▨		√			
Water				√			▨				
Coconut			√					▨			
Bare Land									▨		√
Urban					√					▨	
Industry										√	▨

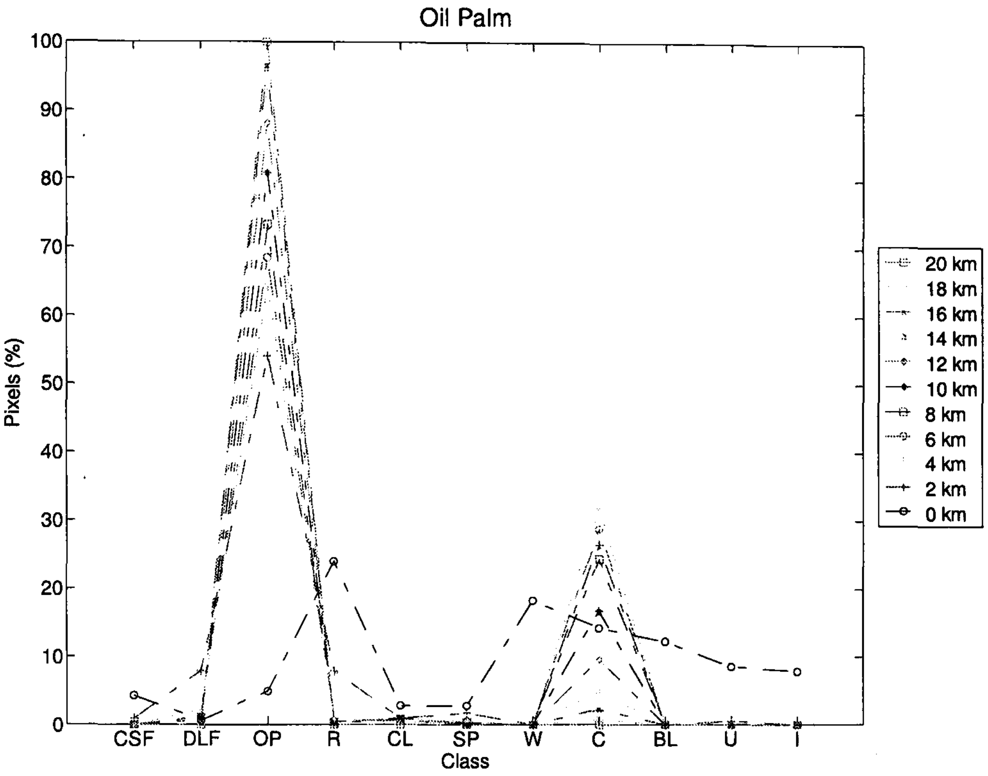
Surprisingly, from Figure 4.26(d), (g), (i) and (k), quite a large number of pixels are still classified to the correct class even under very hazy conditions (i.e. 0 km visibility). The obvious ones are rubber (20%), water (50%) and bare land. This suggests that the modification of spectral properties of these classes due to very thick haze is not as severe as other classes.



(a)

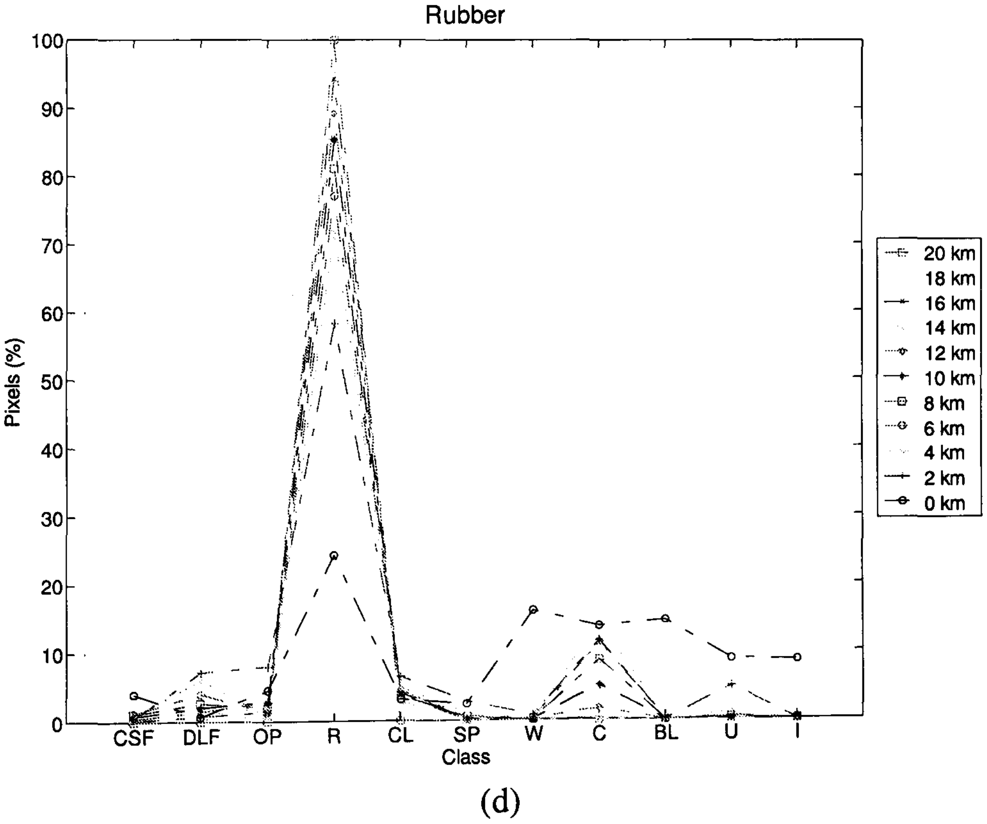


(b)



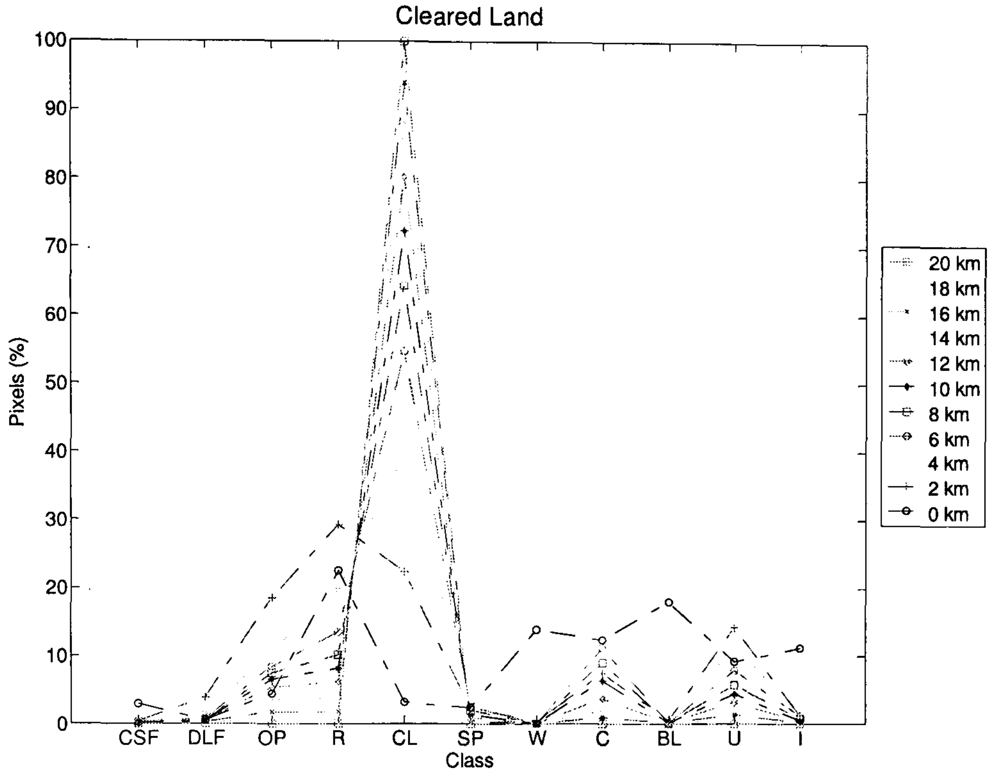
Oil Palm
Pixels (%)
Class
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
0
10
20
30
40
50
60
70
80
90
100
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

(c)

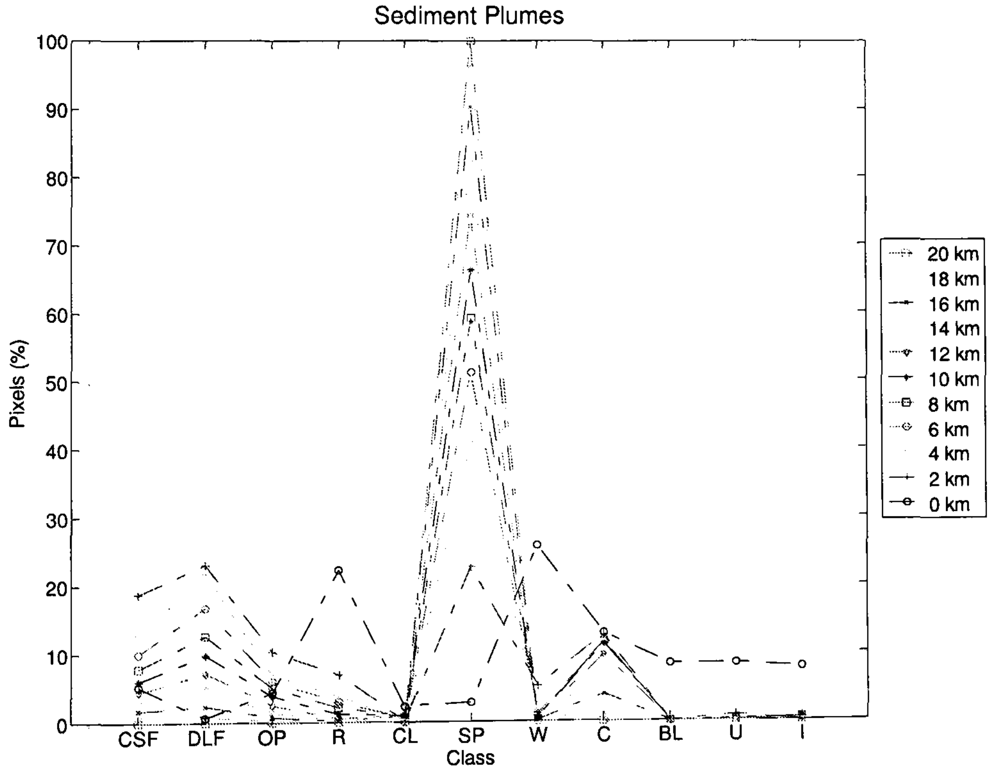


Rubber
Pixels (%)
Class
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
0
10
20
30
40
50
60
70
80
90
100
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

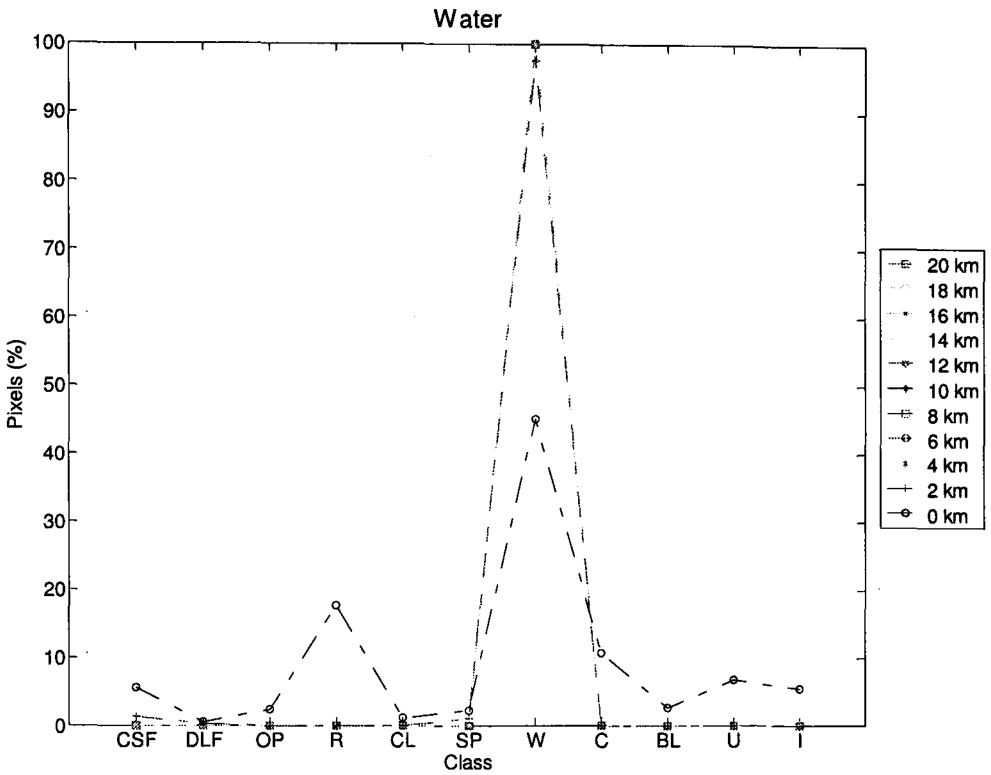
(d)


Cleared Land
Pixels (%)
100
90
80
70
60
50
40
30
20
10
0
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

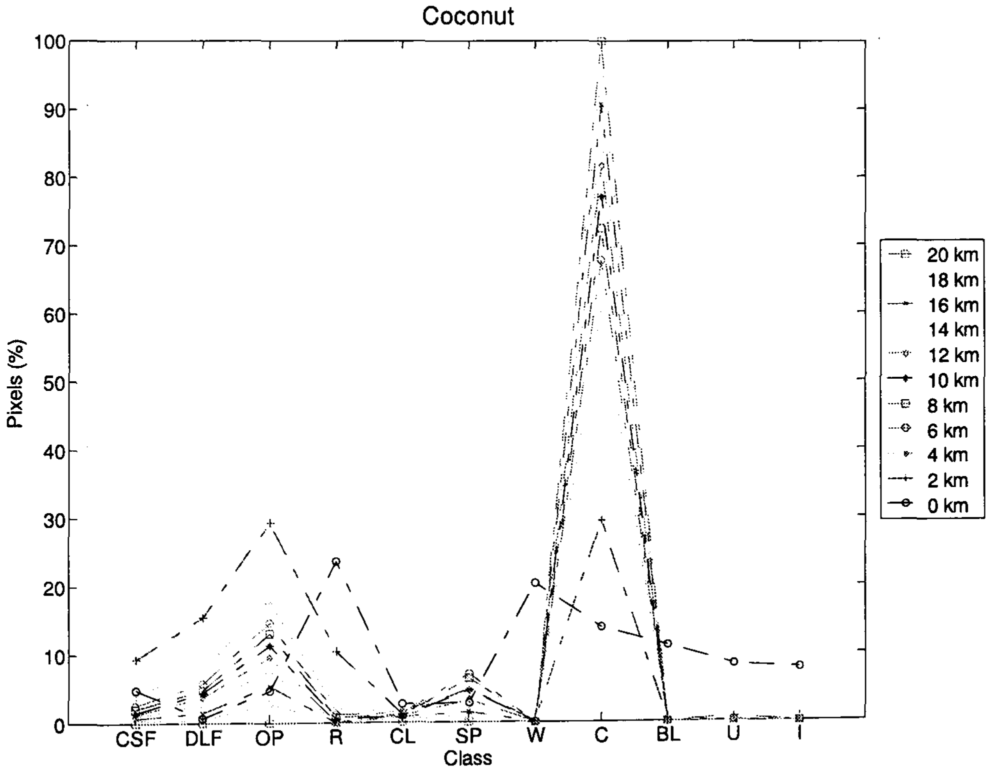
(e)


Sediment Plumes
Pixels (%)
100
90
80
70
60
50
40
30
20
10
0
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

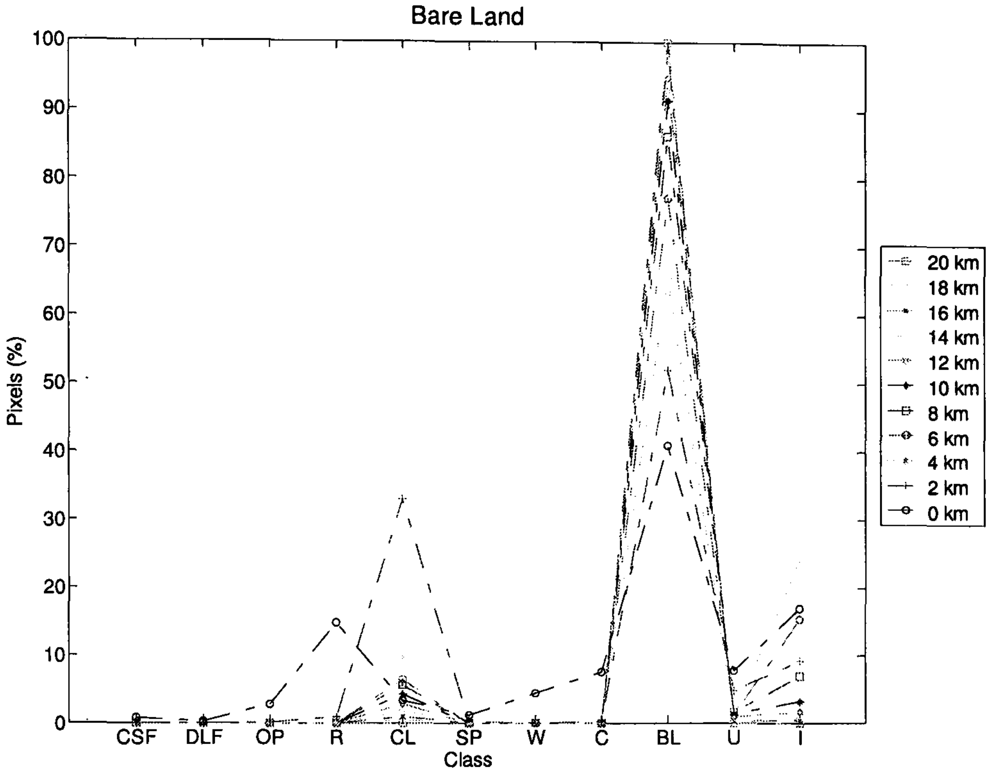
(f)


Water
100
90
80
70
60
50
40
30
20
10
0
Pixels (%)
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

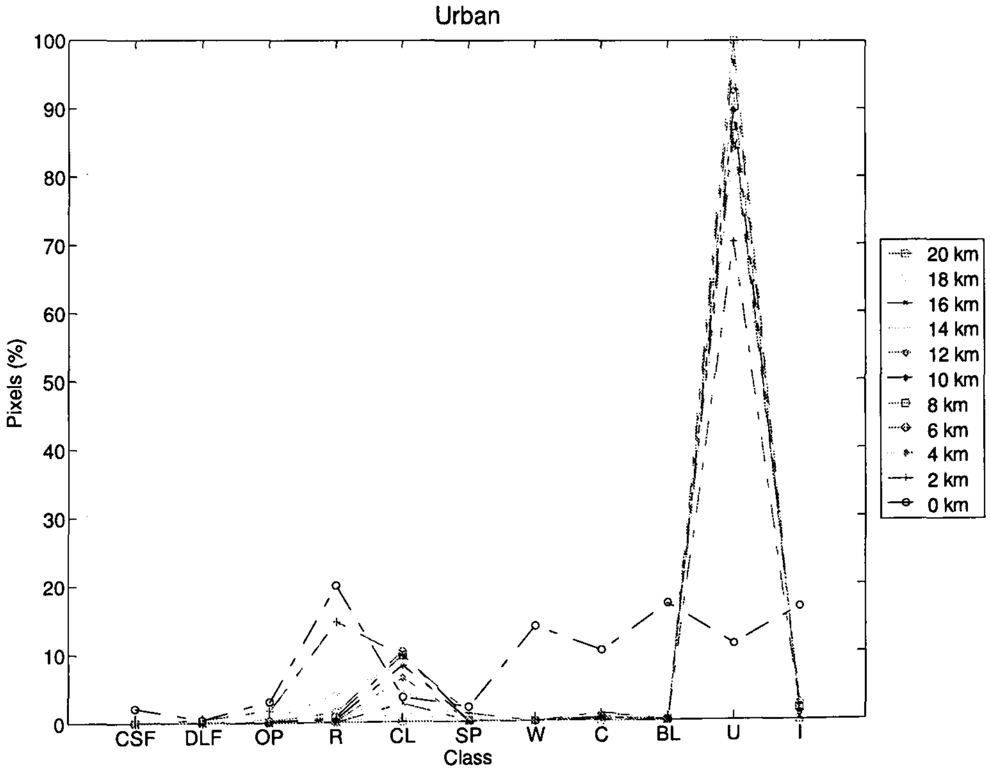
(g)


Coconut
100
90
80
70
60
50
40
30
20
10
0
Pixels (%)
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

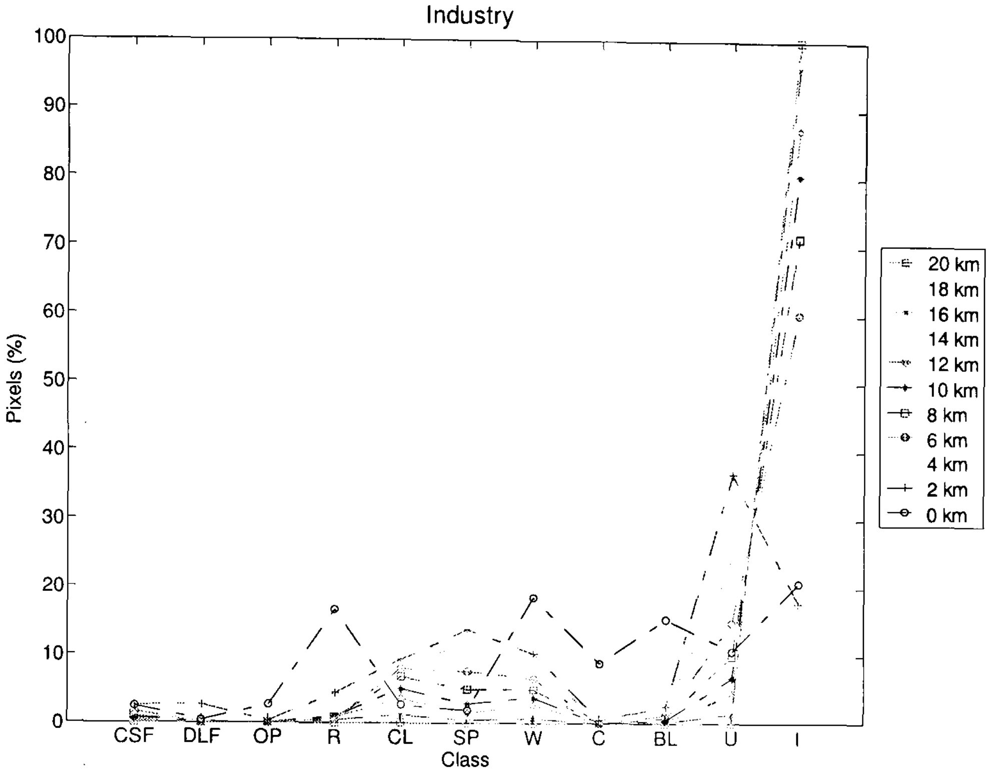
(h)


Bare Land
Pixels (%)
100
90
80
70
60
50
40
30
20
10
0
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

(i)


Urban
Pixels (%)
100
90
80
70
60
50
40
30
20
10
0
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

(j)



(k)

Figure 4.26: *Percentage of pixels for (a) coastal swamp forest, (b) dryland forest, (c) oil palm, (d) rubber, (e) cleared land, (f) sediment plumes, (g) water, (h) coconut, (i) bare land, (j) urban and (k) industry, against ground truth classes. 100% for a given class type, represents all the pixels from that class.*

Figure 4.27 shows plots of user accuracy against visibility for all classes. A nearly linear decrease occurs for rubber, sediment plumes and coconut. The remaining classes have a much slower decline for visibilities greater than 6 km, but a more rapid decline for visibilities less than 6 km.

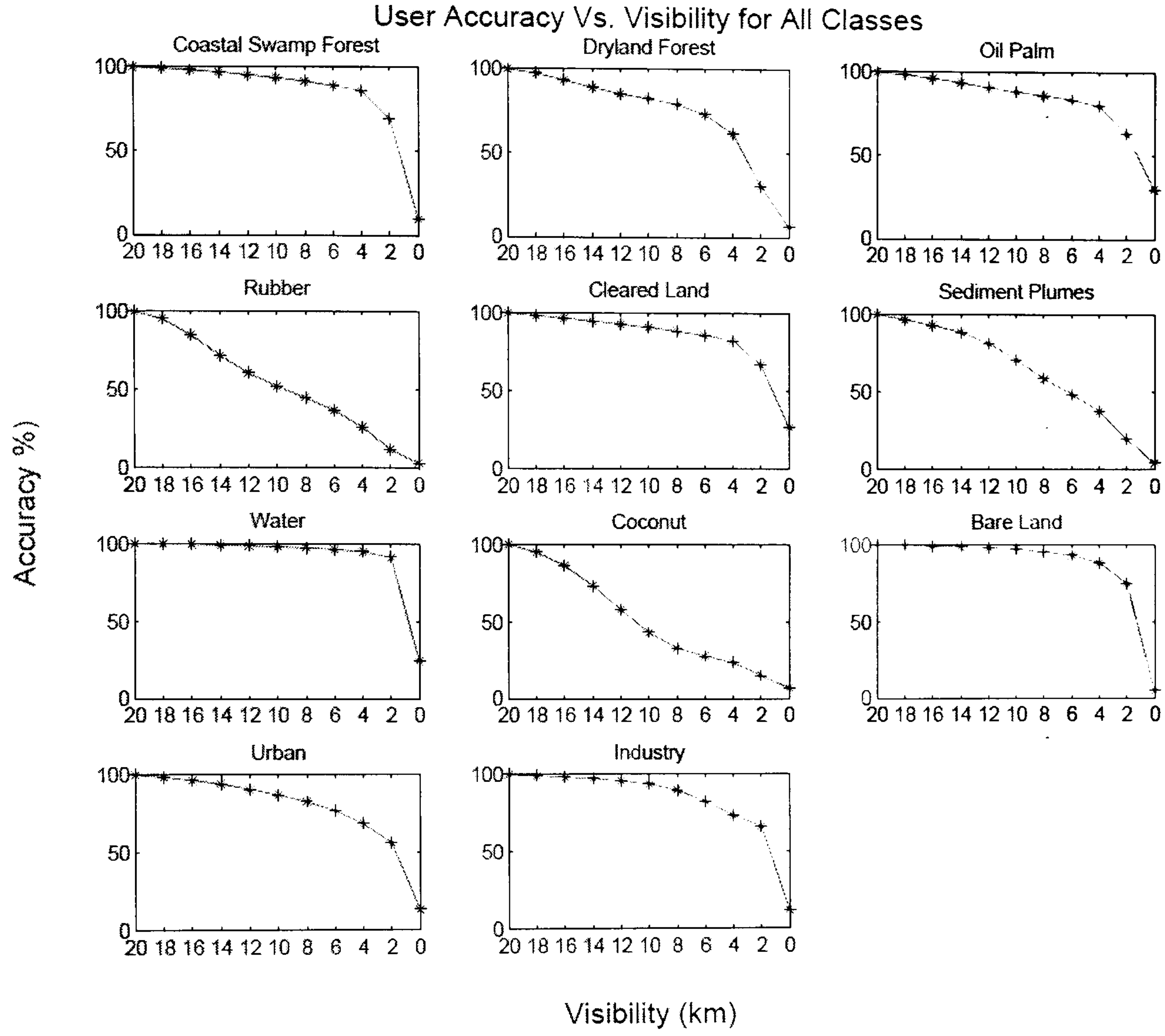


Figure 4.27: *User accuracy for each class with reducing visibility.*

Figure 4.28 shows a plot of overall classification accuracy and kappa coefficient against visibility; both decline as visibility drops. The classification accuracy degrades at a faster rate as visibility gets poorer. The haze becomes intolerable at visibilities less than about 11 km (i.e. ≈ 85% accuracy). For 8 km visibility (moderate haze), accuracy reduces by about 20%. About 70% drop in accuracy occurs between 8 and 0 km visibility. A much sharper decline can be observed for visibilities less than 4 km, with only 50% classification accuracy remaining at about 2 km visibility. It is clear that the kappa coefficient plot shows a consistent result with the classification accuracy plot.

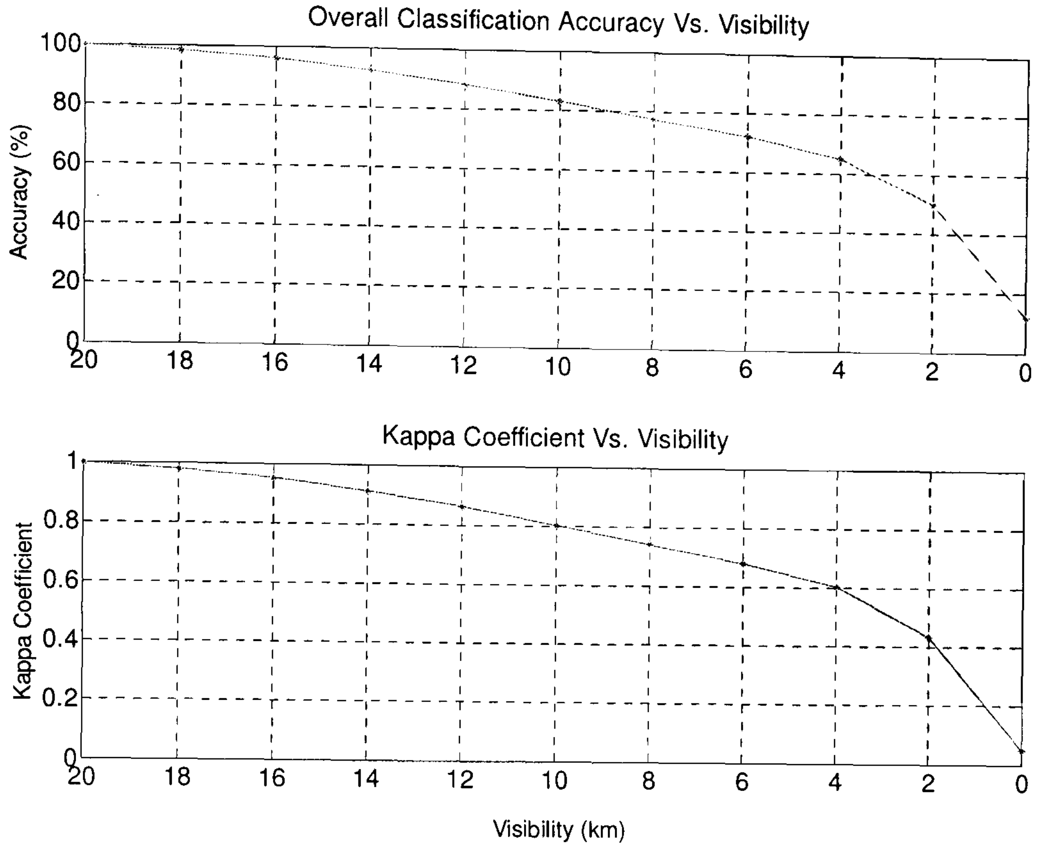


Figure 4.28: *Overall classification accuracy (top) and Kappa coefficient (bottom) versus visibility.*

4.8 ML Classification when Using Training Pixels from the Clear Dataset

In Section 4.5, the effect of haze on classification when training pixels are taken from the haze-affected dataset itself is investigated. In this section, the outcome of using training pixels from the clear dataset in classifying a hazy scene is examined. This is illustrated by Figure 4.18 (column iii) wherein is shown an ML classification for 10 km, 6 km, 2 km and 0 km visibility (b(iii) to e(iii)), using training pixels chosen from a clear scene. The quality of the classifications is poorer than those using training pixels from the hazy dataset itself (Figure 4.18(b(ii)) to (e(ii))). The differences between the outcomes of the two approaches are more evident as visibility gets very short. For example, at 6 km visibility, rubber and coconut cannot be recognised, at 2 km visibility, all land classes become inseparable and at 0 km visibility, no classes can be recognised at all. At extremely short visibilities, most pixels, are influenced by the bright properties of haze and tend to match the properties of industry, consequently are classified as industry.

Figure 4.29 shows plots of the producer accuracy against visibility for the corresponding ML classifications, which show a faster decline for most classes than in Figure 4.24, i.e. the accuracy of the classification degrades more rapidly than when using training pixels from the hazy dataset itself. Some classes reach zero accuracy at visibilities greater than 0 km visibilities. A strange trend occurs for sediment plumes and industry at about 10 km to 6 km visibilities and 6 km to 0 km visibilities, where there is an unexpected increase in the proportion of pixels being correctly classified. We will address this issue later, making use of Figure 4.30.

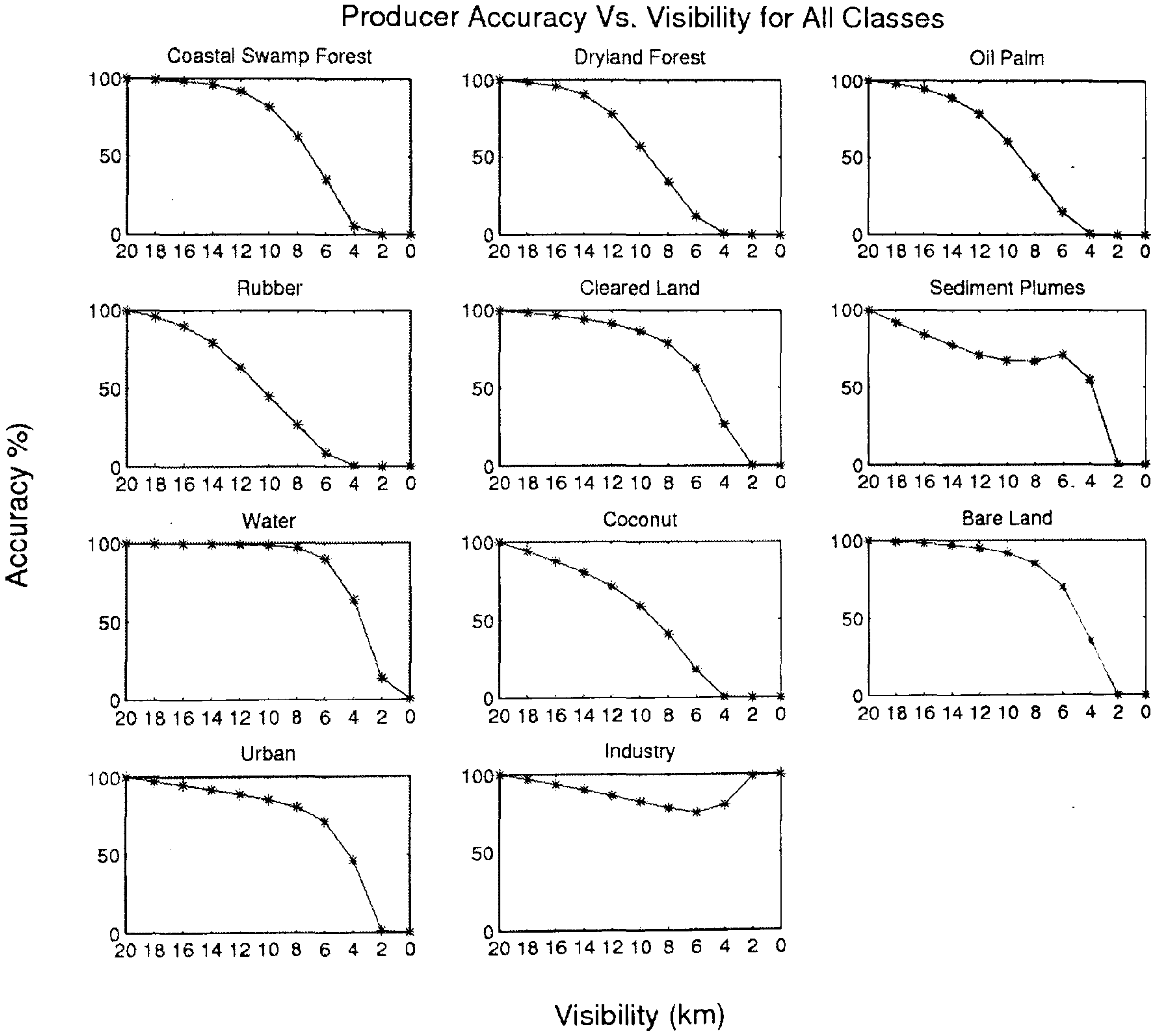


Figure 4.29: *Same as Figure 4.24 but using training pixels from the clear dataset.*

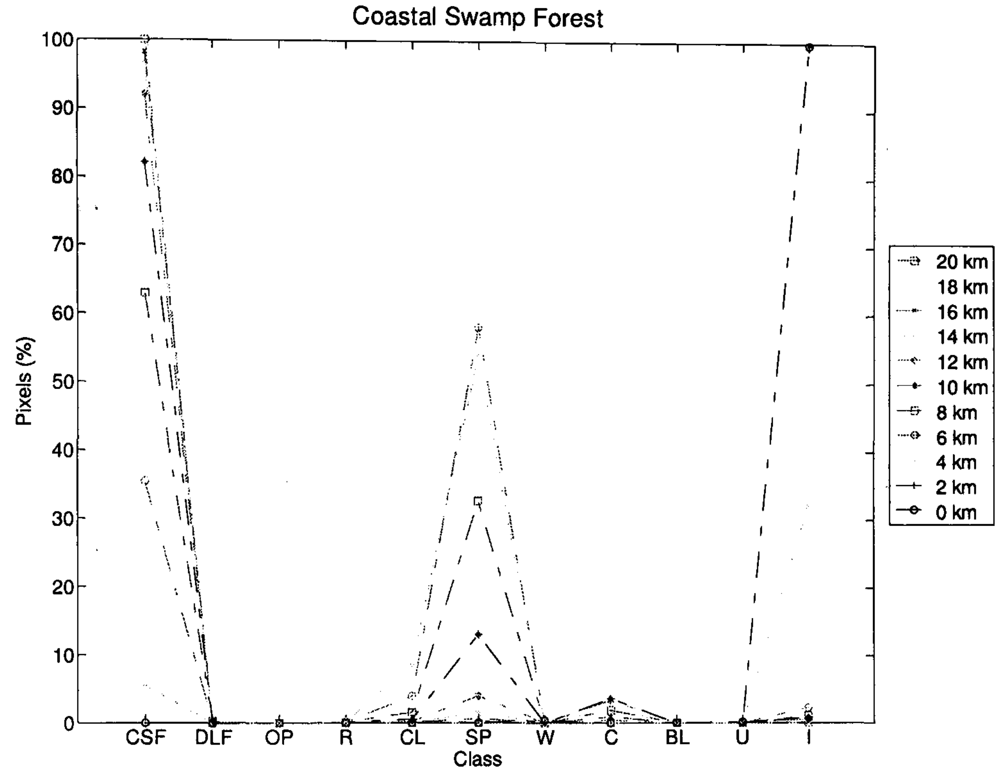
In the plots shown in Figure 4.30, there is a more severe upward trend compared to those in Figure 4.26. This is due to more pixels being misclassified as visibility reduces. The main incorrect classes, which the pixels migrate to, when the visibility reduces are shown in Table 4.9.

Table 4.9: *The main incorrect classes, which the pixels migrate to, as the visibility reduces. The grey shaded boxes are not relevant for this analysis.*

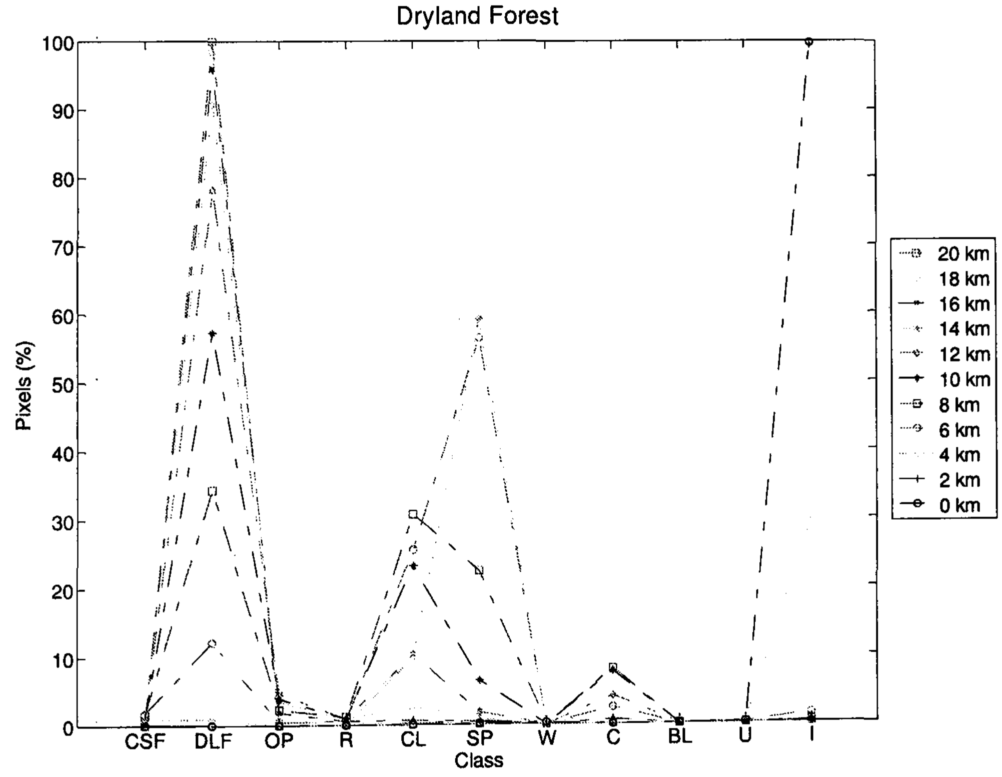
Ground Truth Pixels	Incorrect ML Class which the pixels fall in										
	Coastal Swamp Forest	Dryland Forest	Oil Palm	Rubber	Cleared Land	Sediment Plumes	Water	Coconut	Bare land	Urban	Industry
Coastal Swamp Forest	▨					√					
Dryland Forest		▨			√	√					
Oil Palm			▨		√			√			
Rubber				▨	√						
Cleared Land					▨					√	
Sediment Plumes					√	▨		√			√
Water							▨				√
Coconut			√		√	√		▨			
Bare Land					√				▨		√
Urban					√					▨	
Industry					√					√	▨

A large number of coconut pixels are misclassified as oil palm, cleared land, sediment plumes and industry as visibility reduces ((Figure 4.30(h)). A large number of coastal swamp forest pixels are misclassified as sediment plumes when the visibility drops to less than 10 km (Figure 4.30(a). Dryland forest pixels tend to be misclassified as cleared land and sediment plumes at shorter visibilities (Figure 4.30(b). At 12 km visibility, about 65% of rubber pixels are misclassified as cleared land (Figure 4.30(d)). About 30% of oil palm pixels are misclassified as cleared and coconut at 6 km visibility (Figure 4.30(c)). Urban pixels are misclassified as cleared land for visibilities less than 6 km. About 95%

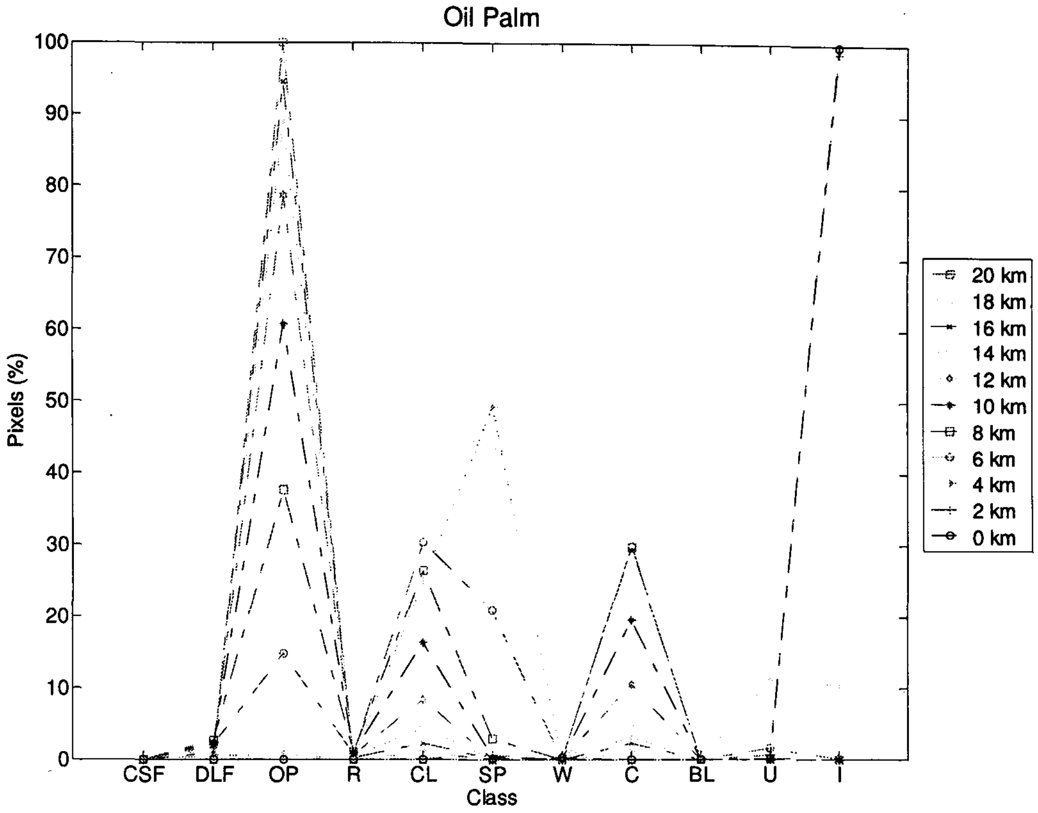
to 100% of non-industry pixels are misclassified as industry for visibilities 2 km and 0 km.



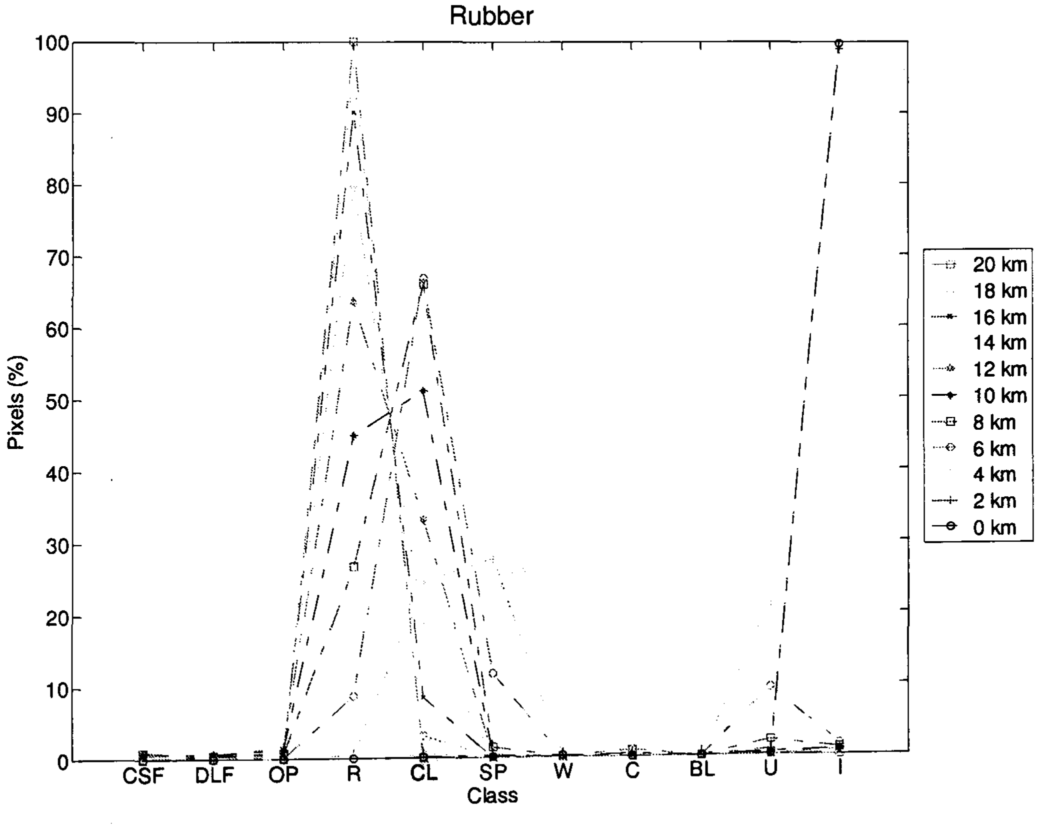
(a)



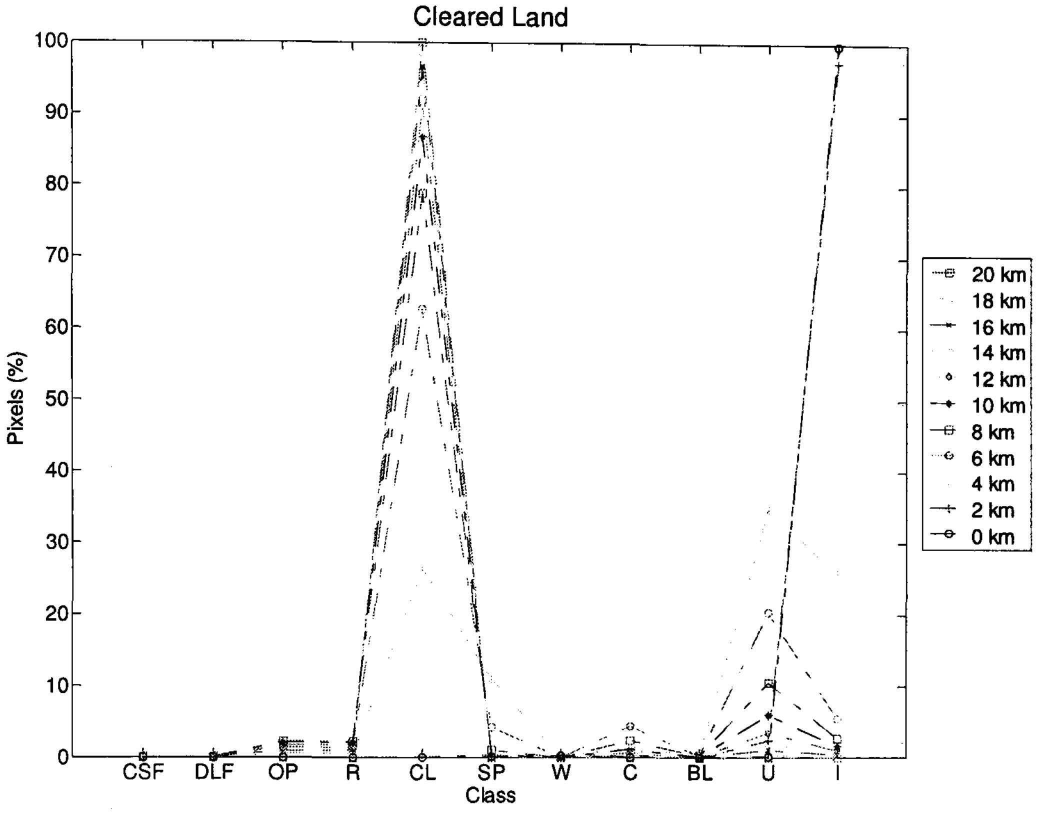
(b)


Oil Palm
100
90
80
70
60
50
40
30
20
10
0
Pixels (%)
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

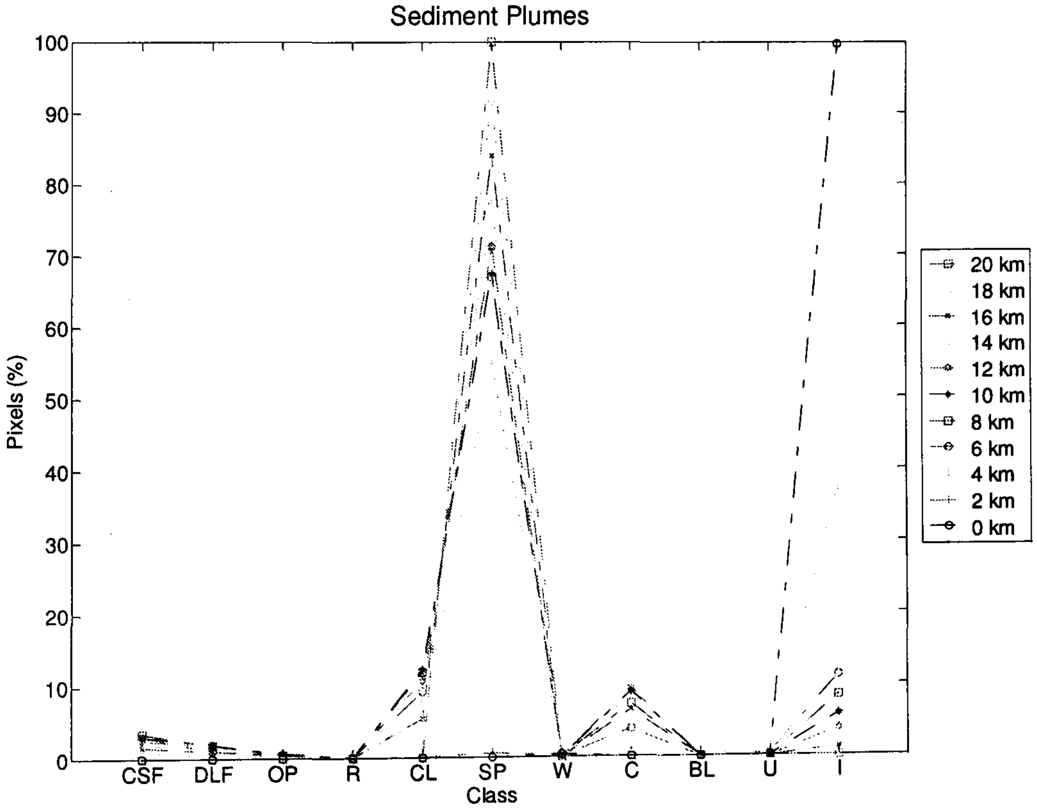
(c)


Rubber
100
90
80
70
60
50
40
30
20
10
0
Pixels (%)
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

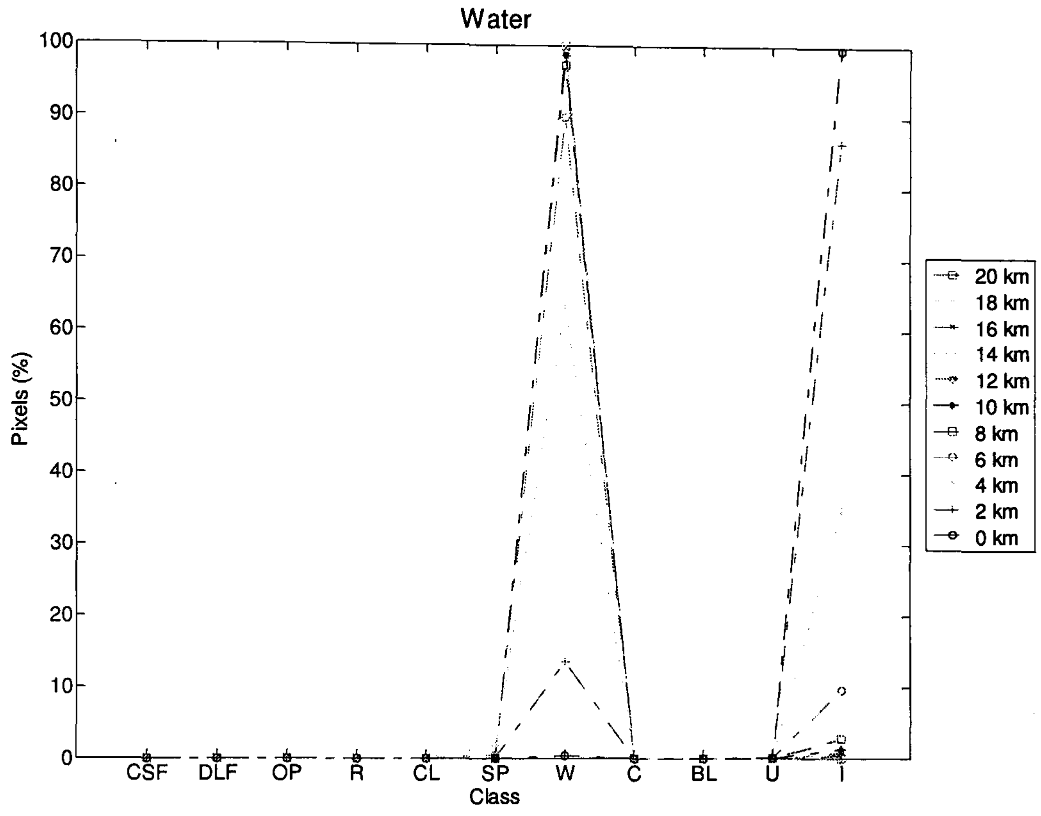
(d)


Cleared Land
100
90
80
70
60
50
40
30
20
10
0
Pixels (%)
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

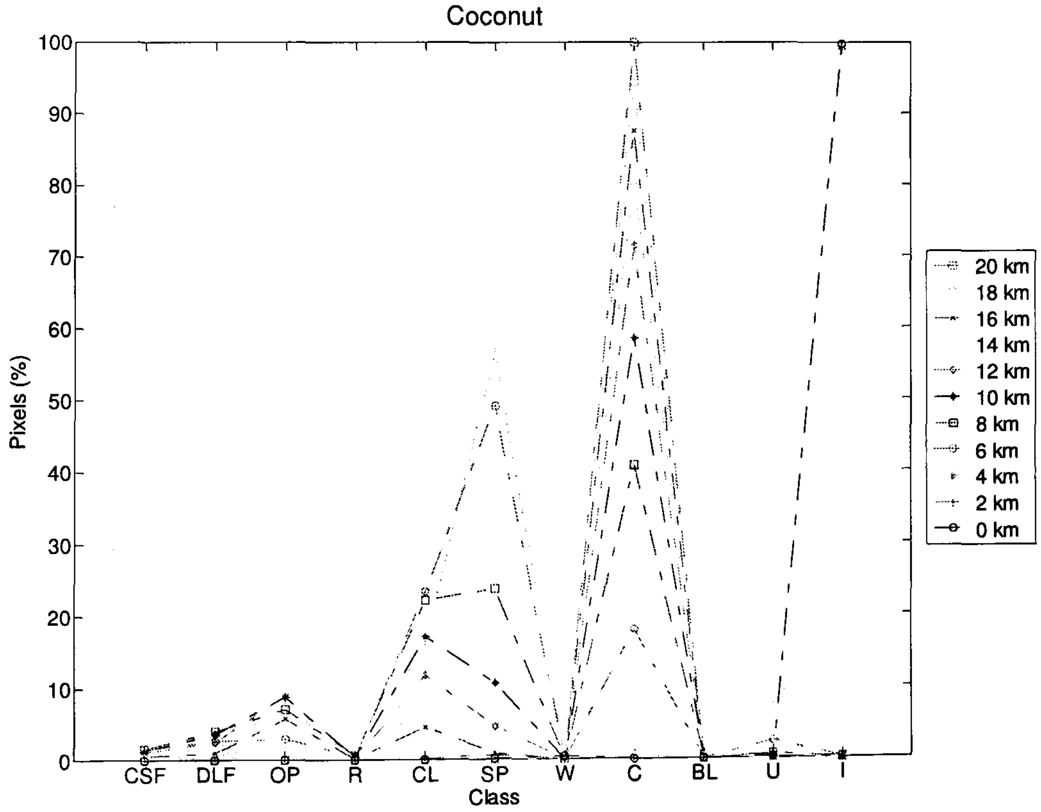
(e)


Sediment Plumes
100
90
80
70
60
50
40
30
20
10
0
Pixels (%)
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

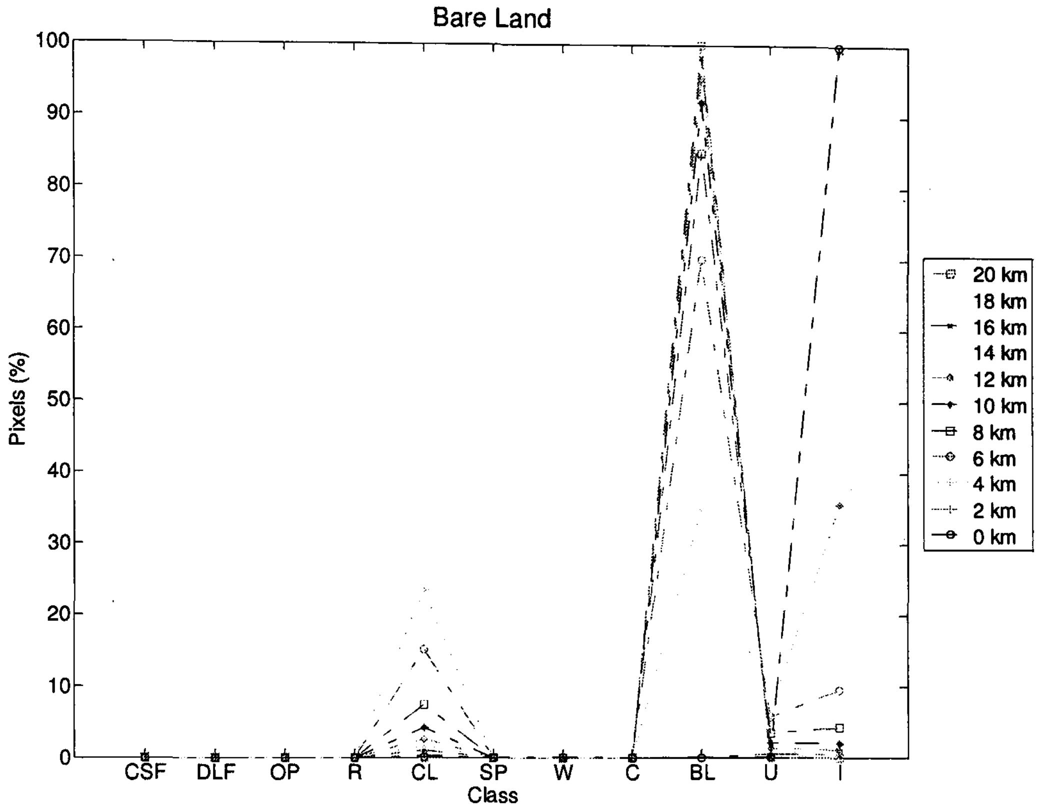
(f)


Water
Pixels (%)
0
10
20
30
40
50
60
70
80
90
100
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

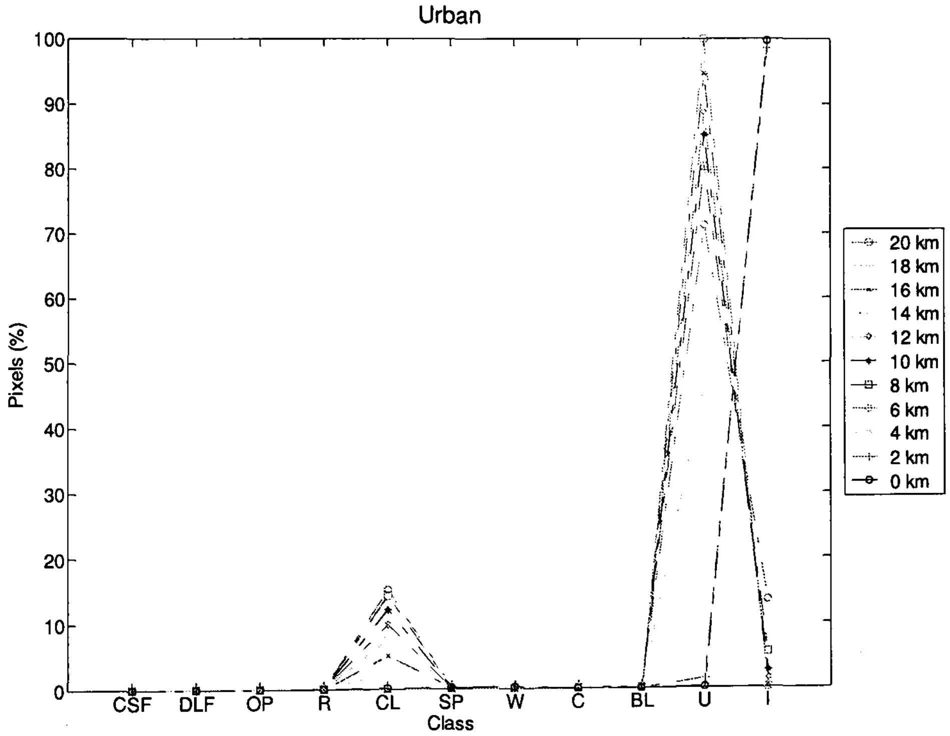
(g)


Coconut
Pixels (%)
0
10
20
30
40
50
60
70
80
90
100
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

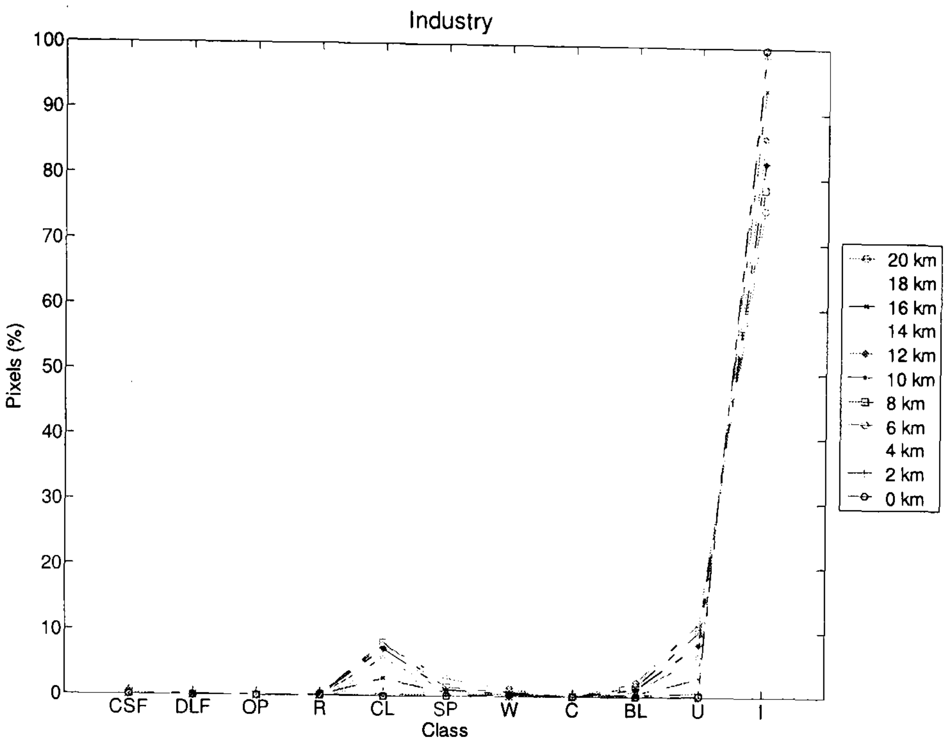
(h)


Bare Land
100
90
80
70
60
50
40
30
20
10
0
Pixels (%)
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

(i)


Urban
100
90
80
70
60
50
40
30
20
10
0
Pixels (%)
CSF
DLF
OP
R
CL
SP
W
C
BL
U
I
Class
20 km
18 km
16 km
14 km
12 km
10 km
8 km
6 km
4 km
2 km
0 km

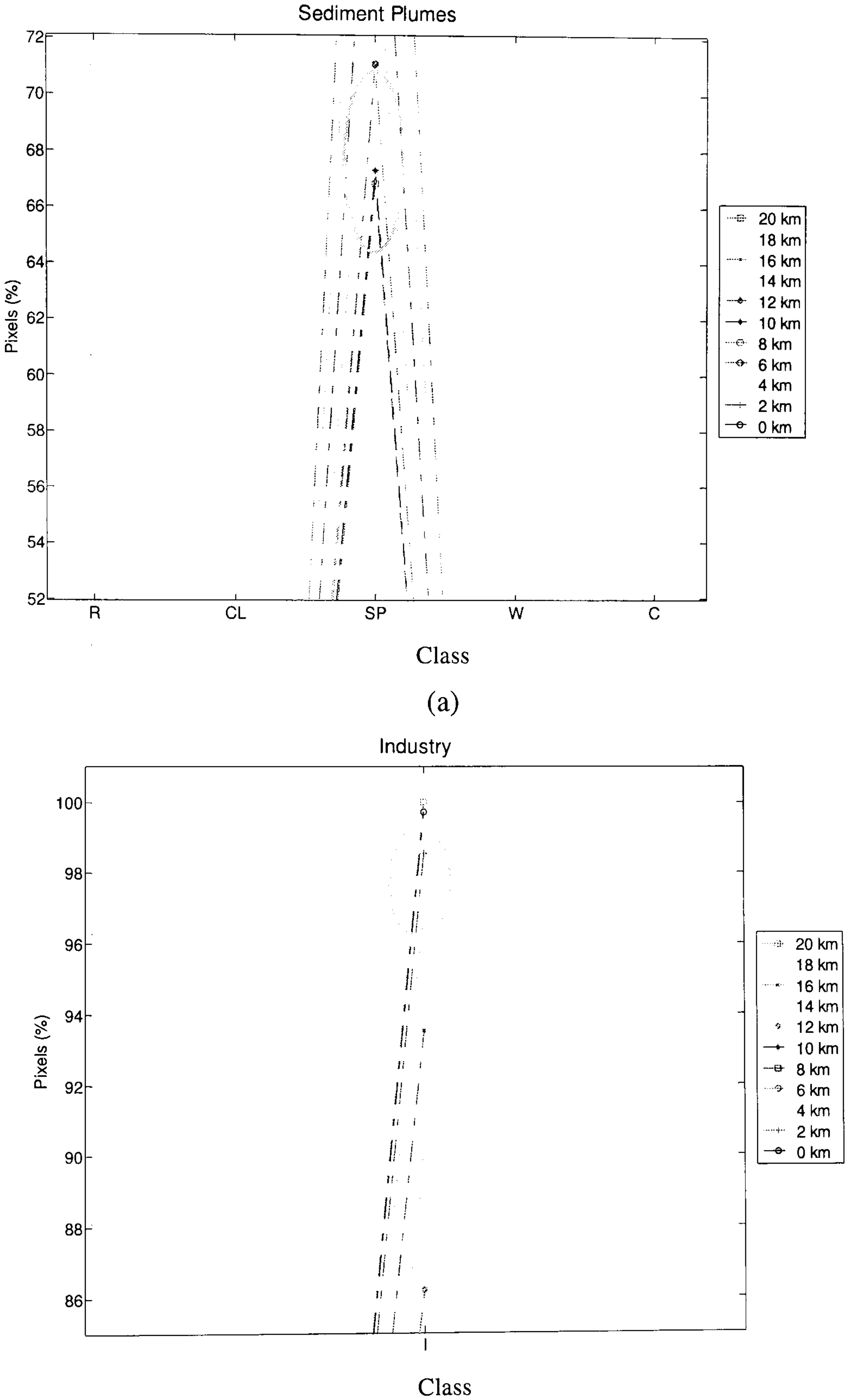
(j)



(k)

Figure 4.30: *Percentage of pixels for (a) coastal swamp forest, (b) dryland forest, (c) oil palm, (d) rubber, (e) cleared land, (f) sediment plumes, (g) water, (h) coconut, (i) bare land, (j) urban and (k) industry, against ground truth classes when ML classification uses training pixels from the clear dataset. 100% for a given class type, represents all the pixels from that class.*

The increase in the producer accuracy of industry from 6 km to 2 km (see producer accuracy plot for industry in Figure 4.29) is due to more pixels being correctly classified as industry at 0 km than 2 km, 4 km and 6 km visibility (Figure 4.31(b)). For 0 km visibility, every pixel experiences very severe signal attenuation and scattering due to haze, eventually posseses spectral properties of the pure haze itself. Since the means and covariance structure of the pure haze pixels across the image match those of industry, consequently, most pixels are classified as industry (see Figure 4.18(f(iii)). This risen the probability of the industry pixel being correctly classified and therefore causing 'strange' increase the producer accuracy. Nevertheless, spatially this is not accurate since non-industry pixels are also classified as industry – we will show this by using user accuracy measure later.



(a)

(b)

Figure 4.31: *An enlarged version of Figure 4.30 (f) and (k) associated with (a) sediment plumes and (b) industry respectively. The unusual trend is located within the green circle.*

Figure 4.32 shows the user accuracies of the classes when using training pixels from the clear dataset. It can be seen that the user accuracy of coastal swamp forest, dryland forest, oil palm, rubber and coconut, reaches 0 at about 2 km visibility (Figure 4.32). The accuracy of water is almost unaffected by haze until about 4 km visibility. Sediment plumes, bare land, urban and industry show a gradual decline at extremely long and short visibilities, but a relatively quick decline at moderate visibilities. Compare to Figure 4.27, the decline is faster particularly at moderate and shorter visibilities, due to the much different condition between the training pixels (i.e. clear) and the pixels (i.e. hazy) of the dataset to be classified. For industry, the unexpected increase in producer accuracy (Figure 4.29) is not consistent with the corresponding user accuracy which is nearly zero at 0 km visibility. This indicate that most pixels on the image that are classified as industry actually does not really represent the class on the ground (spatially), e.g. urban, oil palm and forests pixels being incorrectly classified as industry (see Figure 4.18(f(iii)).

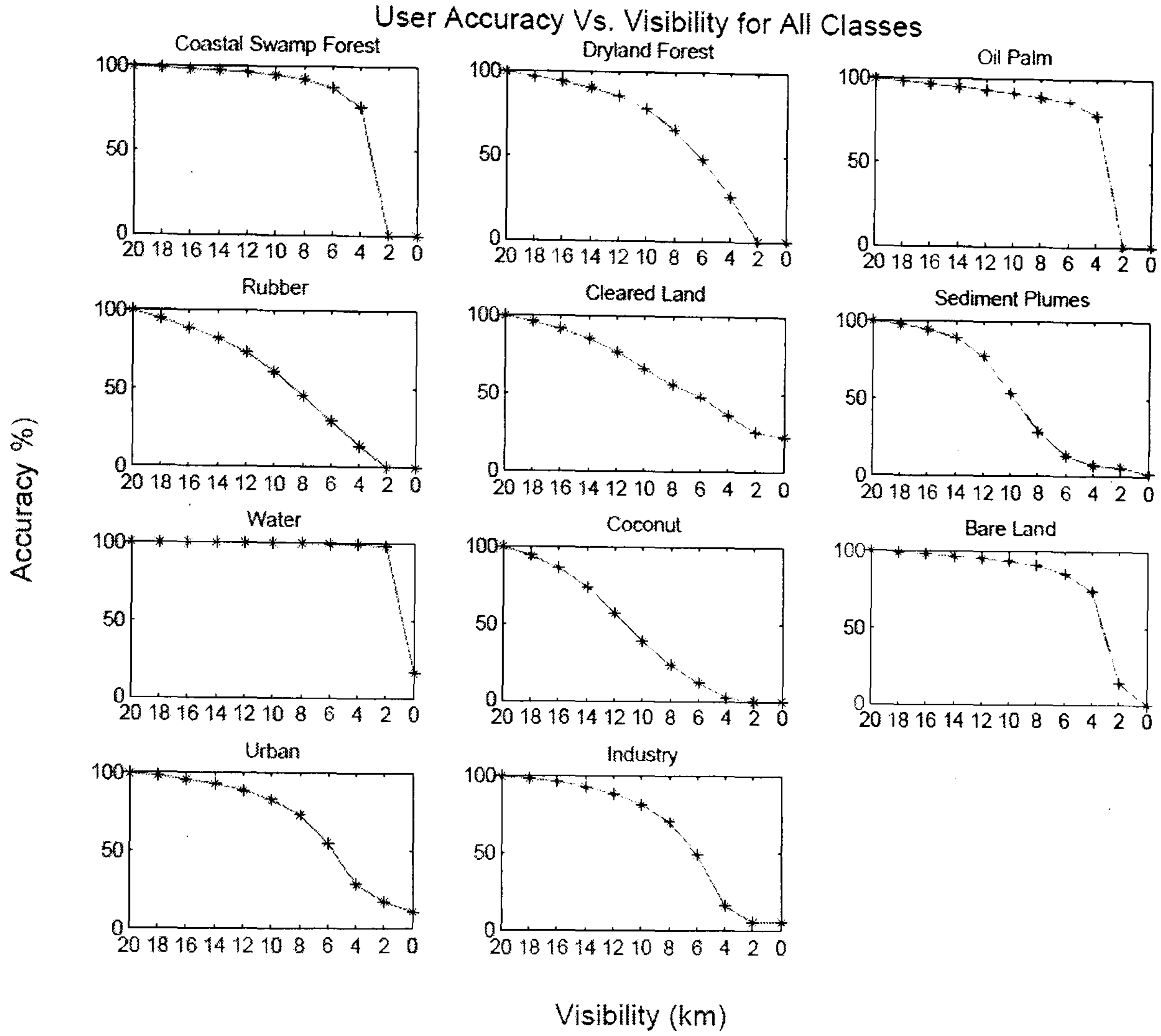


Figure 4.32: *Same as Figure 4.27 but using training pixels from the clear dataset.*

The overall classification accuracy and kappa coefficient (Figure 4.33) drops more quickly than that of Figure 4.28, which shows that haze has more significant effects on ML classification that uses training pixels from the clear dataset than the hazy dataset itself. For visibilities longer than 12 km, about the same accuracies are attained by both approaches, but they differ noticeably as visibility becomes shorter; for example, at 6 km visibility ML classification gives only about 50% accuracy compared with 70% when using training pixels from the hazy dataset. The haze becomes intolerable at visibilities less than about 12 km (i.e. corresponds to 85% accuracy), which indicates an increase of 1 km more where the classification assumed unacceptable compared to that of Figure 4.28.

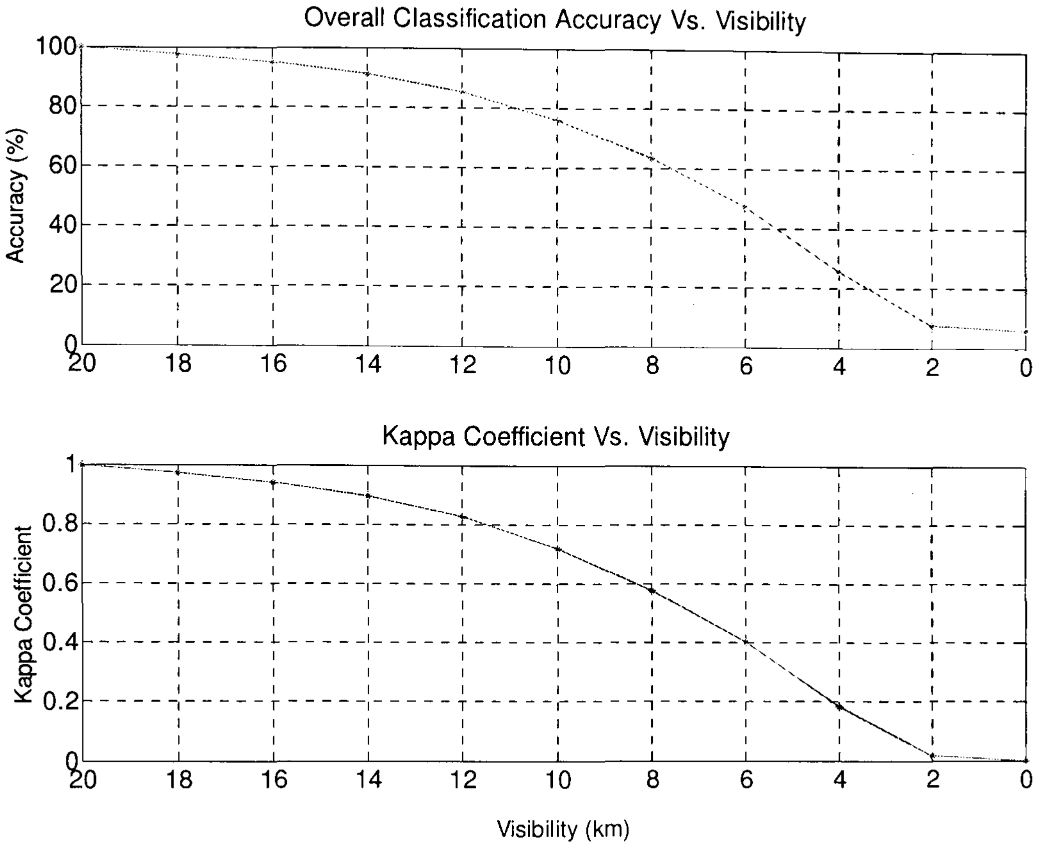


Figure 4.33: *Overall classification accuracy (top) and Kappa coefficient (bottom) versus visibility when training pixels are drawn from the clear dataset.*

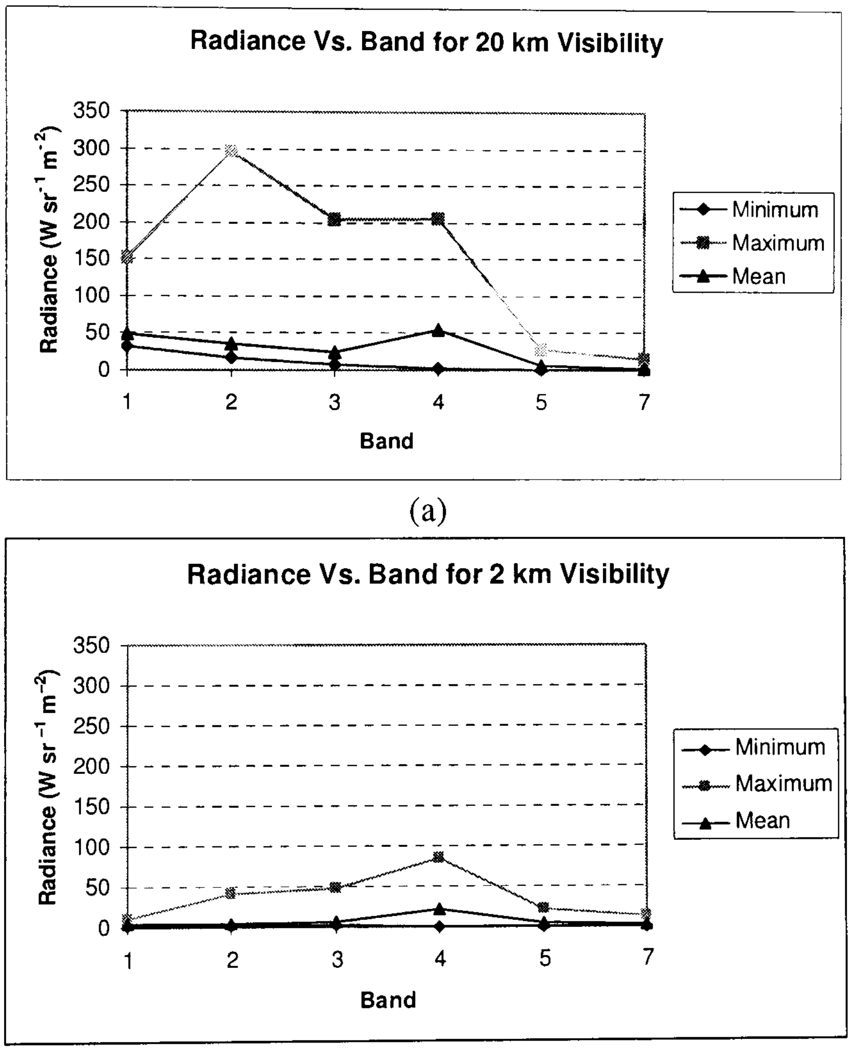
Hence, classification accuracy and producer accuracy decreases faster when the training pixels are drawn from the clear dataset rather than the hazy dataset itself. This suggests that when hazy conditions are unavoidable, it is better to use training pixels from a hazy dataset rather than clear dataset for performing ML classification.

4.9 Classification Accuracy when Neglecting the Haze Scattering Component

Here, we investigate the effects of haze on classification when the haze scattering component is not considered, hence Equation (4.28) becomes:

$$L_i(V) = \left(1 - \beta_i^{(1)}(V)\right)T_i + L_O \qquad \text{... (4.30)}$$

In other word, this equation expressed the observed radiance $L_i(V)$ when taking into account only $\left(1-\beta_i^{(1)}(V)\right)T$ and L_O. Figure 4.34 shows minimum, maximum and mean radiances versus band for (a) 20 km (clear) and (b) 2 km visibility (attenuated signal only). Compared to (a), (b) has a quite low radiances due to very severe signal attenuation.



(a)

(b)

Figure 4.34: *Minimum, maximum and mean radiances versus band for 20 km (clear) (a) and 2 km visibility (attenuated signal only) (b).*

Table 4.10 shows image covariances (along and above the diagonal) and correlations (below the diagonal) calculated from Landsat bands 1, 2, 3, 4, 5 and 7 for the (a) 20 km (b) and 2 km visibility. It can be seen that the 20 km visibility has a bigger covariances than 2 km visibility image due to the bigger pixel radiances as a result from the clear condition. Nevertheless, the correlations (i.e. normalised covariance) for the 20 and 2 km visibility images are the same. When confusion matrix between the 2 km visibility and clear image was drawn, the overall accuracy for the 2 km visibility image was still 100%. Table 4.11 shows the confusion matrix of ML classification for 2 km visibility data (without haze scattering component against 20 km visibility data (clear). The analysis shows that signal attenuation due to haze does not alter the structure of means and covariances that govern ML classification (Equation 3.8); therefore, the accuracy of the classification is not affected.

Table 4.10: *Image covariances (along and above the diagonal) and correlations (below the diagonal) calculated from Landsat bands 1, 2, 3, 4, 5 and 7 for 20 km (clear) (a) and 2 km visibility (without haze scattering component) (b).*

Covariance/ Correlation	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Band 1	118.22	135.99	143.57	-39.73	14.85	7.85
Band 2	0.96	171.12	183.12	-28.95	20.47	9.74
Band 3	0.91	0.96	211.40	-6.90	28.25	12.14
Band 4	-0.17	-0.10	-0.02	465.04	43.09	6.26
Band 5	0.41	0.48	0.59	0.61	10.83	3.20
Band 7	0.66	0.68	0.76	0.27	0.89	1.19

(a)

Covariance/ Correlation	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Band 1	0.72	1.44	2.62	-1.28	0.91	0.51
Band 2	0.96	3.14	5.82	-1.62	2.17	1.10
Band 3	0.91	0.96	11.61	-0.67	5.18	2.37
Band 4	-0.17	-0.10	-0.02	79.65	13.95	2.16
Band 5	0.41	0.48	0.59	0.61	6.63	2.09
Band 7	0.66	0.68	0.76	0.27	0.89	0.83

(b)

Table 4.11: *Confusion matrix of ML classification of 2 km visibility data (without haze scattering component against 20 km visibility data (clear).*

	Ground Truth (Pixel)											
Class	Coastal Swamp Forest	Dryland Forest	Oil Palm	Cleared land	Sediment Plumes	Water	Coconut	Bare Land	Urban	Industry	Rubber	Total
Coastal Swamp Forest	39219	0	0	0	0	0	0	0	0	0	0	39219
Dryland Forest	0	32616	0	0	0	0	0	0	0	0	0	32616
Oil Palm	0	0	147305	0	0	0	0	0	0	0	0	147305
Cleared land	0	0	0	115038	0	0	0	0	0	0	0	115038
Sediment Plumes	0	0	0	0	24759	0	0	0	0	0	0	24759
Water	0	0	0	0	0	68846	0	0	0	0	0	68846
Coconut	0	0	0	0	0	0	41398	0	0	0	0	41398
Bare Land	0	0	0	0	0	0	0	8896	0	0	0	8896
Urban	0	0	0	0	0	0	0	0	65174	0	0	65174
Industry	0	0	0	0	0	0	0	0	0	37281	0	37281
Rubber	0	0	0	0	0	0	0	0	0	0	19804	19804
Total	39219	32616	147305	115038	24759	68846	41398	8896	65174	37281	19804	600336

4.10 Summary and Conclusions

This chapter modelled and simulated hazy datasets, as summations of a weighted signal component and a weighted pure haze component. It then examined the effects of haze on land classification.

1. Classification accuracy and producer accuracy decreases faster with visibility when the training pixels are drawn from the clear dataset rather from the hazy dataset itself.

2. The haze becomes intolerable (i.e. the accuracy fell to below 85%) at visibilities less than about 11 km when using training pixels from a hazy dataset and 12 km visibility when using training pixels from a clear dataset.

3. When haze gets very thick, spectral signatures curves of land covers become very close to each other, approximating the pure haze spectral signature.

4. The increase in class mean and standard deviation as haze increases are particularly significant for less reflective classes because the radiance is scattered by the haze directly into the satellite's field of view (i.e. path radiance) dominates the radiance observed by the satellite.

5. A decrease in radiance tends to occur for bright classes because the haze scatters some of the solar radiance out of the satellite's field of view before reaching the ground, and attenuates the reflected radiation on its way back.

6. The modification that haze made to band correlations varies, depending on the radiance of the classes and the severity of the haze; correlations are very high in very hazy conditions because the band covariances are dominated by the high correlation of the pure haze.

7. The weightings used is due to the apparent difference between attenuation of target radiance (i.e. corresponds to β_1) and total scattering of incoming radiance (i.e. corresponds to β_2) as haze severity increases, particularly for measurement made from shorter wavelengths.

Chapter 5

Haze Removal from Satellite Data

5.1 Introduction

This chapter addresses one of the most crucial aims of this thesis, which is to develop methods to mitigate the effects of haze on land cover classification. We have demonstrated that haze modifies the spectral signatures of land classes and reduces classification accuracy, so causing problems to users of remote sensing data. Hence, we need to reduce the haze effects to improve the usefulness of the data. In Chapter 4, we modelled hazy satellite data as $L_i(V) = \left(1-\beta_i^{(1)}(V)\right)T_i + L_O + \beta_i^{(2)}(V)H_i$, where, $L_i(V)$, T_i, H_i, $\beta_i^{(1)}(V)$, $\beta_i^{(2)}(V)$, L_O and V are the true signal component, the pure haze component, the signal attenuation factor, the haze weighting, the radiance scattered by the atmosphere and the visibility for band i. From this equation, it is clear that the degradation of hazy satellite data is caused by haze scattering and signal attenuation characterised by $\beta_i^{(2)}(V)$ and $\beta_i^{(1)}(V)$ respectively. Ideally, to reduce the haze effects and restore the surface information, we need to reduce the former so that $\beta_i^{(2)}(V)H_i \approx 0$ and restore the latter so that $\left(1-\beta_i^{(1)}(V)\right)T_i \approx T_i$.

In practice, the effects of signal attenuation through $\beta_i^{(1)}(V)$ are not significant, as shown in Section 4.9, so their removal is not important. On the other hand, the effects of $\beta_i^{(2)}(V)H_i$ is very significant; therefore, this chapter is concerned mainly with reducing $\beta_i^{(2)}(V)H_i$.

The primary aim of this chapter is to develop and test a haze removal method. In developing the method, we need to know the existing methods and issue encountered; this leads us to a review of a number of haze removal methods (Section 5.2).

Since the primary issue is to develop haze removal, we need to define physical processes for removing haze. Section 5.3 clarifies the concepts of haze removal and mathematically analyses these processes, and in Section 5.4 the haze removal procedures are carried out.

An important issue for haze removal is to assess its performance. In Section 5.5, we discuss how this is defined and how we can measure it given the hazy and reference satellite data.

We are left wanting to know the actual performances of the haze removal and how they compare to other methods. In Section 5.6, we describe procedures for testing haze removal onto real hazy data, and give an extended analysis of haze removal by comparing the results with Liang et al. (2001) method.

5.2 Previous Studies

Haze removal, in practice, should be usable at any time and independent from auxiliary information, e.g. haze path radiance and meteorological information (Du et al., 2002), which is unavailable in most cases due to a lack of ground stations. Initially, studies attempted to determine and remove uniform haze path radiance, but later spatially-varying haze was taken into account.

Initially, the most popular procedure was dark-object subtraction (DOS) which considers uniform haze (Chavez, 1988). In order to determine the haze path radiance, it is assumed that there are some pixels within the image that are totally black (dark objects); this is usually caused by topography or cloud shadows. A dark object is assumed to be unable to reflect any solar energy and thus should possess zero DN or zero reflectance. If haze exists, these pixels do not appear completely dark because solar energy is scattered into the satellite's field of view by the haze. From the histogram of a particular visible band, this effect can be seen as a sharp increase in occurrence frequency in the lower DN region. The DN value that corresponds to this increase is assumed to be the amount of haze in that particular band. This needs to be subtracted from the entire image for that band to correct for the haze, although smaller occurring DNs may also represent haze. An example is given in Figure 5.1, where frequency of occurrence suddenly increases from 10 to 110 between DN value 60 and DN 61. Hence, the haze value is taken to be 61. Although this is easy and practical, it is quite ambiguous in most cases, since the shadow pixels caused by topography and clouds may not actually have zero DN due to secondary energy scattered from other objects into the shadowed area; thus the haze value selected from the histogram may not correspond to a real dark object. This can lead to over-correction for haze and consequently cause truncation of the values for some surface pixels. Hence, this method is not considered further.

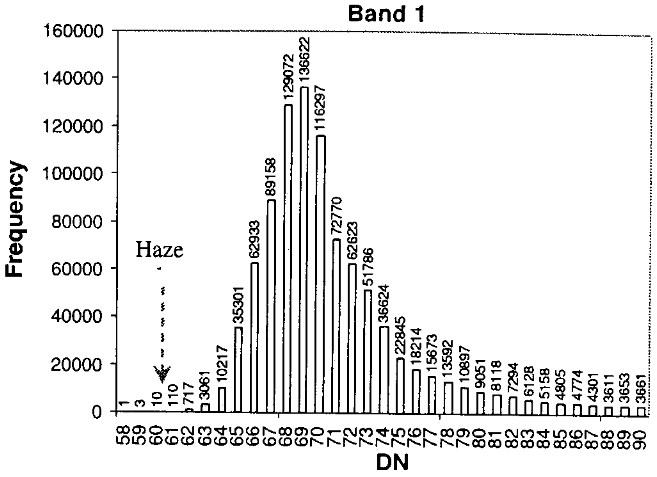


Figure 5.1: *A sharp increase in a Landsat-5 TM band 1 histogram indicating the haze value.*

Chavez (1988) proposed an improvement to the standard DOS, where an atmospheric scattering model based on wavelength power law models was introduced. In this approach, the wavelength dependency of atmospheric scattering for different haze severity levels is considered, i.e. very clear (λ^{-4}), clear (λ^{-2}), moderate (λ^{-1}), hazy ($\lambda^{-0.7}$) and very hazy ($\lambda^{-0.5}$) (where λ is the centre wavelength of a satellite band). For a satellite band, these relationships approximate true atmospheric scattering. From this model, by knowing the haze value for a single band (i.e. using the original DOS method as proposed by Chavez (1988)), the haze values for the rest of the bands can be determined. An example using band 1 as the starting haze band is given in Table 5.1. Subsequently, multiplication factors can be calculated by dividing the scattering contributions in Table 5.1 (bold) by the corresponding total scattering contribution. Based on the haze severity, the haze values in bands 2, 3, 4, 5 and 7 can be determined by multiplying the haze value in band 1 by the multiplication factor associated with that particular band.

Table 5.1 : *The scattering contributions and multiplication factors associated with different relative scattering models for Landsat bands.*

TM	λ (μm)	Scattering Contributions					Multiplication Factors				
		Very Clear λ^{-4}	Clear λ^{-2}	Moderate λ^{-1}	Hazy $\lambda^{-0.7}$	Very Hazy $\lambda^{-0.5}$	Very Clear λ^{-4}	Clear λ^{-2}	Moderate λ^{-1}	Hazy $\lambda^{-0.7}$	Very Hazy $\lambda^{-0.5}$
1	0.485	**18.073**	**4.251**	**2.062**	**1.660**	**1.436**	1.000	1.000	1.000	1.000	1.000
2	0.560	10.168	3.189	1.786	1.501	1.336	0.563	0.750	0.866	0.904	0.931
3	0.660	5.270	2.296	1.515	1.338	1.231	0.292	0.540	0.735	0.806	0.857
4	0.830	2.107	1.452	1.205	1.139	1.098	0.117	0.341	0.584	0.687	0.764
5	1.650	0.135	0.367	0.606	0.704	0.778	0.007	0.086	0.294	0.424	0.542
7	2.215	0.042	0.204	0.451	0.573	0.672	0.002	0.048	0.219	0.345	0.468
Total contributions		35.795	11.758	7.625	6.914	6.551	-	-	-	-	-

Chavez (1988) claimed that the improved DOS method produces a more realistic haze DN for all the satellite bands compared to the standard DOS.

Scott et al. (1988) developed a scene-to-scene radiometric normalisation technique for haze correction that is based on the statistical invariance of the reflectance possessed by objects known as pseudoinvariant features (PIF). For a scene, Scott et al. (1988) suggested that the PIF pixels can be chosen from man-made objects such as road surfaces, roof tops and parking lots. For a feature (e.g. road surfaces), one or more pixels can be chosen, depending on the size of the feature (more pixels can be chosen from a large PIF provided they have nearly constant reflectance); however, Scott et al. (1988) did not discuss in detail the minimum size requirement for a single PIF. The relationship between PIF pixels from a hazy scene and from a clear reference scene is assumed to be represented by a linear equation, which can be generated by regressing the DNs of PIF pixels from a hazy dataset against those of a reference dataset from the same band. The entire hazy dataset is then transformed using this linear equation. In other words, this transformed dataset predicts what the hazy dataset would look like if it possessed the same atmospheric condition as the clear dataset. Hence, the relationship between the current and hazy dataset is expressed as:

$x_t = m_c x_h + b_c$ where x_t and x_h are pixel brightness for band k in the current dataset and hazy dataset respectively; $m_c = \frac{\sigma_r}{\sigma_h}$ and $b_c = \overline{x}_r - m_c \overline{x}_h$ where σ_r and $\overline{x}_r$ are the estimated standard deviation and mean of the PIF pixels for the reference dataset, and σ_h and $\overline{x}_h$ are those of the hazy dataset.

$$\sigma_r = \sqrt{\frac{1}{(n_r - 1)} \sum_{i=1}^{n_r} \left(x_r(i) - \overline{x_r}\right)^2} \text{ and } \sigma_h = \sqrt{\frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} \left(x_h(i) - \overline{x_h}\right)^2}$$

n_r and n_h are the numbers of PIF pixels (from different features) in the reference and hazy datasets respectively. An example using a Landsat band 1 data for Bukit Beruntung, located in Selangor, Malaysia, is illustrated in Figure 5.2, which shows the regression between the DN of PIF pixels from a hazy dataset from 6 August 2005 with 6 km visibility and from a dataset from 22 August 2005 with 12 km visibility which is free of haze. The PIFs were selected based on knowledge of the study area aided by Google Maps and consist of road surfaces (DN 70 -80), the rooftops of industrial buildings (DN 90 – 100) and houses (DN 55 - 65). Note that these DNs were taken from the clear data. It can be seen that for each point, the DN from the hazy data (6 August 2005) is bigger than that of the clear data, indicating the existence of haze. There is a good linear correlation between hazy and clear PIF pixels with $R^2 = 0.9032$. Consequently, the haze-reduced dataset can be obtained by using:

$$\text{Corrected dataset} = 0.7422 \times \text{Hazy dataset} + 13.072$$

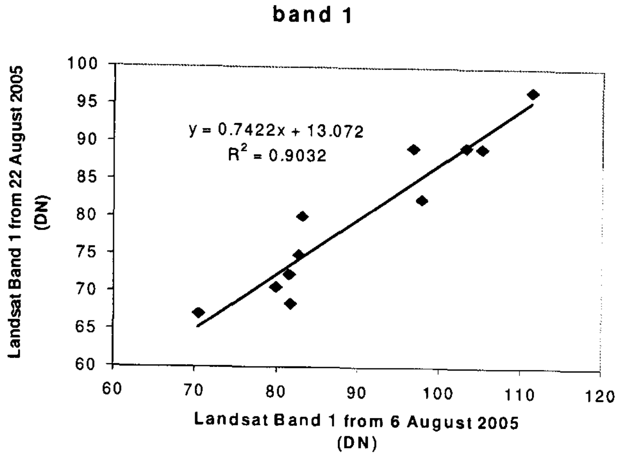


Figure 5.2: *An example of applying the PIF method to 400 x 400 pixels from a band 1 Landsat data for Bukit Beruntung, located in Selangor, Malaysia; the acquisition dates of the hazy (6 km visibility) and 12 km visibility datasets are from 6 August 2005 and 22 August 2005 respectively. The DN values for the PIF pixels for the clear dataset (y-axis) are plotted against those from the hazy dataset (x-axis). The solid line is the linear regression line.*

Unlike the DOS method which uses dark features that produce a weak radiance, the PIF method uses bright surfaces. Hence, the additive effects of secondary scattering on the PIFs can be neglected.

In Sections 5.5 and 5.6, we will use the PIF concept to estimate haze path radiance from simulated and real hazy datasets. This method can be extended to non-uniform haze by subdividing the hazy scene into smaller subscenes by using the Minimum Noise Fraction (MNF) method (Green et al., 1988) in which the haze is relatively uniform. However, it is essential to ensure that these subscenes contain PIFs so that the haze path radiance can be calculated and the individual segments can be corrected for haze.

A similar method was presented by Eckhardt et al. (1990) who used normalisation targets, which have the same functions as the PIFs described by Scott et al. (1988). The main improvement is that Eckhardt et al. (1990) suggested the criteria that should be met by these targets:

- The visual appearance (from satellite imagery) of the normalisation targets should not change over time; the method requires their reflectance to be constant over time.
- Targets must be roughly at the same elevation as other objects' surfaces in the study area and should not be too far above sea level. This is because most haze scattering occurs within the lowest 1000 m of the atmosphere, so choosing a target at a relatively high altitude (e.g. a mountain top) will miss most of the haze.
- Targets should contain little or no vegetation, because environmental stress and plant phenology can affect the spectral radiance of vegetation.
- Targets should be approximately flat so that they will have the same proportional changes in sun angle for different acquisition dates.
- Targets should consist of pixels with a wide range of radiance values so that a representable regression model of pixels from reference and hazy datasets can be produced.

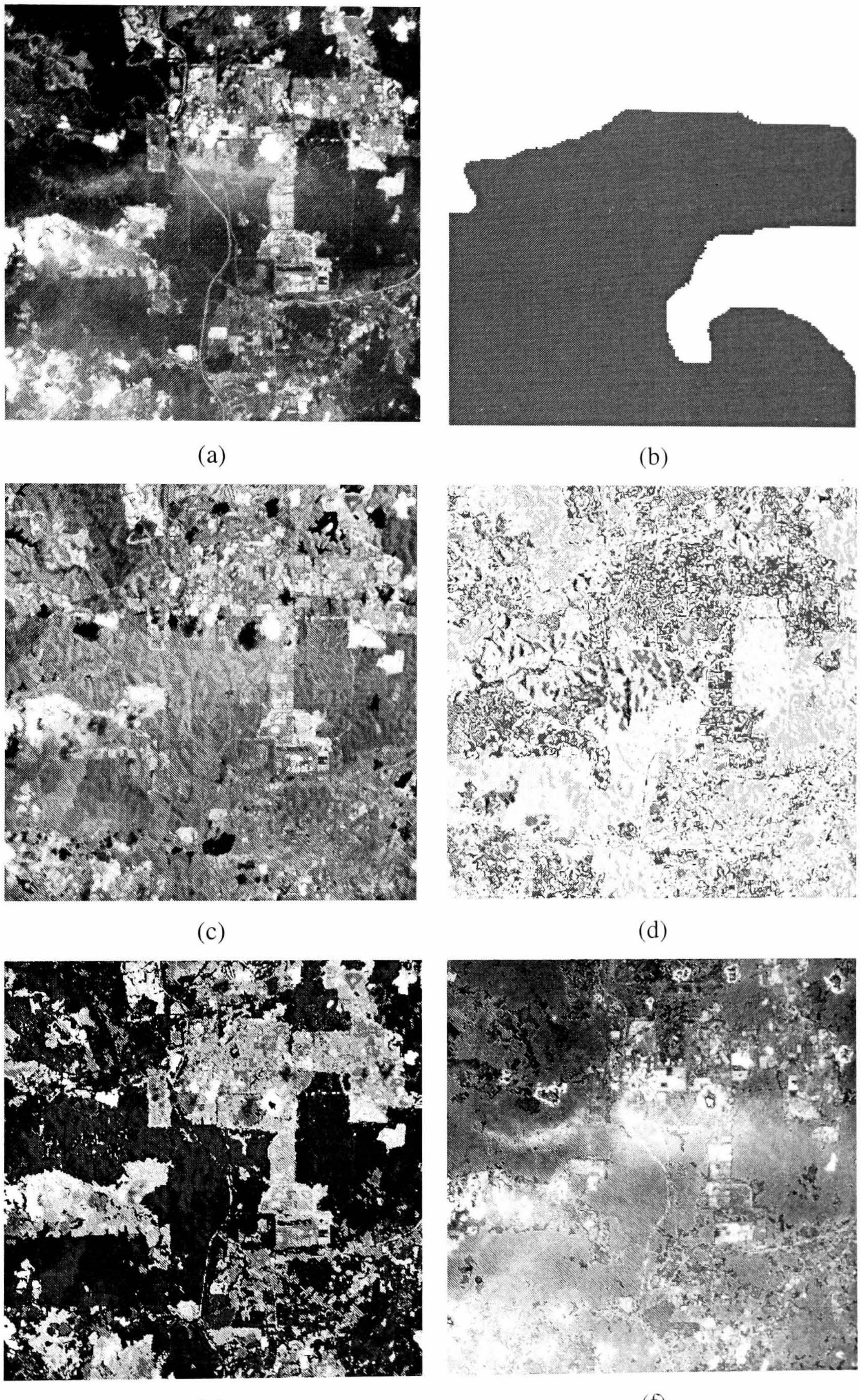
Eckhardt et al. (1990) then applied the regression equation to the test scene in order to produce image normalisation. The criteria described by Eckhardt et al. (1990) can be applied to the PIF method and are relevant to this study. However, the suitability of the size of a PIF and variability of DN within a PIF are not discussed. These criteria and other issues related to the PIFs used in our study will be discussed in detail in Section 5.4.

Liang et al. (2001) presented an atmospheric correction method that takes into account non-uniform haze within a Landsat dataset, by combining image-based and radiative transfer equation approaches. Initially, the near-infrared bands 4, 5 and 7, which are less affected by haze, are used to classify pixels into cover types. For this purpose, they used an unsupervised classification method, where 20 to 50 clusters are generated, depending on the complexity of the landscape. They then separated clear and the hazy regions by enhancing the boundaries between hazy and clear regions and then visually analysing and drawing the hazy regions using image processing software. Liang et al. (2001) suggested that the boundaries can be enhanced using one of the following methods: (i) the fourth component of the Tasseled Cap transformation (Crist and Cicone, 1984), (ii) the ratio of bands 1 and 4 or (iii) the visible bands 1, 2 and 3; the last one is often used because it is simple and effective. Next, they determined the mean reflectances of clear regions and matched with those of the hazy region from the same cluster. They then subtracted the mean reflectances of the cluster from the hazy reflectances in order to determine the haze reflectance. With the assumption that the distribution of haze reflectance is smoother than surface reflectance, Liang et al. (2001) subsequently used a low-pass smoothing (i.e. using 5 x 5 window) to determine the distribution of the haze reflectance in each band. Finally, they determined the corrected surface reflectance for each band by subtracting the corresponding haze reflectance from the hazy data. Liang et al. (2001) claimed the method visually removed non-uniform haze from bands 1, 2 and 3.

An example of the application of this method for Bukit Beruntung in Selangor, Malaysia from a 400 x 400 Landsat dataset from 6 August 2005 is given here (this was the same data as was used in the PIF analysis above). Bands 3, 2 and 1 (Figure 5.3(a)), which are significantly affected by haze, are first used to visually determine the hazy (blue) and clear (yellow) regions (Figure 5.3(b)); this is done by enhancing the hazy-clear regions using contrast manipulation method (i.e. contrast stretching) (Lillesand et al. 2004) and delineating a boundary between them using the built-in applications in the ENVI software. Next, Bands 4, 5 and 7 (Figure 5.3(c)), which are less affected by haze, are used as input for ISODATA clustering to produce 35 clusters for the entire area (Figure

5.3(d)). Subsequently, the hazy-clear regions developed in (a) are used as a template for the 35-clusters map. By doing so, each cluster was then subdivided into two parts: the parts that fall within the hazy region and the clear region.

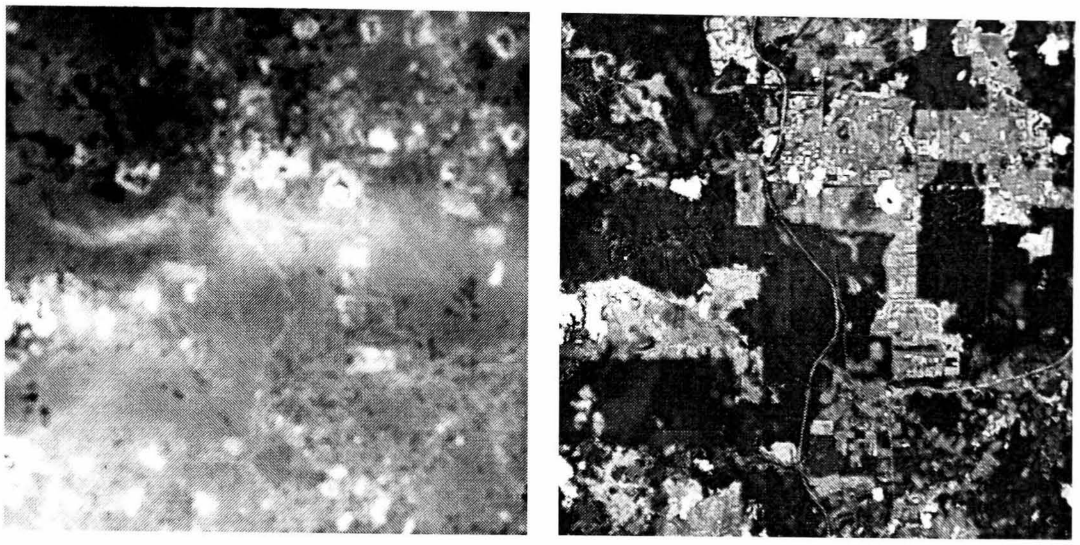
For each cluster, pixel reflectances from the hazy region are replaced with the mean reflectance from the clear region; this process was carried out for all 35 clusters and for bands 1, 2 and 3 (bands 4, 5 and 7 are almost unaffected by the haze). Figure 5.3(e) shows bands 3, 2 and 1 assigned to red, green and blue channels after the mean reflectance replacement. Next, the haze reflectance in each band is determined by subtracting the mean reflectances from the hazy data and then filtering it with a 5 x 5 average filter. The results for band 1 after the subtraction of mean reflectances and filtering are shown in Figure 5.3(f) and (g); the latter represent the haze reflectance for band 1. Finally, for each band, the haze reflectance is subtracted from the hazy data in order to determine the corrected surface reflectances for that band. Figure 5.3(h) shows the corrected bands 3, 2 and 1 assigned to red, green and blue channels. It is clear that the haze has been removed and the edges of certain features, e.g. roads and urban areas, have been restored, but some detailed structures within the urban areas are lost. This is mainly due to the effects of clustering and mean reflectances replacement, which heavily depends on the accuracy of the hazy-clear boundary. Besides that, a clear weakness of the method of Liang et al. (2001) is the use of visual analysis in determining the hazy-clear boundary, which is very subjective and exposed to inaccuracy.



(a) (b)

(c) (d)

(e) (f)



(g) (h)

Figure 5.3: *The outcome of applying Liang's method to Landsat dataset from Bukit Beruntung: (a) bands 3, 2 and 1 assigned to red, green and blue channel, (b) the corresponding hazy (blue) and clear (yellow) regions, (c) bands 4, 5 and 7 assigned to red, green and blue channel, (d) the 35-clusters map generated using the ISODATA clustering, (e)same as (a) but after mean reflectances replacement, (f) the result after subtracting (e) from band 1 in (a), (g) same as (f) but after 5 x 5 average filtering and (h) same as (a) but after subtracting the haze reflectance in each band.*

Liang et al. (2002) carried out validation for the study done by Liang et al. (2001) with several approaches, viz. use of ground radiometric measurements, applying the method on datasets from MODIS and SeaWIFS, and analysing the classification, change detection and broadband albedo performances. Liang et al. (2002) indicated that the method successfully removed haze effects but not all the validation work was quantitative, e.g. in terms of classification performance, only visual assessment was performed but no quantitative analysis was carried out.

From the above discussion, a major issue in removing haze is to deal with the case of non-uniform haze. It is obvious that the PIF methods have been used mainly in the case of uniform haze. On the other hand, the method of Liang et al. (2001) takes into account non-uniform haze, but the main weakness is the use of visual analysis to separate hazy and clear regions, which is exposed to inaccuracy. Another weakness is the assumption that clear (or less hazy) regions exist within a hazy image, which is not true in most cases. Although visually successful in removing haze, there is a lack of quantitative analysis in validating their method. The primary aims of our study are to develop and test a haze removal method. The haze removal will exploit the PIF method, whereas the non-uniformity of haze will be dealt with the MNF method. The testing of the developed method will make use of classification accuracy.

In order to achieve these aims, we first need to explore the related development on haze removal to date; this will lead to a review of existing methods (Section 5.2). Next, we need to describe the fundamental processes concerning haze removal; these are explained systematically by making use of mathematical models (Section 5.3). In implementing the haze removal, a procedure is developed by making use of the PIF method to estimate the haze radiance (Section 5.4). Subsequently, the performance of the haze removal is measured by means of classification accuracy (Section 5.5). Finally, testing of the haze removal on a real hazy dataset is carried out and comparison with the other method is made (Section 5.6).

5.3 General Concepts of Haze Removal

In Chapter 4, we developed a statistical model for hazy satellite data, which can be expressed as:

$$L_i(V) = \left(1 - \beta_i^{(1)}(V)\right)T_i + L_O + \beta_i^{(2)}(V)H_i \qquad \text{... (5.1)}$$

where $L_i(V)$, T_i, H_i, L_O, $\beta_i^{(1)}(V)$ and $\beta_i^{(2)}(V)$ are the hazy dataset, the signal component, the pure haze component, the radiance scattered by the atmosphere, the signal attenuation factor and the haze weighting in satellite band i, respectively. H_i can be expressed as:

$$H_i = \overline{H_i} + H_{i_v} \qquad \ldots (5.2)$$

Where $\overline{H_i}$ is the haze mean, which is assumed to be uniform within the image or sub-region of the image, and H_{i_v} is a zero-mean random variable corresponding to haze randomness. Hence:

$$\mathrm{Var}\left(H_{i_v}\right) = \mathrm{Var}\left(H_i\right) \qquad \ldots (5.3)$$

So Equation (5.1) can be written as:

$$L_i(V) = \left[1 - \beta_i^{(1)}(V)\right]T_i + L_O + \beta_i^{(2)}(V)\left[\overline{H_i} + H_{i_v}\right] \qquad \ldots (5.4)$$

In order to remove the haze effects, we need to remove both the weighted haze mean $\beta_i^{(2)}(V)\overline{H_i}$ and the varying component $\beta_i^{(2)}(V)H_{i_v}$ and deal with the signal attenuation factor $\beta_i^{(1)}(V)$.

From Chapter 4, the effects of $\beta_i^{(1)}(V)$ and L_O to classification accuracy are not significant (see Section 4.9), so we will not consider their removal throughout the analysis. We normally do not have prior knowledge about $\beta_i^{(2)}(V)\overline{H_i}$ therefore we need to estimate it from the hazy data itself. If the estimate is $\widehat{\beta_i^{(2)}(V)\overline{H_i}}$, subtracting it from $L_i(V)$ yields:

$$\widehat{L_{i_Z}(V)} = L_i(V) - \widehat{\beta_i^{(2)}(V)\overline{H_i}} = \left[1 - \beta_i^{(1)}(V)\right]T_i + L_O + \beta_i^{(2)}(V)\left[\overline{H_i} + H_{i_v}\right] - \widehat{\beta_i^{(2)}(V)\overline{H_i}} \qquad \ldots (5.5)$$

Equation (5.5) becomes:

$$\widehat{L_{i_Z}(V)} = \left[1-\beta_i^{(1)}(V)\right]T_i + L_O + \left[\beta_i^{(2)}(V)\overline{H_i} - \widehat{\beta_i^{(2)}(V)\overline{H_i}}\right] + \beta_i^{(2)}(V)H_{i_v} \qquad \text{... (5.6)}$$

where $\left[\beta_i^{(2)}(V)\overline{H_i} - \widehat{\beta_i^{(2)}(V)\overline{H_i}}\right]$ is the error associated with the difference between the ideal and estimated weighted haze mean. The haze randomness component $\beta_i^{(2)}(V)H_{i_v}$ can then be smoothed by applying a spatial filter:

$$\widehat{f_i(V)} = h\left(\widehat{L_{i_Z}(V)}\right) \qquad \text{... (5.7)}$$

where h is the filter function and $\hat{f}(V)$ is the restored data. Note that this also smoothes the signal component; we will show in Section 5.5 that filtering is only necessary for thick haze where the haze variability is much greater than the surface. For thin haze, the surface variability is much greater than the haze; filtering causes degradation to the surface and therefore is not required.

In this chapter, we consider three types of filter, viz. average, Gaussian and median (see Section 5.4.2). For the linear filters, such as the average and Gaussian filters, since $\beta_i^{(k)}(V)$ is assumed to be constant, we have:

$$\begin{aligned}\widehat{f_i(V)} &= h_{\text{linear}}\left(\widehat{L_{i_Z}(V)}\right)\\ &= \left[1-\beta_i^{(1)}(V)\right]h_{\text{linear}}(T_i) + h_{\text{linear}}\left(\left[\beta_i^{(2)}(V)\overline{H_i} - \widehat{\beta_i^{(2)}(V)\overline{H_i}}\right]\right) +\\ &\quad \beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right) + h_{\text{linear}}(L_O) \qquad \text{... (5.8)}\\ &= \left[1-\beta_i^{(1)}(V)\right]h_{\text{linear}}(T_i) + h_{\text{linear}}\left(\beta_i^{(2)}(V)\overline{H_i}\right) - h_{\text{linear}}\left(\widehat{\beta_i^{(2)}(V)\overline{H_i}}\right) +\\ &\quad \beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right) + L_O\end{aligned}$$

Linear filters are usually normalised to 1 i.e the sum of the filter coefficients is 1; since the haze mean $\overline{H_i}$ is assumed to be constant, hence:

$$\begin{aligned}\widehat{f_i(V)} = {} & \left[1-\beta_i^{(1)}(V)\right]h_{\text{linear}}(T_i)+\beta_i^{(2)}(V)\overline{H_i}-h_{\text{linear}}\left(\widehat{\beta_i^{(2)}(V)\overline{H_i}}\right)+ \\ & \beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right)+L_O\end{aligned} \qquad \text{... (5.9)}$$

The median filter is non-linear, so this separation is not possible:

$$\begin{aligned}\widehat{f_i(V)} &= h_{\text{Median}}\left(\widehat{L_{i_Z}(V)}\right) \\ &= \text{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i+\left[\beta_i^{(2)}(V)\overline{H_i}-\widehat{\beta_i^{(2)}(V)\overline{H_i}}\right]+\beta_i^{(2)}(V)H_{i_v}+L_O\right)\end{aligned} \qquad \text{... (5.10)}$$

For an ideal case where the haze mean is known exactly, $\beta_i^{(2)}(V)\overline{H_i}=\widehat{\beta_i^{(2)}(V)\overline{H_i}}$, so the degraded data after subtracting the haze mean becomes:

$$\widehat{L_{i_{Z(\text{ideal})}}(V)}=\left(1-\beta_i^{(1)}(V)\right)T_i+\beta_i^{(2)}(V)H_{i_v}+L_O \qquad \text{... (5.11)}$$

Consequently, when using average and Gaussian filters, we have:

$$\begin{aligned}\widehat{f_{i_{Z(\text{ideal})}}}(V) &= h_{\text{linear}}\left(\widehat{L_{i_{Z(\text{ideal})}}(V)}\right) \\ &= \left[1-\beta_i^{(1)}(V)\right]h_{\text{linear}}(T_i)+\beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right)+L_O\end{aligned} \qquad \text{... (5.12)}$$

but when using a median filter, the restored data becomes:

$$\begin{aligned}\widehat{f_{i_{Z(\text{ideal})}}}(V) &= h_{\text{Median}}\left(\widehat{L_{i_{Z(\text{ideal})}}(V)}\right) \\ &= \text{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i+\beta_i^{(2)}(V)H_{i_v}+L_O\right)\end{aligned} \qquad \text{... (5.13)}$$

From Equation (5.12), it is clear that a linear filter filters not only the haze randomness H_{i_v}, but also the surface information T_i. For thin haze (i.e. small $\beta_i^{(1)}(V)$ and $\beta_i^{(2)}(V)$), filtering will cause degradation to T_i. For thick haze, we have big $\beta_i^{(1)}(V)$ and $\beta_i^{(2)}(V)$; the effect of filtering to H_{i_v} is more significant than T_i. Although small, $\left[1-\beta_i^{(1)}(V)\right]h_{\text{linear}}(T_i)$, the structure of $(\omega-\mu_i)$ and C in T_i (Equation (3.8)) is still preserved and therefore will be useful for classification purpose (Section 5.5).

For non-linear filters such as median filtering (Equation (5.13)), the filtering affects the linear summation of the signal and haze components as a whole, i.e. $\text{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i+\beta_i^{(2)}(V)H_{i_v}+L_O\right)$. For thick haze, the input of median filtering is dominated by H_{i_v}; therefore, the effects of haze will be reduced to some extent. For thin haze, the input of the filtering dominated by T_i; therefore, degradation of surface occurs. Section 5.5 will exploit this and try to find the point where filtering starts to degrade the image, instead of improving it.

Based on this analysis, haze removal consists of (a) estimating the haze mean from hazy data using PIFs, (b) subtracting the haze mean from the data in order to remove the haze path radiance and (c) applying spatial filtering in order to reduce the haze randomness within the data (Section 5.4). To assess the performance of the haze removal, we use (a) a measure of signal-to-noise-ratio (SNR) and (b) classification accuracy (Section 5.5). The procedures for the haze removal and quality assessment are illustrated in a flowchart in Figure 5.4. At this stage, we consider only uniform haze within simulated datasets, we will discuss methods of dealing with the spatial variation of haze in Section 5.6, when applying the haze removal approach to real hazy datasets.

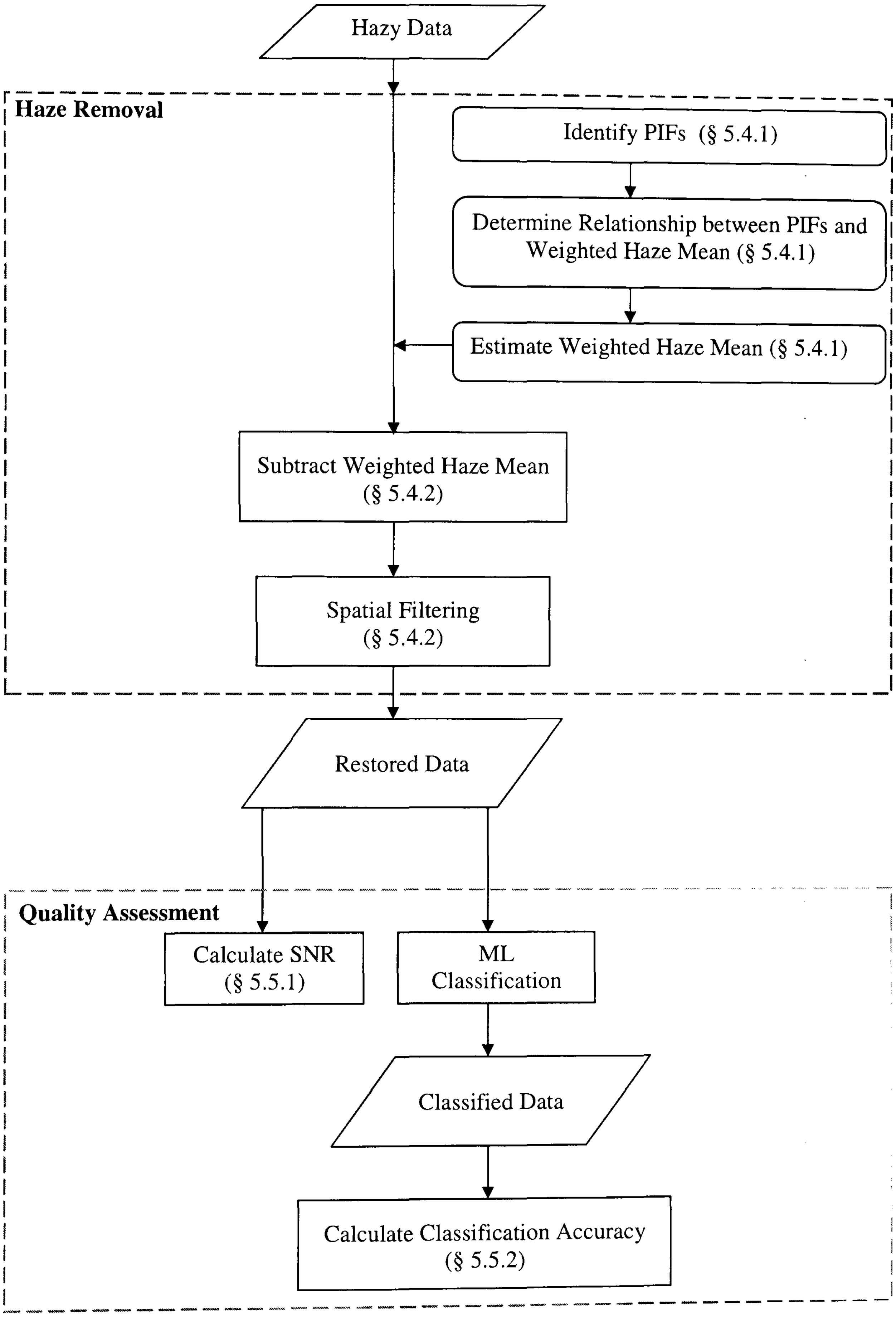

Hazy Data
Haze Removal
Identify PIFs (§ 5.4.1)
Determine Relationship between PIFs and Weighted Haze Mean (§ 5.4.1)
Estimate Weighted Haze Mean (§ 5.4.1)
Subtract Weighted Haze Mean (§ 5.4.2)
Spatial Filtering (§ 5.4.2)
Restored Data
Quality Assessment
Calculate SNR (§ 5.5.1)
ML Classification
Classified Data
Calculate Classification Accuracy (§ 5.5.2)

Figure 5.4: *Flowchart of haze removal and quality assessment procedures using simulated hazy datasets.*

5.4 Haze Removal

To test the haze removal procedures, we first make use of the simulated hazy datasets described in Chapter 4, for which the visibilities and values of $\beta_i^{(2)}(V)\overline{H_i}$ are known. The assumptions are:

(a) The haze is spatially uniform.

(b) The haziness within the data is mainly associated with additive effects due to scattering from particles; therefore, most of the efforts done here are to deal with the haze scattering term $\beta_i^{(2)}(V)H_i$.

(c) We have shown in Section 4.9 that multiplicative effects of haze (mainly due to absorption by smaller constituents) on land classification are not significant; therefore, we assume the effect of $\beta_i^{(1)}(V)$ is negligible.

5.4.1 Estimation of Haze Mean Radiance

In order to estimate $\beta_i^{(2)}(V)\overline{H_i}$, we first need to establish relationships between the exact $\beta_i^{(2)}(V)\overline{H_i}$ and the corresponding PIF radiances within the simulated hazy datasets for different levels of visibility.

In equatorial countries such as Malaysia, only a small amount of variation occurs in BRDF throughout a year, so their effects on the target radiance for different acquisition dates are assumed to be negligible. Variation in PIF spectral radiance from multi-date datasets is therefore assumed to be due to only atmospheric conditions (i.e. signal attenuation and haze scattering).

In our study, the study area (Klang, Selangor) is located within a flat region, near the west coast of Malaysia. The PIF pixels are chosen from rooftops of terrace houses, which are not too high and covers about 44% of Malaysian houses (Isa et al. 2010). The typical area of this type of houses is about 20 feet in wide and 70 feet in length, with built-up area of about 1200 square feet while the remaining area is garden. Most of these houses were built using clay bricks and have clay roof tiles (Isa et al. 2010). The houses are usually built in rows that are separated by roads made of tarmac. It is clear that most of the housing areas are covered with impervious surfaces and have very little vegetation; therefore are little affected by biological changes (Scott 1988). The study area is located in a sub urban region; such housing area is normally surrounded by distinct features such as rubber and oil palm plantation, so is quite easy to identify.

It is important to note that in order to minimised mixed pixel problem, the PIF should have at least a few pixels in size (Hadjimitsis et al. 2009) so that measurement made from a satellite IFOV (i.e. 30 m by 30 m for Landsat) will not fall out of the chosen features. In practice mixed pixel problem in PIF is unavoidable, however, this was minimised because the objects within a single PIF pixel are mostly impervious surfaces (clay bricks and tiles and tarmac). A schematic diagram on what is in a PIF pixel is illustrated in Figure 5.5.

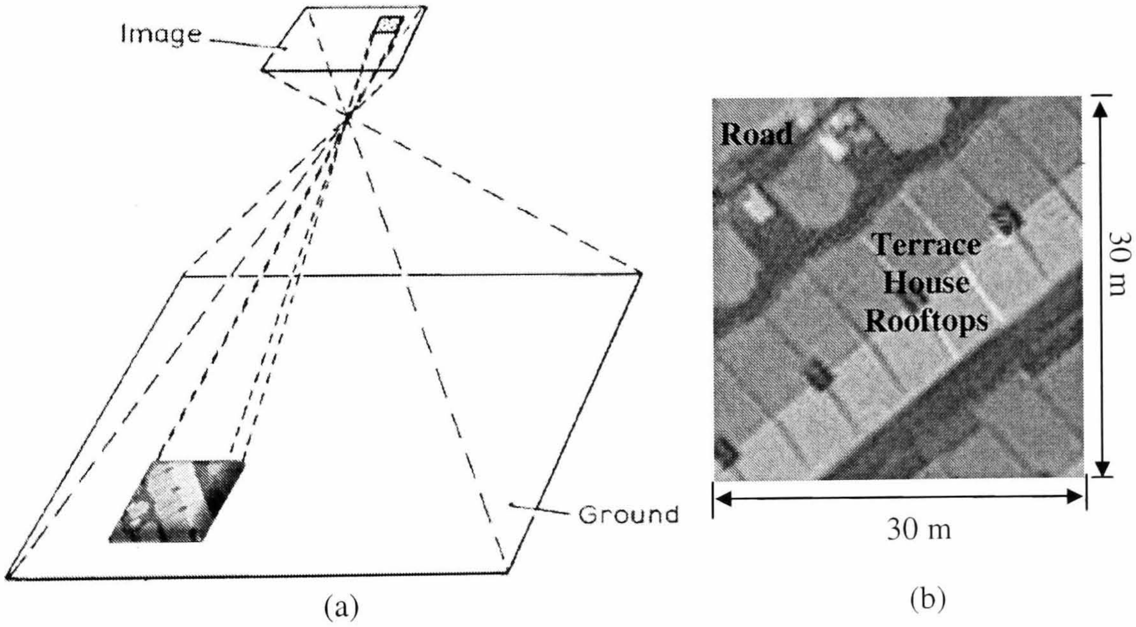
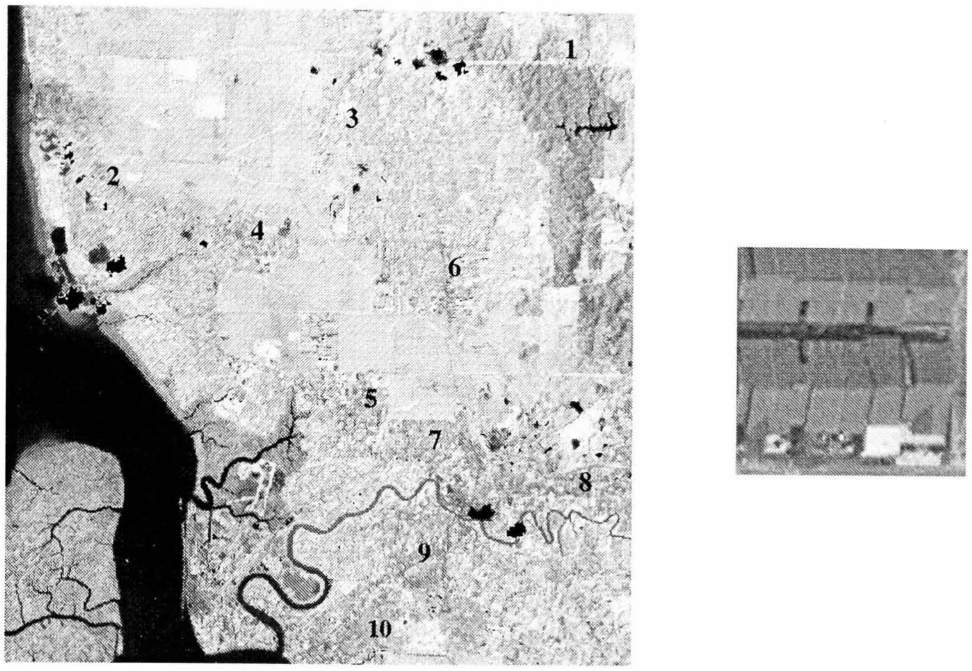


Figure 5.5: *(a) A schematic diagram of a single PIF pixe. (b) Close-up of the pixel in (a).*

For visibilities from 18 km to 2 km, ten PIFs that are distributed throughout the image were selected and their radiances were extracted. Figure 5.6 (a) shows Landsat bands 4, 5 and 3 assigned to red, green and blue channels from 11 February 1999 for Klang, in Selangor, Malaysia. The white box indicates the location of the PIFs which are indicated by the red squares. Figure 5.6 (b) shows the enlarged version of PIF number

7. The PIFs are selected from the rooftops of houses that have nearly constant radiances. It can be seen that the PIF consists of house rooftops and roads, with little vegetation.



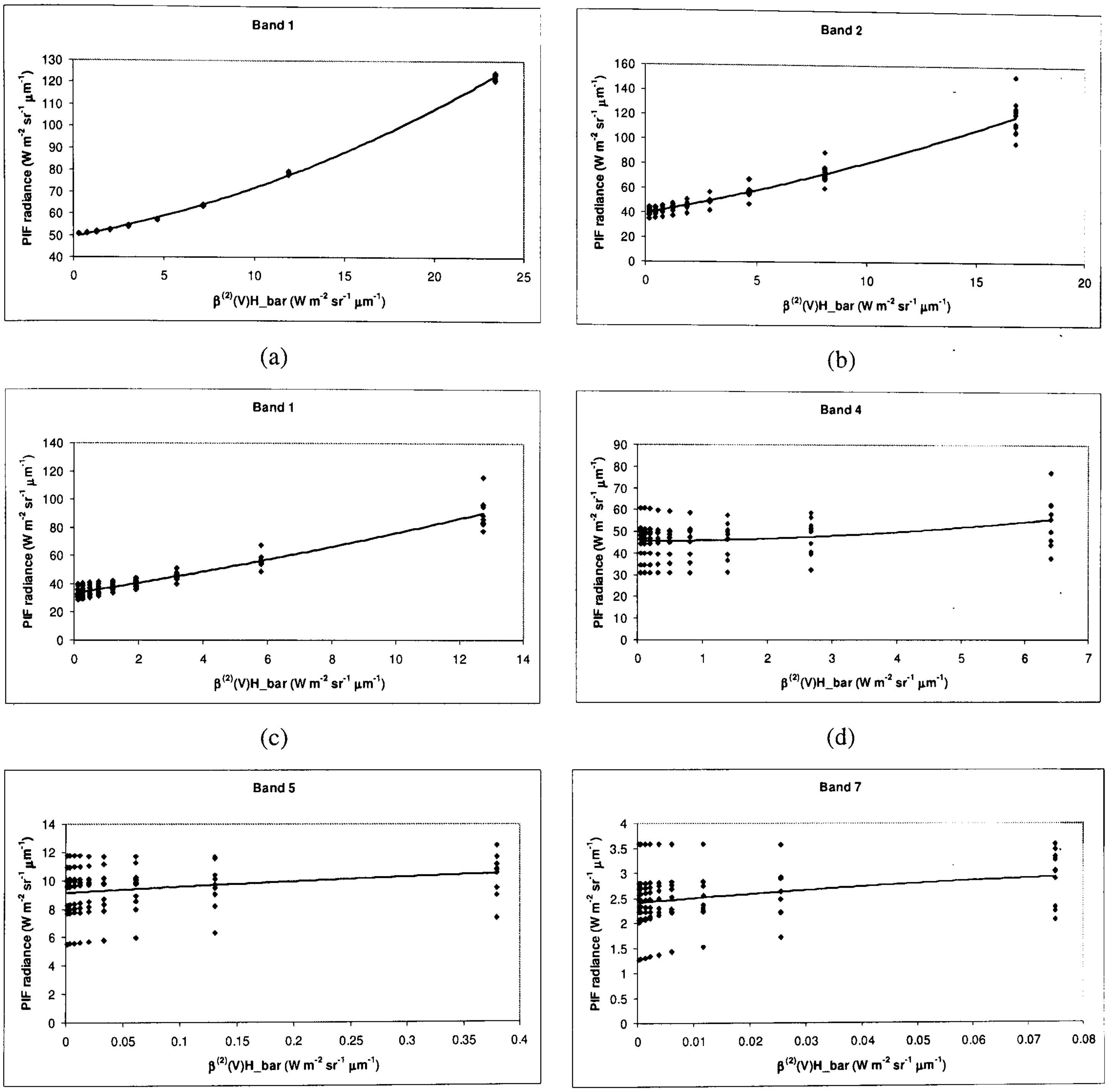
(a) (b)

Figure 5.6: *(a) Landsat bands 5, 4 and 3 assigned to red, green and blue channels from 11 February 1999 for Klang, in Selangor, Malaysia. (b) Enlarged version of PIF number 7 taken from Google Maps.*

Scatterplots of $\beta_i^{(2)}(V)\overline{H_i}$ versus the PIF radiance, for bands 1, 2, 3, 4, 5 and 7 are plotted in Figure 5.7. The PIF radiance values are indicated by '♦'. The PIF radiance increases steadily as $\beta_i^{(2)}(V)\overline{H_i}$ gets larger (i.e. haze gets thicker) due to the increasing atmospheric effects (haze scattering and signal attenuation).

Hence, by knowing the radiances of the PIF pixels, it is possible to predict the corresponding $\beta_i^{(2)}(V)\overline{H_i}$. In order to do so, we carried out regression between $\beta_i^{(2)}(V)\overline{H_i}$ and the PIF radiance. The solid curves in Figure 5.7 are the regression curves which represent the predicted $\beta_i^{(2)}(V)\overline{H_i}$, i.e. $\widehat{\beta_i^{(2)}(V)\overline{H_i}}$. It can be seen that the regression curves for all the bands have similar trends and therefore can be modelled by the same regression equation: $\widehat{\beta_i^{(2)}(V)\overline{H_1}} = a\left(L_{PIF_i}\right)^2 + bL_{PIF_i} + c$, where a, b, and c are the regression variables and L_{PIF_i} is the PIF radiance for band i. The regression variables, a, b, and c and the coefficient of determination, R^2, are given in Table 5.2.

Overall the R^2 values are greater than 0.9, indicating a good fit between the regression curve and the data in all the bands. The estimated weighted haze mean radiance $\widehat{\beta_i^{(2)}(V)\overline{H_i}}$ for visibilities 2 to 18 km, calculated using the regression equation, is given in Table 5.3. In this table, the ideal weighted haze mean radiance $\beta_i^{(2)}(V)\overline{H_i}$ is also given for comparison. Bands with shorter wavelengths possess larger L_{PIF_i} and $\widehat{\beta_i^{(2)}(V)\overline{H_i}}$ values due to the greater haze scattering than longer wavelengths. There is a sharper increase in L_{PIF_i} in bands with shorter wavelengths (bands 1, 2 and 3) compared to those with longer wavelengths (bands 4, 5 and 7), indicated that the former are affected by haze while the later almost not being affected by haze. It is clear that bands with shorter wavelengths have larger L_{PIF_i}, therefore brighter PIFs, than clear or less hazy image. Bands with longer wavelengths have a somewhat constant L_{PIF_i}, indicating that haze has almost not effects on the PIFs.



(a) (b)

(c) (d)

(e) (f)

Figure 5.7: *Regression analysis, of PIF radiance from the simulated hazy data against* $\beta_i^{(2)}(V)\overline{H_i}$ *, for (a) band 1, (b) band 2, (c) band 3, (d) band 4, (e) band 5 and (f) band 7. The corresponding regression equations are in Table 5.2.*

Table 5.2: *Regression model to predict* $\widehat{\beta_i^{(2)}(V)\overline{H_i}}$ *for bands 1, 2, 3, 4, 5, and 7;* L_{PIF} *is the PIF radiance.*

Band	Regression Model	R^2
1	$\widehat{\beta_1^{(2)}(V)\overline{H_1}} = -0.0023*\left(L_{PIF_1}\right)^2 + 0.7176*L_{PIF_1} - 29.397$	0.9956
2	$\widehat{\beta_2^{(2)}(V)\overline{H_2}} = -0.0011*\left(L_{PIF_2}\right)^2 + 0.3771*L_{PIF_2} - 13.103$	0.9539
3	$\widehat{\beta_3^{(2)}(V)\overline{H_3}} = -0.001*\left(L_{PIF_3}\right)^2 + 0.334*L_{PIF_3} - 9.8765$	0.9395
4	$\widehat{\beta_4^{(2)}(V)\overline{H_4}} = 0.0041*\left(L_{PIF_4}\right)^2 - 0.3157*L_{PIF_4} + 6.8741$	0.1827
5	$\widehat{\beta_5^{(2)}(V)\overline{H_5}} = 0.0046*\left(L_{PIF_5}\right)^2 - 0.0647*L_{PIF_5} + 0.2629$	0.0788
7	$\widehat{\beta_7^{(2)}(V)\overline{H_7}} = 0.0028*\left(L_{PIF_7}\right)^2 - 0.0026*L_{PIF_7} + 0.0021$	0.0826

Table 5.3: *Comparison between* $\beta_i^{(2)}(V)\overline{H_i}$ *(exact) and* $\widehat{\beta_i^{(2)}(V)\overline{H_i}}$ *(estimated).*

Visibility (km)	Weighted Haze Mean (W m^{-2} sr^{-1}µm^{-1})											
	Band 1 (0.49 µm)		Band 2 (0.56 µm)		Band 3 (0.66 µm)		Band 4 (0.83 µm)		Band 5 (1.67 µm)		Band 7 (2.24 µm)	
	$\beta_1^{(2)}(V)\overline{H_1}$	$\widehat{\beta_1^{(2)}(V)\overline{H_1}}$	$\beta_2^{(2)}(V)\overline{H_2}$	$\widehat{\beta_2^{(2)}(V)\overline{H_2}}$	$\beta_3^{(2)}(V)\overline{H_3}$	$\widehat{\beta_3^{(2)}(V)\overline{H_3}}$	$\beta_4^{(2)}(V)\overline{H_4}$	$\widehat{\beta_4^{(2)}(V)\overline{H_4}}$	$\beta_5^{(2)}(V)\overline{H_5}$	$\widehat{\beta_5^{(2)}(V)\overline{H_5}}$	$\beta_7^{(2)}(V)\overline{H_7}$	$\widehat{\beta_7^{(2)}(V)\overline{H_7}}$
2	23.4055	21.9717	16.8127	16.1936	12.7475	11.9776	6.4084	2.3435	0.3794	0.0896	0.0750	0.0213
4	11.9197	11.0431	8.0862	8.5720	5.7942	5.6754	2.6693	1.4052	0.1305	0.0713	0.0254	0.0173
6	7.1821	6.3763	4.6531	5.0435	3.1836	3.1571	1.3875	1.2908	0.0615	0.0667	0.0116	0.0161
8	4.6351	4.4832	2.9044	3.3441	1.9225	1.9730	0.8074	1.2678	0.0333	0.0647	0.0060	0.0156
10	3.0535	3.6924	1.8689	2.4140	1.2099	1.3165	0.4982	1.2685	0.0204	0.0636	0.0038	0.0153
12	2.0323	3.3734	1.2207	1.8537	0.7758	0.9081	0.3126	1.2755	0.0123	0.0629	0.0023	0.0151
14	1.3031	3.2689	0.7720	1.4922	0.4837	0.6338	0.1924	1.2838	0.0074	0.0623	0.0014	0.0149
16	0.7612	3.2674	0.4435	1.2460	0.2739	0.4393	0.1067	1.2919	0.0037	0.0620	0.0006	0.0149
18	0.3304	3.3163	0.1910	1.0708	0.1167	0.2946	0.0450	1.2993	0.0016	0.0617	0.0003	0.0147

5.4.2 Restoration of Surface Information Using Spatial Filtering

After subtracting the estimated haze mean component, the haze noise within the image is expected to behave as a zero-mean random variable associated with haze randomness, $\beta_i^{(2)}(V)H_{i_v}$ (see Equation (5.7)) (although errors in the haze mean estimate will cause a bias). If we assume the estimate of $\widehat{\beta_i^{(2)}(V)\overline{H_i}}$ is good enough that can be neglected, our concern now is to reduce $\beta_i^{(2)}(V)H_{i_v}$ by using spatial filtering. Here, three types of filtering are considered, i.e. average, median and Gaussian.

The Average Filter

The main advantages of average filtering are that it is simple, intuitive and easy to use, but still effective in reducing noise. Average filtering simply replaces each pixel value in an image with the average value of its neighbours, including itself. The average filter depends on the size of the window used, and the size can be increased to suit the severity of the haze. Here square averaging windows are used and, in order to determine the best window size for a specific visibility, we use window sizes from 3 x 3 to 21 x 21 and calculate the SNR of the resulting filtered data.

The Gaussian Filter

A continous Gaussian filter has the form:

$$h_g(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \quad \text{... (5.14)}$$

where x and y are distance from the origin in the horizontal and vertical direction respectively and σ is the standard deviation of the Gaussian filter. In order to preserve the mean energy in the digital case, the form shown in Equation (5.14) is normalised by the sum of the filter coefficients, to give:

$$h(r,s)=\frac{h_g(r,s)}{\sum_{r=-M/2}^{M/2}\sum_{s=-N/2}^{N/2}h_g(r,s)} \qquad \text{... (5.15)}$$

During filtering, the centre pixel receives the heaviest weight, and pixels receive smaller weights as the distance from the window centre increases. This study uses built-in Gaussian filters in the ENVI image processing software, in which σ is related to window size by $\sigma=\frac{M}{8}$ for an M×M window. Plots of the 1-dimensional weighting distribution for 3 x 3, 5 x 5, 7 x 7, 11 x 11 and 21 x 21 window sizes are shown in Figure 5.8 and examples of filter windows for 3 x 3, 5 x 5 and 21 x 21 are given in Figure 5.9. Note that the weighting of the centre location for a 3 x 3-window is 0.9, which implies that the filtered image is likely to be very similar to the original image. On the other hand, the 21 x 21-window gives much lower weighting across the filter (the highest is 0.02 and lowest is nearly 0) so is likely to resemble an average filter.

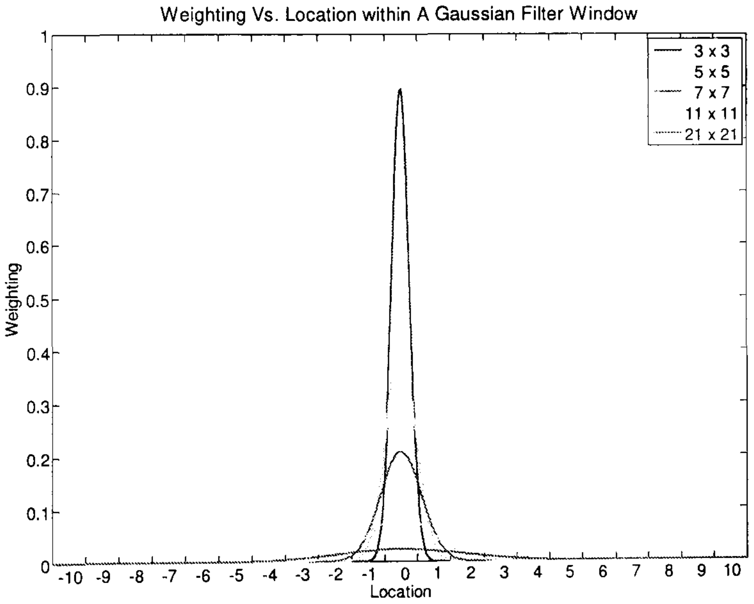


Figure 5.8: *Distribution of pixel weighting for 3 by 3, 5 by 5, 7 by 7, 11 by 11 and 21 by 21 window sizes, of a Gaussian filter in 1-dimension.*

0.00073	0.025561	0.00073
0.025561	0.894834	0.025561
0.00073	0.025561	0.00073

(a)

0.000015	0.000676	0.002431	0.000676	0.000015
0.000676	0.031441	0.113083	0.031441	0.000676
0.002431	0.113083	0.406718	0.113083	0.002431
0.000676	0.031441	0.113083	0.031441	0.000676
0.000015	0.000676	0.002431	0.000676	0.000015

(b)

0.016071	0.017281	0.016071
0.019979	0.021483	0.019979
0.021483	0.0231	0.021483
0.019979	0.021483	0.019979
0.016071	0.017281	0.016071

(c)

Figure 5.9: *Gaussian filter (a) 3 by 3, (b)5 by 5 and (c) 21 by 21 but only part of centre cells.*

The Median Filter

Median filtering is often used to remove noise from a degraded image and at the same time preserve edges. It replaces the central pixel with the median value in the window. A similar approach to that for average and Gaussian filtering is used to determine the best window size for a specific visibility.

5.5 Quality Assessment of Restored Data

A common way to measure the accuracy of restored data is to compare its quality with uncorrupted data. Visual analysis offers a fast and simple way to do this, but suffers from possible analyst bias. Hence we develop here two quantitative approaches.

5.5.1 SNR

One measure of performance for single band data is the signal-to-noise ratio (SNR), which quantifies how severely data have been degraded by noise. SNR is defined as the ratio between the squared ratio of signal amplitude and noise amplitude:

$$\mathrm{SNR} = \left(\frac{\mathrm{A_S}}{\mathrm{A_N}} \right)^2 \qquad \ldots (5.16)$$

where P_S and A_S are signal power and amplitude respectively, and similarly for noise. SNR also can be measured on a decibel scale (dB):

$$\mathrm{SNR(dB)} = 10\log_{10}(\mathrm{SNR}) = 10\log_{10}\left(\frac{P_S}{P_N}\right) = 20\log_{10}\left(\frac{A_S}{A_N}\right) \qquad \ldots (5.17)$$

The expression for SNR and its estimates vary between: (a) original hazy data (with nonzero-mean noise), (b) hazy data after subtracting the haze mean and (c) restored data (after filtering).

From Equation (5.1), the SNR of hazy data with nonzero-mean haze noise can be expressed as:

$$\begin{aligned} \mathrm{SNR} &= \frac{\left\langle \left\{ \left[1-\beta_i^{(1)}(V)\right] T_i + L_O \right\}^2 \right\rangle}{\left\langle \beta_i^{(2)2}(V) H_i{}^2 \right\rangle} \\ &= \frac{\left\langle \left[1-\beta_i^{(1)}(V)\right]^2 T_i{}^2 + 2\left[1-\beta_i^{(1)}(V)\right] T_i L_O + L_O{}^2 \right\rangle}{\left\langle \beta_i^{(2)2}(V) H_i{}^2 \right\rangle} \\ &= \frac{\left[1-\beta_i^{(1)}(V)\right]^2 \left\langle T_i{}^2 \right\rangle + 2L_O\left[1-\beta_i^{(1)}(V)\right]\left\langle T_i \right\rangle + L_O{}^2}{\left\langle \beta_i^{(2)}(V)^2 \left(\overline{H}_i + H_{i_v}\right)^2 \right\rangle} \\ &= \frac{\left[1-\beta_i^{(1)}(V)\right]^2 \left\langle T_i{}^2 \right\rangle + 2L_O\left[1-\beta_i^{(1)}(V)\right]\left\langle T_i \right\rangle + L_O{}^2}{\beta_i^{(2)}(V)^2 \left[\overline{H}_i{}^2 + \mathrm{Var}\left(H_{i_v}\right)\right]} \end{aligned} \qquad \ldots (5.18)$$

since by assumption $\beta_i^{(1)}(V)$ and $\beta_i^{(2)}(V)$ are the same for all pixels in the scene. Note that here we assume $\left[1-\beta_i^{(1)}(V)\right]T_i$ and L_O from the hazy data to be the signal amplitude because the effects of $\left[1-\beta_i^{(1)}(V)\right]$ to classification is negligible (see Section 4.9); this applies for all cases. Due to the descrete properties of the hazy data, the exact values are replaced by their estimates:

$$\widehat{\mathrm{SNR}} = \frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i + L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\beta_i^{(2)}(V)^2\left(\overline{H}_i + H_{i_v}\right)^2} \qquad \dots (5.19)$$

where Q_m and Q_n are the numbers of pixels in the rows and columns of the image respectively. Note that such calculation is only possible if the values of T_i, $\overline{H}_i$, H_{i_v}, $\beta_i^{(1)}(V)$, $\beta_i^{(2)}(V)$, Q_m and Q_n are known apriori (e.g. simulated dataset). From Equation (5.6), the exact SNR of degraded data after subtraction of the weighted haze mean can be expressed as:

$$\mathrm{SNR} = \frac{\left\langle\left\{\left[1-\beta_i^{(1)}(V)\right]T_i + L_O\right\}^2\right\rangle}{\left\langle\left\{\left[\beta_i^{(2)}(V)\overline{H}_i - \widehat{\beta_i^{(2)}(V)\overline{H}_i}\right] + \beta_i^{(2)}(V)H_{i_v}\right\}^2\right\rangle} \qquad \dots (5.20)$$

and can be estimated by:

$$\widehat{\mathrm{SNR}} = \frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i + L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[\beta_i^{(2)}(V)\overline{H}_i - \widehat{\beta_i^{(2)}(V)\overline{H}_i}\right] + \beta_i^{(2)}(V)H_{i_v}\right\}^2} \qquad \dots (5.21)$$

Subsequently, the degraded data undergo spatial filtering. From Equation (5.9), for linear filtering, the exact SNR of restored data can be expressed as:

$$\mathrm{SNR}=\frac{\left\langle\left\{\left[1-\beta_i^{(1)}(V)\right]T_i+L_O\right\}^2\right\rangle}{\left\langle\left[\hat{f}(V)-\left(1-\beta_i^{(1)}(V)\right)T_i+L_O\right]^2\right\rangle}$$

$$=\frac{\left\langle\left\{\left[1-\beta_i^{(1)}(V)\right]T_i+L_O\right\}^2\right\rangle}{\left\langle\begin{bmatrix}\left(1-\beta^{(1)}(V)\right)h_{\text{linear}}(T_i)+h_{\text{linear}}\left(\left[\beta_i^{(2)}(V)\overline{H_i}-\widehat{\beta_i^{(2)}(V)\overline{H_i}}\right]\right)+\\ \beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right)+L_O-\left(1-\beta_i^{(1)}(V)\right)T_i-L_O\end{bmatrix}^2\right\rangle}$$

$$=\frac{\left\langle\left\{\left[1-\beta_i^{(1)}(V)\right]T_i+L_O\right\}^2\right\rangle}{\left\langle\begin{Bmatrix}\left(1-\beta_i^{(1)}(V)\right)\left[h_{\text{linear}}(T_i)-T_i\right]+h_{\text{linear}}\left(\left[\beta_i^{(2)}(V)\overline{H_i}-\widehat{\beta_i^{(2)}(V)\overline{H_i}}\right]\right)+\\ \beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right)\end{Bmatrix}^2\right\rangle}\quad\ldots(5.22)$$

and can be estimated by:

$$\widehat{\mathrm{SNR}}=\frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i+L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\begin{Bmatrix}\left(1-\beta_i^{(1)}(V)\right)\left[h_{\text{linear}}(T_i)-T_i\right]+\\ h_{\text{linear}}\left(\left[\beta_i^{(2)}(V)\overline{H_i}-\widehat{\beta_i^{(2)}(V)\overline{H_i}}\right]\right)+\beta_i^{(2)}(V)h_{\text{linear}}(H_v)\end{Bmatrix}^2}\quad\ldots(5.23)$$

For median filtering (see Equation (5.10)), the exact SNR can be expressed as:

$$\mathrm{SNR} = \frac{\left\langle \left\{ \left[1-\beta_i^{(1)}(V)\right]T_i + L_O \right\}^2 \right\rangle}{\left\langle \left[\hat{f}(V) - \left(1-\beta_i^{(1)}(V)\right)T_i\right]^2 \right\rangle}$$

$$= \frac{\left\langle \left\{ \left[1-\beta_i^{(1)}(V)\right]T_i + L_O \right\}^2 \right\rangle}{\left\langle \left\{ \begin{array}{l} \mathrm{Median}\left(\begin{array}{l} \left[1-\beta_i^{(1)}(V)\right]T_i + \left[\beta_i^{(2)}(V)\overline{H_i} - \widehat{\beta_i^{(2)}(V)\overline{H_i}}\right] + \\ \beta_i^{(2)}(V)H_{i_v} + L_O \end{array} \right) - \\ \qquad \left[1-\beta_i^{(1)}(V)\right]T_i - L_O \end{array} \right\}^2 \right\rangle} \qquad \ldots (5.24)$$

and its estimate by:

$$\widehat{\mathrm{SNR}} = \frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n} \left\{ \left[1-\beta_i^{(1)}(V)\right]T_i + L_O \right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n} \left\{ \begin{array}{l} \mathrm{Median}\left(\begin{array}{l} \left[1-\beta_i^{(1)}(V)\right]T_i + \left[\beta_i^{(2)}(V)\overline{H_i} - \widehat{\beta_i^{(2)}(V)\overline{H_i}}\right] + \\ \beta_i^{(2)}(V)H_{i_v} + L_O \end{array} \right) - \\ \left[1-\beta_i^{(1)}(V)\right]T_i - L_O \end{array} \right\}^2} \quad .. (5.25)$$

$\widehat{\mathrm{SNR}}$ can also be expressed in dB by taking 10 $\log_{10}$ of Equation (5.19), (5.21), (5.23) and (5.25).

The SNR of Restored Data when the Haze Mean is Known Exactly

When the haze mean is known exactly, $\left[\beta_i^{(2)}(V)\overline{H_i} - \widehat{\beta_i^{(2)}(V)\overline{H_i}}\right] = 0$ and therefore can be eliminated (see Equation (5.21)). Hence the SNR after subtraction of the haze mean is:

$$\widehat{\mathrm{SNR}} = \frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i + L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\beta_i^{(2)}(V)^2 H_{i_v}^{\ 2}} \qquad \text{... (5.26)}$$

For linear filtering (see Equation (5.23)) we have:

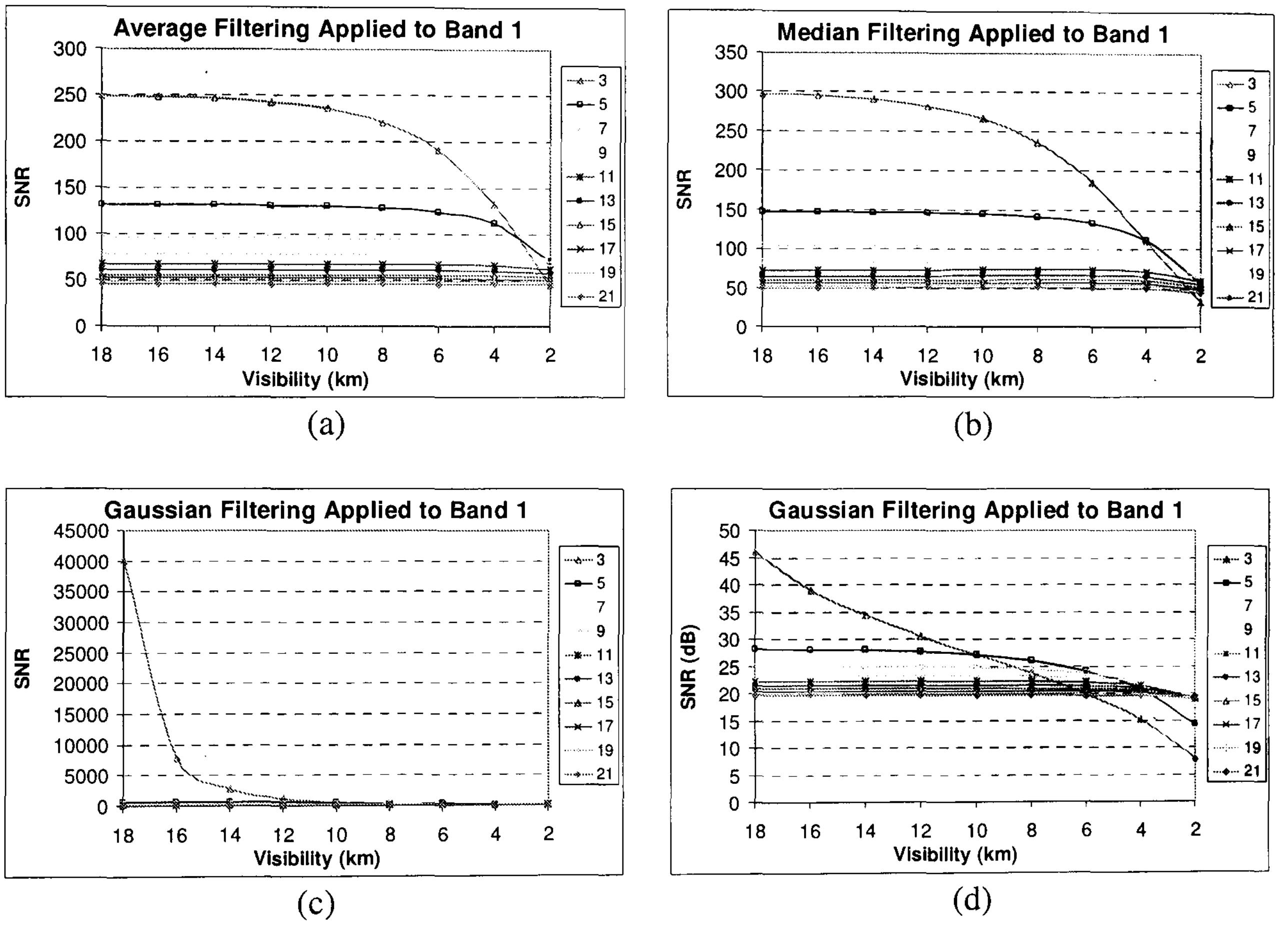
$$\widehat{\mathrm{SNR}} = \frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i + L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]\left[h_{\mathrm{linear}}(T_i) - T_i\right] + \beta_i^{(2)}(V)\,h_{\mathrm{linear}}\left(H_{i_v}\right)\right\}^2} \qquad \text{.. (5.27)}$$

For median filtering (see Equation (5.25)) we have:

$$\widehat{\mathrm{SNR}} = \frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i + L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\begin{matrix}\mathrm{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i + \beta_i^{(2)}(V)H_{i_v} + L_O\right) - \\ \left[1-\beta_i^{(1)}(V)\right]T_i - L_O\end{matrix}\right\}^2} \qquad \text{... (5.28)}$$

Calculation of SNR

In this section, we calculate the SNR of the data after weighted mean subtraction and filtering for the case when the haze mean is known exactly. The SNR calculations for bands 1 are given first and the explanation is given after that. These are then followed by the SNR calculations for bands 2, 3, 4, 5 and 7. This makes use of the simulated dataset for visibilities 2 km to 18 km described in Chapter 4.



(a)

(b)

(c)

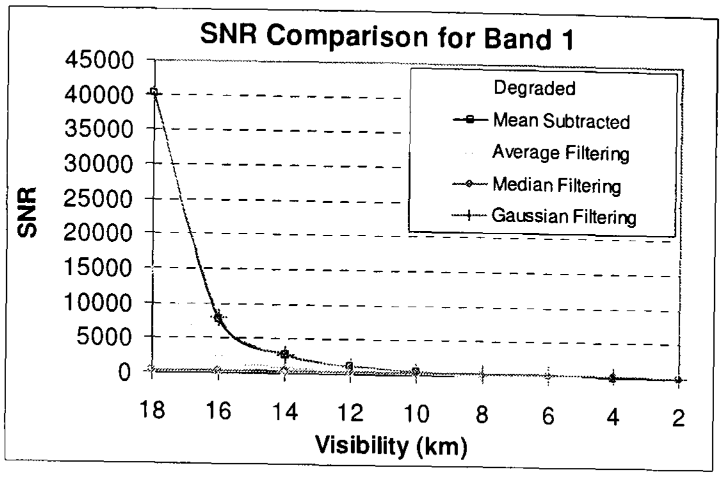
(d)

Figure 5.10: *SNR for Band 1 after applying (a) average filtering, (b) median filtering, (c) Gaussian filtering and (d) Same as (c) but in dB.*

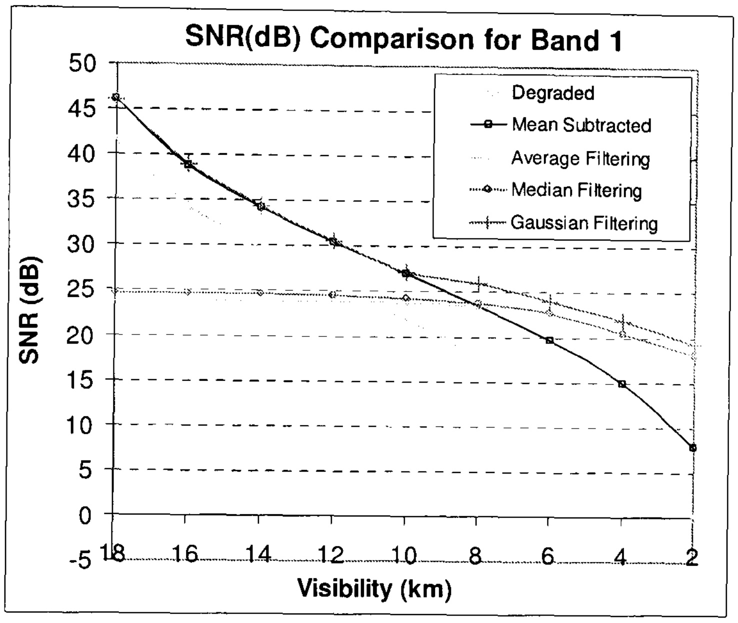
Figure 5.10 shows the SNR for band 1 with the exact mean removed, after applying average, median and Gaussian filtering. These plots help to determine the window size that produces the highest SNR at a particular visibility. For average and median filtering (Figure 5.10(a and b)), for smaller window sizes, the drop in SNR gets more rapid as the visibility reduces, but for bigger sizes, the SNR is nearly constant for all visibilities. For longer and moderate visibilities, 3 x 3 windows give the highest SNR, but the SNR drops when the window size is increased. For very short visibilities, bigger windows produce higher SNRs.

For Gaussian filtering (Figure 5.10(c)), the 3 x 3 window shows a sharp decrease in SNR for long visibilities, but then a slow decline for moderate visibilities. A big difference in SNR is observed between the 3 x 3 window and the rest of the windows, particularly for long visibilities. The larger-sized windows show a relatively flat trend towards shorter visibilities. The separation of the effect of window sizes is much better in the dB plot (Figure 5.10(d)). It can be seen that, for longer visibilities, smaller windows show higher SNR than bigger windows, while for shorter visibilities, the bigger windows exhibit higher SNRs, but the separation between windows is relatively narrow.

For all types of filtering, the highest SNR for a particular visibility (associated with the corresponding optimal window size in Table 5.4) is plotted in Figure 5.11(a). The SNR for Gaussian filtered data is very close to weighted-mean subtracted data and noticeably improves the original degraded data at shorter visibilities. The dB plot in Figure 5.11(b) provides a better separation for all types of filtering, where Gaussian filtering shows the best SNR for all visibilities. The changes in trend in the middle of the Gaussian filtering curve is due to a transition of the corresponding window sizes, i.e. the window size changes from 3 x 3 (at 10 km visibility) to 5 x 5 (at 8 km visibility). The improvement made by the Gaussian filtering with respect to weighted-mean subtracted data and degraded data curves is likely to increase as visibility reduces. The average and median filtering show a lower SNR than the degraded and weighted-mean subtracted data, for longer visibilities, indicating that the quality of the data becomes poorer after filtering compared to before filtering. However, the SNR of the average-filtered and the median-filtered data is better than the degraded and mean subtracted data for shorter visibilities.



(a)



(b)

Figure 5.11: *Comparison of filter performances for band 1.*

Table 5.4: *Optimal window sizes for band 1.*

Visibility	Filter Types / Sizes		
(km)	Average	Median	Gaussian
2	7	7	15
4	3	5	7
6	3	3	5
8	3	3	5
10	3	3	3
12	3	3	3
14	3	3	3
16	3	3	3
18	3	3	3

Explanation of the SNR Results

In order for the filtered data to have higher SNR than the mean subtracted data, the denominator in Equations (5.27) and (5.28) should be smaller than that of (5.26).

From (5.26) and (5.27), the denominator difference is:

$$\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[\beta_i^{(2)}(V)\right]^2 H_{i_v}{}^2-\left\{\left(1-\beta_i^{(1)}(V)\right)\left[h_{\text{linear}}(T_i)-T_i\right]+\beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right)\right\}^2\right\}$$

$$\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[\beta_i^{(2)}(V)\right]^2 H_{i_v}{}^2-\left[\begin{array}{l}\left(1-\beta_i^{(1)}(V)\right)^2\left[h_{\text{linear}}(T_i)-T_i\right]^2+\\ 2\left(1-\beta_i^{(1)}(V)\right)\left[h_{\text{linear}}(T_i)-T_i\right]\beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right)+\\ \left[\beta_i^{(2)}(V)\right]^2\left[h_{\text{linear}}\left(H_{i_v}\right)\right]^2\end{array}\right]\right\}$$

$$\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\begin{array}{l}-\left(1-\beta_i^{(1)}(V)\right)^2\left[h_{\text{linear}}(T_i)-T_i\right]^2-2\left(1-\beta_i^{(1)}(V)\right)\left[h_{\text{linear}}(T_i)-T_i\right]\beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right)+\\ \left[\beta_i^{(2)}(V)\right]^2\left[H_{i_v}{}^2-\left[h_{\text{linear}}\left(H_{i_v}\right)\right]^2\right]\end{array}\right\}$$

$$\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\underbrace{-\left(1-\beta_i^{(1)}(V)\right)^2\left[h_{\text{linear}}(T_i)-T_i\right]^2}_{A}+\underbrace{\left\{\begin{array}{l}-2\left(1-\beta_i^{(1)}(V)\right)\left[h_{\text{linear}}(T_i)-T_i\right]\beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_v}\right)+\\ \left[\beta_i^{(2)}(V)\right]^2\left[H_{i_v}{}^2-\left[h_{\text{linear}}\left(H_{i_v}\right)\right]^2\right]\end{array}\right\}}_{B}\right\}$$

For the denominator in (5.27) to be smaller than the denominator in (5.26), they must be positive. This means the term B should be larger than A; it seems that this is possible if $H_{i_v}{}^2>\left[h_{\text{linear}}\left(H_{i_v}\right)\right]^2$ and $\left(h_{\text{linear}}(T_i)-T_i\right)\approx 0$. However, if the term B is smaller than A, the SNR of the linear filtered data will be smaller than the SNR after subtraction of the haze mean.

Similarly, for the median filtering, the denominator in (5.28) should be less than that of (5.26), in order for the filtered data to have larger SNR compared to before filtering. However, this is not easy to predict because separation of $\text{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i+\beta_i^{(2)}(V)H_{i_v}+L_O\right)-\left[1-\beta_i^{(1)}(V)\right]T_i-L_O$ is not possible.

Here we carry out detail analysis on Equation (5.27) and (5.28) for extreme cases, i.e. very thin haze (good visibilities) and very severe haze. When there is good visibility, the term $\beta^{(2)}(V)h_{\text{linear}}(H_V)$ is very small (Figure 5.13(left)), therefore its contribution in Equation (5.27) is very small and can be ignored. Consequently, we have:

$$\widehat{SNR} = \frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i+L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]\left[h_{\text{linear}}(T_i)-T_i\right]\right\}^2} \qquad \text{... (5.29)}$$

Equation (5.29) indicates that the $\widehat{SNR}$ depends only on the scene itself. For average filtering, at good visibilities, the filter significantly reduces the variability within the scene. Therefore $\left[h_{\text{linear}}(T_i)-T_i\right]^2$ tends to be bigger than $\left[\beta_i^{(2)}(V)H_{i_v}\right]^2$ in Equation (5.26) and $\beta_i^{(2)}(V)^2\left(\overline{H}_i+H_{i_v}\right)^2$ in Equation (5.19); consequently, the SNR for the average filtered data tends to be lower than that of the mean subtracted data and original degraded data respectively.

For Gaussian filtering using a 3 x 3 window, since the weight of the centre window is 0.9 (see Figure 5.9), the filtering hardly alters the original pixel, therefore $h_{\text{linear}}(T_i)$ is almost equal to T_i, consequently $\left[h_{\text{linear}}(T_i)-T_i\right]^2$ is very small and almost zero. This explains why at good visibility, the SNR of Gaussian filtered data is higher than the average filtered data.

For median filtering, at good visibilities, $\beta_i^{(2)}(V)H_{i_v}$ is very small compared to $\left(1-\beta_i^{(1)}(V)\right)T_i$ and can be neglected, hence Equation (5.28) becomes:

$$\widehat{\mathrm{SNR}} = \frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i + L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\mathrm{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i + L_O\right)-\left[1-\beta_i^{(1)}(V)\right]T_i - L_O\right\}^2} \qquad \text{... (5.30)}$$

Due to the non-uniformity of the signal in the data (mainly caused by variability in land features), $\left\{\mathrm{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i + L_O\right)-\left[1-\beta_i^{(1)}(V)\right]T_i - L_O\right\}^2$ tend to be bigger than $\left[\beta_i^{(2)}(V)H_{i_v}\right]^2$ in Equation (5.26) and $\left[\beta_i^{(2)}(V)\left(\overline{H_i}+H_{i_v}\right)\right]^2$ in Equation (5.19), consequently, the SNR of median filtered data is smaller compared to the mean subtracted data (Figure 5.11(b)).

The results for all three filters suggest that for good visibilities, it is better not to filter the data at all, because the filtering will either decrease (as in the case of average and median filters) or give about the same SNR (as for Gaussian filtering). Visibilities considered good for bands 1, 2, 3, 4, 5 and 7 are given in Table 5.5. The key point is that it seems to be the scene is vary variable at pixel scale, so the extra variability introduced by haze noise is not detectable by the filter, unless the scene is very hazy, at which the haze variability is greater than the scene.

Table 5.5: *Visibility ranges at which filtering is not required.*

Band	Visibility (km)		
	Average	Median	Gaussian
1	> 8	> 8	> 10
2	> 12	> 12	> 14
3	> 10	> 10	> 12
4	> 8	> 8	> 10
5	> 8	> 8	> 12
7	> 10	> 10	> 12

Figure 5.12 shows (a) Hazy data, $L_i(V)$ (b) horizontal radiance profile for $L_i(V)$ and (c) horizontal radiance profile for $\beta_i^{(2)}(V)(H_i)$ associated with 18 km, 8 km and 2 km visibility in band 1. The vertical lines in (b) and (c) represent the cut along the horizontal line in (a). $\beta_i^{(2)}(V)(H_i)$ is obtained from the corresponding haze layers developed in Chapter 4. It can be seen that at 18 km visibility, since $\beta_i^{(2)}(V)(H_i)$ is very small and almost not variable, the variation in the $L_i(V)$ is caused mainly by the scene itself, $\left[1-\beta_i^{(1)}(V)\right]T_i$ while at 2 km visibility, the variation in the $L_i(V)$ is dominated by the haze, $\beta_i^{(2)}(V)(H_i)$.

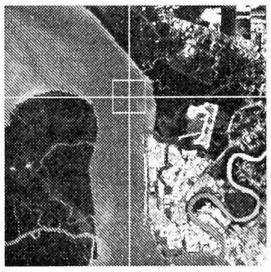
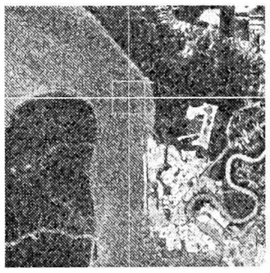
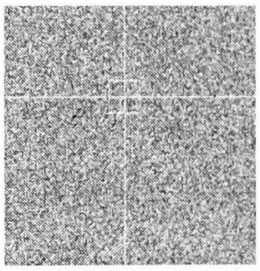
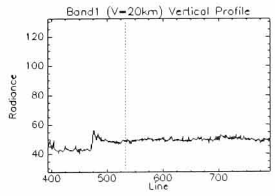
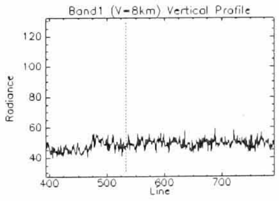
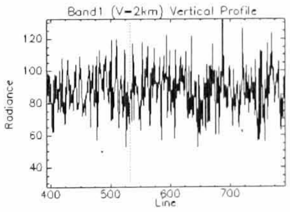
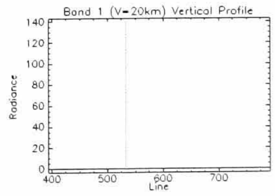
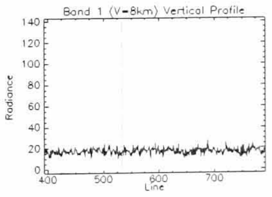
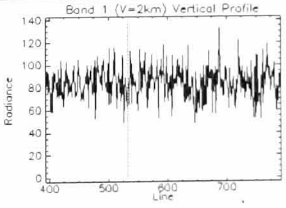
		18 km visibility (little haze)	8 km visibility (moderate haze)	2 km visibility (severe haze)
(a)	Hazy data, $L_i(V)$			
(b)	Horizontal profile of $L_i(V)$ $=\left[1-\beta_i^{(1)}(V)\right]T_i + L_O$ $+\beta_i^{(2)}(V)H_i$			
(c)	Horizontal profile of $\beta_i^{(2)}(V)(H_i)$ $=\beta_i^{(2)}(V)\left(\overline{H_i}+H_{i_v}\right)$			

Figure 5.13: *(a) Hazy data,* $L(V)$ *(b) horizontal profile for* $L_i(V)$ *and (c) horizontal profile for* $\beta_i^{(2)}(V)(H_i)$ *associated with 18 km, 8 km and 2 km visibility in band 1.*

For linear filtering, when the haze is very severe (i.e. short visibilities), $\beta_i^{(2)}(V)h_{\text{linear}}(H_{i_v})$ will tend to be very variable (Figure 5.13 (right)) and $\left[1-\beta_i^{(1)}(V)\right]\left[h_{\text{linear}}(T_i)-T_i\right]$ in Equation (5.27) is very small because of the strong signal attenuation ($\beta_i^{(1)}(V)\approx 1$) and so can be ignored; hence Equation (5.27) becomes:

$$\widehat{SNR}=\frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i+L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left[\beta_i^{(2)}(V)h_{\text{linear}}(H_{i_v})\right]^2} \qquad \text{... (5.31)}$$

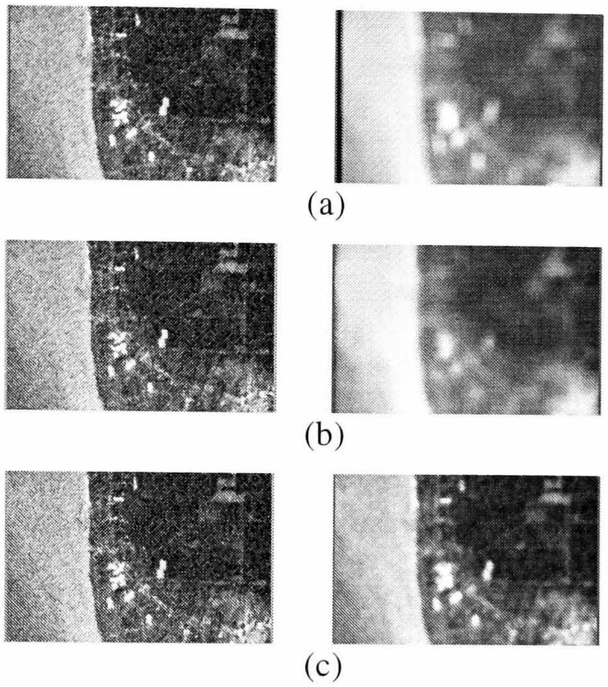
Because of the very severe haze and the effect of averaging, here $h_{\text{linear}}^2(H_{i_v})$ tends to be smaller than ${H_{i_v}}^2$ in Equation (5.26) and $(\overline{H_i}+H_{i_v})^2$ in Equation (5.19), therefore the average and Gaussian filtering are likely to have higher SNR than the mean subtracted data and original degraded data (Figure 5.11(b)). For median filtering, $\left[1-\beta_i^{(1)}(V)\right]T_i$ in the denominator of Equation (5.28) is very small compared to $\beta_i^{(2)}(V)H_{i_v}$ and so can be neglected. Hence we have:

$$\widehat{SNR}=\frac{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left\{\left[1-\beta_i^{(1)}(V)\right]T_i+L_O\right\}^2}{\sum_{m=1}^{Q_m}\sum_{n=1}^{Q_n}\left[\text{Median}\left(\beta_i^{(2)}(V)H_{i_v}+L_O\right)-L_O\right]^2} \qquad \text{... (5.32)}$$

Due to the very severe haze, $\left[\text{Median}\left(\beta_i^{(2)}(V)H_{i_v}\right)\right]^2$ tends to be less than $\left[\beta_i^{(2)}(V)H_{i_v}\right]^2$ in Equation (5.26) and $\left[\beta_i^{(2)}(V)\left(\overline{H_i}+H_{i_v}\right)\right]^2$ in Equation (5.19). This is due to the removal of extreme values by the median filter. Consequently, the SNR of the median filtered data is likely to be higher than the mean subtracted and original degraded data (Figure 5.11).

For linear filtering, for moderate haze, H_{i_V} in Equation (5.27) is more variable than for little haze. An optimal SNR can be achieved by keeping the denominator in Equation (5.27) low. In order to do so, the window size needs to be increased to effectively reduce variation in H_V, but, at the same time, not to cause significant increase in $\left[h_{linear}(T_i)-T_i\right]$. This explains why the optimal window size of the average and Gaussian filters needs to be increased as the visibility reduces (Table 5.4). The larger the window, the more effectively the variation in H_{i_V} will be reduced, but at some points, this may also cause $\left[h_{linear}(T_i)-T_i\right]$ to increase, causing the SNR to drop below the optimal value. An example of this can be seen when increasing the window size of an average and Gaussian filter for 12 km visibility data (see Figure 5.10(a and d)). This effect is also apparent from visual analysis of average and Gaussian filtering (Figure 5.14(a) and (c) respectively) 12 km visibility data using 3 x 3 and 21 x 21 windows (Figure 5.14 (left) and (right) respectively).

For median filtering, as H_V gets higher, the denominator in Equation (5.28) can be kept low by ensuring the difference between $\text{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i+\beta_i^{(2)}(V)H_{i_V}+L_O\right)$ and $\left[1-\beta_i^{(1)}(V)\right]T_i-L_O$ is not significant; this is possible by using a larger window. This explains why the optimal window size for median filter needs to be increased as the haze gets more severe (see Figure 5.10(b)). However, if the window size is larger than it should be, there is a possibility that $\text{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i+\beta_i^{(2)}(V)H_{i_V}+L_O\right)$ will differ greatly from $\left[1-\beta_i^{(1)}(V)\right]T_i-L_O$, so the corresponding SNR will drop below the optimal SNR. The visual effect of median filtering 12 km visibility data using 3 x 3 and 21 x 21 windows is shown in Figure 5.14(b).



(a)

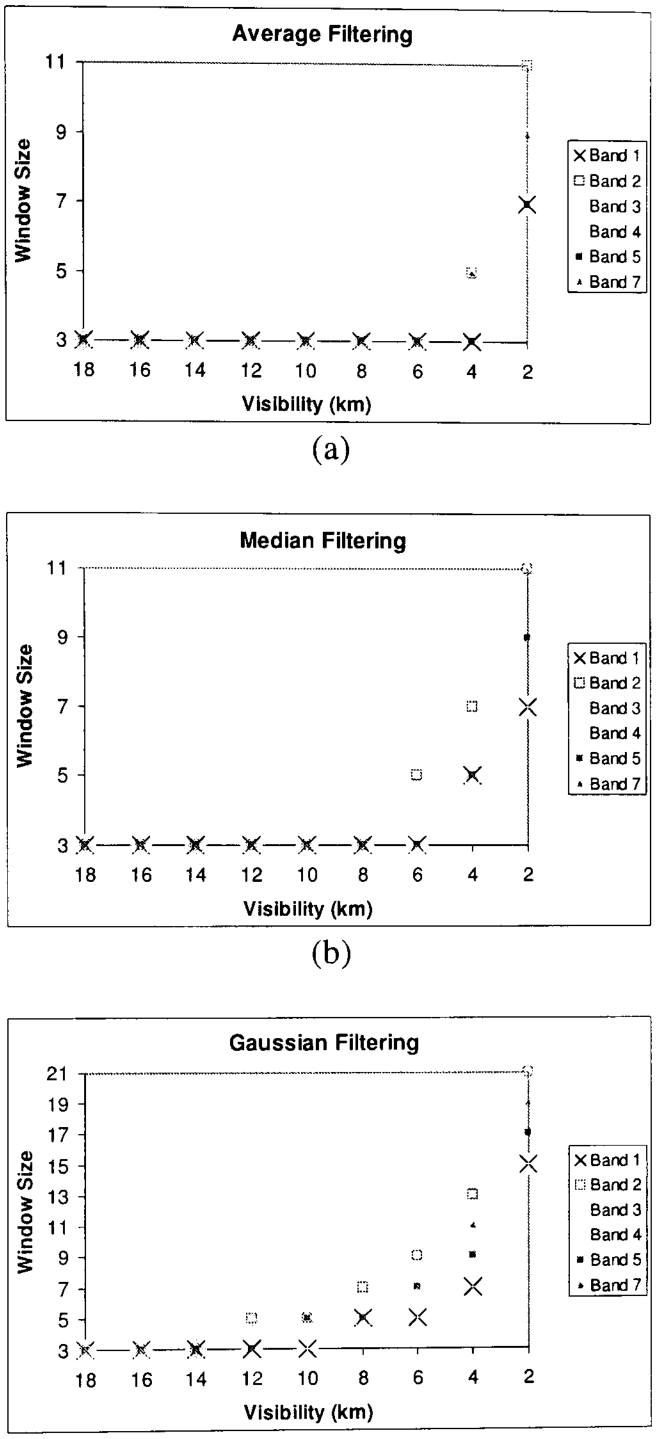
(b)

(c)

Figure 5.14: *Visual effect of (a) average filtering, (b) median filtering and (c) Gaussian filtering for 12 km visibility band 1 with window sizes 3 by 3 (left) and 21 by 21 (right).*

Similar SNR calculations are carried out for bands 2, 3, 4, 5 and 7.

Figure 5.15((a) to (c)) show plots of window size needed to obtain the highest SNR by using average, median and Gaussian filtering respectively for visibilities 18 down to 2 km, for all bands. For average and median filtering, little variation with window size can be seen for long and moderate visibilities but larger windows are needed as visibility drops (Figure 5.15(a) and (b)). For Gaussian filtering, progressively increasing window sizes are needed with reducing visibility (Figure 5.15(c)).



(a)

(b)

(c)

Figure 5.15: *The optimal window size for (a) average filtering (b) median filtering and (c) Gaussian filtering for visibilities 18 to 2 km.*

Overall, the SNR for the Gaussian filtering, is higher than the average and median filtering, but only slightly improves from the original hazy data and weighted-mean subtracted data for good visibilities. The separation between SNR curves for the Gaussian filtered data and that of the weighted-mean subtracted data and original hazy data increases towards shorter visibilities due to the transition from smaller to larger windows, allowing the higher variation rate of $\beta^{(2)}(V)H_{i_v}$ to be reduced more effectively. For the same reason, a similar trend is also produced by the average and median filters.

Comparison between SNR of Filtered Data for Known and Estimated Mean

In this section, we compare the SNR of filtered data when the haze mean is known exactly (the ideal case) and when the haze mean is estimated. Note that the estimated means are those calculated in Section 5.4.1. In this section, each filtering method is assigned the corresponding optimal window size for each band (see Tables 5.5 to 5.11). Figures 5.22, 5.23 and 5.24 show the effect of filtering on the SNR between the ideal and estimated case for the average, median and Gaussian filtering respectively. Due to the very large SNR range, Gaussian filtering plots are in dB (Figure 5.18). Overall, data restoration using average filtering for bands 2, 3, 4, 5, and 7 (Figure 5.16(b to f)), exhibits a very good fit between the estimated and ideal case. For band 1 (Figure 5.16(a)), a relatively good fit is seen from 18 to 6 km visibility, but an increasing discrepancy occurs from 6 to 2 km visibility; the SNR difference at 2 km visibility is about 50. The discrepancy can be linked to the regression plot of band 1 (see Figure 5.7(a)), on weighted haze mean and PIF radiance, where the regression curve does not exactly match the data ('♦') but slightly deviates from the data at the extremely low and high PIF radiance. For the same reason, a similar discrepancy can be seen in the band 1 plot for median (Figure 5.17(a)) and Gaussian filtering (Figure 5.18(a)). The rest of the bands show very good agreement between the estimated and ideal cases.

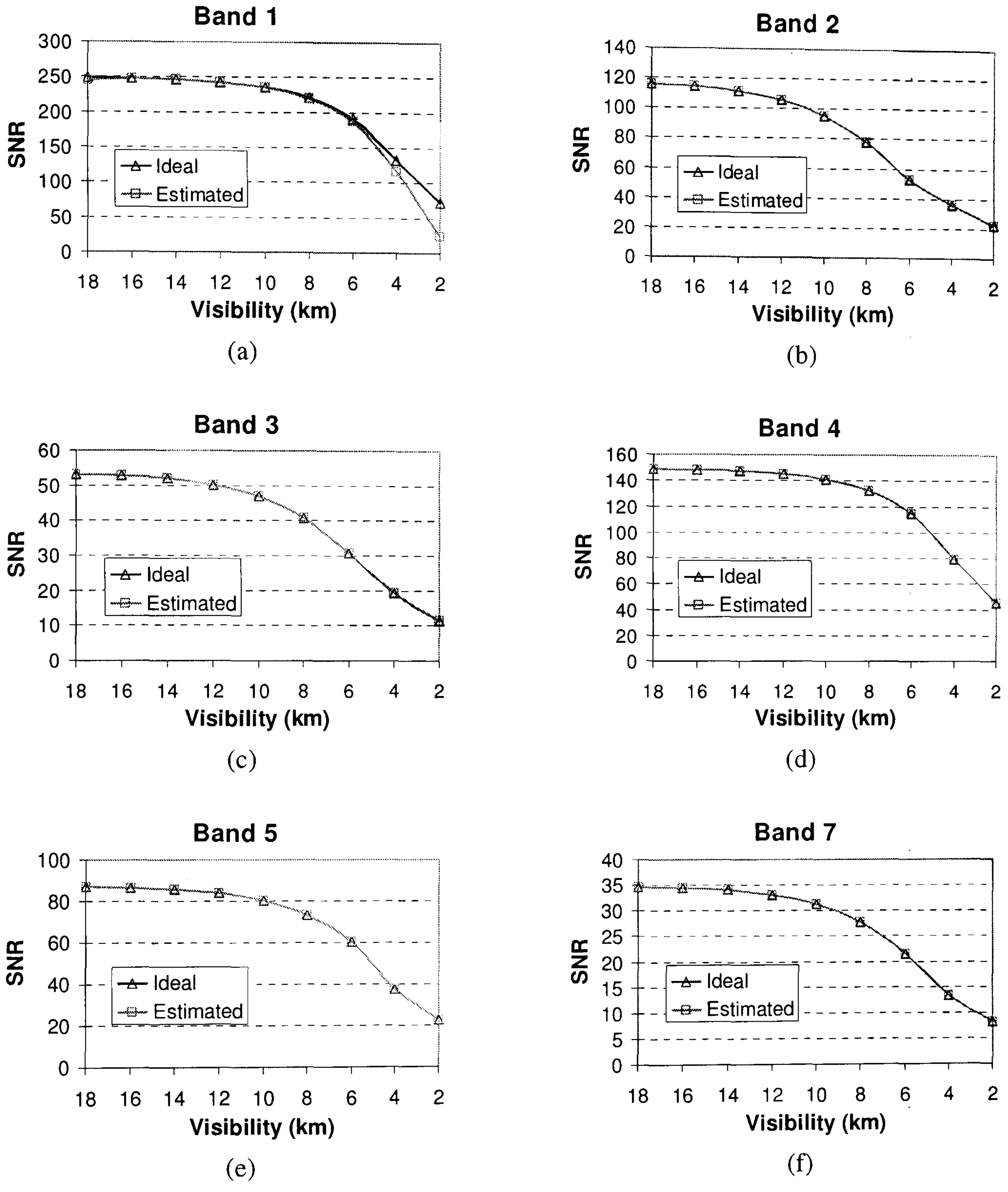
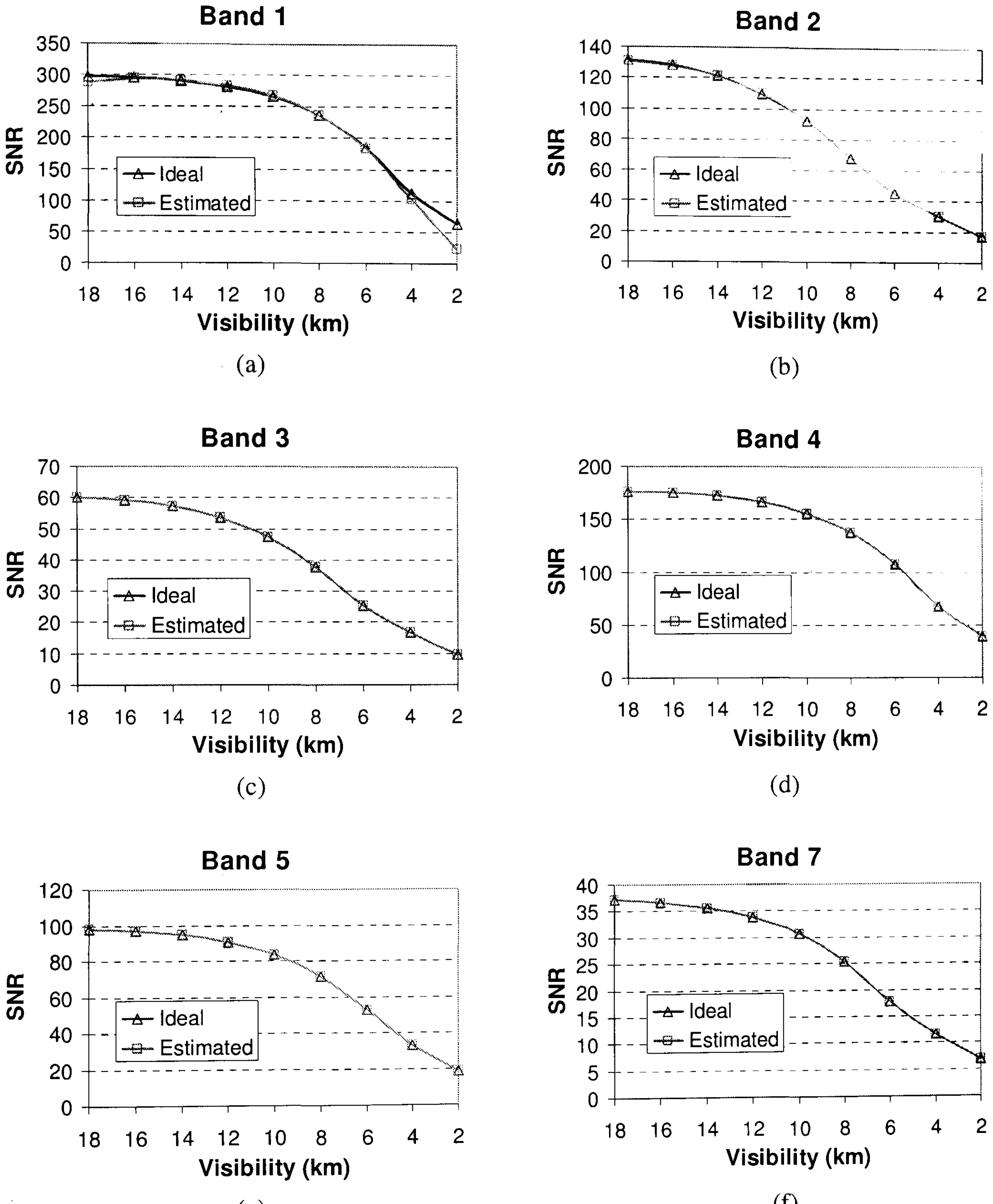


Figure 5.16: *The effect of average filtering on the SNR for the ideal and estimated case. The former corresponds to the case when the haze mean is known exactly and the latter, when the haze mean is estimated.*



(a)

(b)

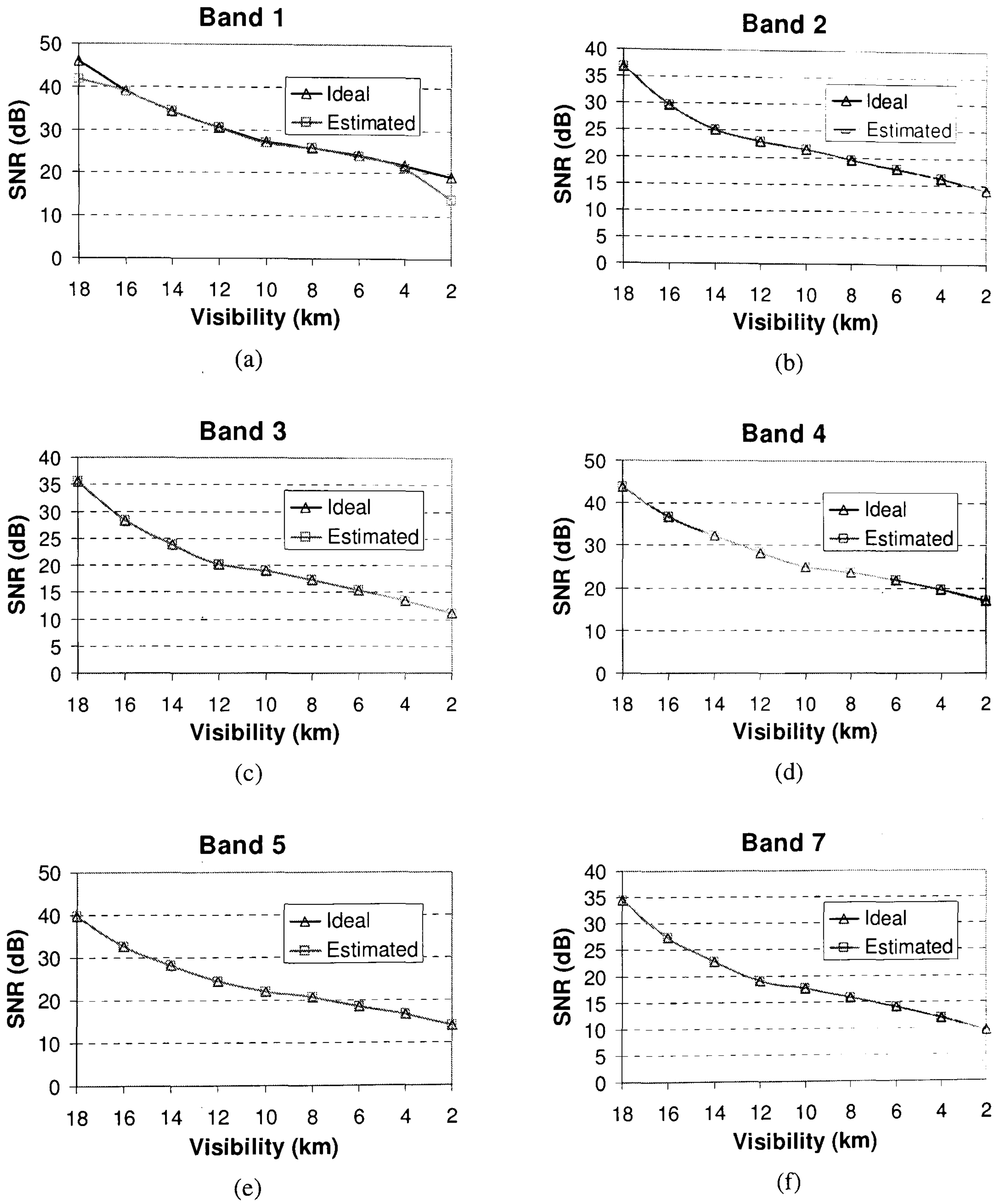
(c)

(d)

(e)

(f)

Figure 5.17: *Same as Figure 5.16 but for median filtering.*



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.18: *Same as Figure 5.17 but for Gaussian filtering.*

5.5.2 ML Classification and Classification Accuracy

The ultimate aim of a restoration process is to produce reliable restored data to be used in land classification. Since classification is normally carried out by making use of band combinations, SNR may not be the best way to measure the quality of multiple bands. Hence here we measure performance by carrying out classification on the restored dataset and then measuring its relative performance by classification accuracy.

In this section, ML classification is applied to the datasets, filtered using average, median and Gaussian filtering for 18 to 2 km visibility, in which the training pixels were chosen from the hazy datasets (see Section 3.4). Classification accuracy is then calculated by comparing the ML classification of the hazy dataset with that of the clear dataset using a confusion matrix.

Classification Accuracy of Restored Data

Figure 5.19, Figure 5.20 and Figure 5.21 show plots of classification accuracy against visibility for average, median and Gaussian filtered datasets respectively. The classification accuracies of unfiltered datasets, i.e. the weighted-mean subtracted data, $L_{i_z}(V)$ are also plotted. The results show that average filtering is able to improve classification accuracy for visibilities 9 km and less, and the improvement increases as the visibility reduces (Figure 5.19). At 2 km visibility, the classification accuracy is 20% better than for the hazy data. At 18 km visibility, average filtered data gives about 15% less classification accuracy than for the hazy data.

From Equation (5.6), when the difference between the estimated and actual weighted haze mean is insignificant $\left(\widehat{\beta_i^{(2)}(V)\overline{H_i}} \approx \beta_i^{(2)}(V)\overline{H_i}\right)$, $\left[\beta_i^{(2)}(V)\overline{H_i} - \widehat{\beta_i^{(2)}(V)\overline{H_i}}\right]$ can be neglected. Hence, the weighted mean subtracted data becomes:

$$\widehat{L_{i_Z}(V)} = \left[1-\beta_i^{(1)}(V)\right]T_i + \beta_i^{(2)}(V)H_{i_V} + L_O \qquad \text{... (5.33)}$$

For linear filtering, Equation (5.8) becomes:

$$\widehat{f_i(V)} = \left[1-\beta_i^{(1)}(V)\right]h_{\text{linear}}(T_i) + \beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_V}\right) + L_O \qquad \text{... (5.34)}$$

For longer visibilities, $\left[1-\beta_i^{(1)}(V)\right]h_{\text{linear}}(T_i)$ is much larger than $\beta_i^{(2)}(V)h_{\text{linear}}\left(H_{i_V}\right)$. Average filtering reduces pixel-to-pixel variations so that T_i , which is more variable, to be more affected than H_{i_V} , which is less variable. Consequently, the restored data experiences loss of surface information, resulting in a more significant difference between $\left[1-\beta_i^{(1)}(V)\right]h_{\text{linear}}(T_i)$ and T_i compared to $\left[1-\beta_i^{(1)}(V)\right]T_i$ and T_i in Equation (5.33). Thus, the classification accuracy of the restored data $\widehat{f_i(V)}$ becomes worse than the weighted-mean subtracted data, $L_{i_Z}(V)$. As the visibility reduces, the variability of H_{i_V} increases and eventually overtakes that of T_i . Hence, as visibility reduces, the classification accuracy of the average filtered data becomes higher than that of the hazy data.

Similar performance is shown by median filtering (Figure 5.20). By eliminating $\left[\beta_i^{(2)}(V)\overline{H_i} - \overline{\beta_i^{(2)}(V)H_i}\right]$ in Equation (5.10), we have:

$$\widehat{f_i(V)} = \text{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i + \beta_i^{(2)}(V)H_{i_V} + L_O\right) \qquad \text{... (5.35)}$$

At longer visibilities, most pixels possess a higher signal, $\left[1-\beta_i^{(1)}(V)\right]T_i$ than the haze randomness, $\beta_i^{(2)}(V)H_{i_V}$. Therefore $\text{Median}\left(\left[1-\beta_i^{(1)}(V)\right]T_i + \beta_i^{(2)}(V)H_{i_V} + L_O\right)$ is more influenced by $\left[1-\beta_i^{(1)}(V)\right]T_i$ than $\beta_i^{(2)}(V)H_{i_v}$. However, due to the behaviour of median filtering, the difference between $\widehat{f_i(V)}$ and T_i is more significant than $\widehat{L_{i_z}(V)}$ and T in Equation (5.33). Consequently, the classification accuracy of $\widehat{f_i(V)}$ becomes lower than that of $\widehat{L_{i_z}(V)}$. At shorter visibilities, although having more severe $\beta_i^{(2)}(V)H_{i_v}$, use of larger windows increases the classification accuracy because the difference between $\widehat{f_i(V)}$ and T_i becomes less significant than between $L_{i_z}(V)$ and T_i.

From 18 to 10 km visibility, the Gaussian filtered data has almost the same classification accuracy as the hazy data, but improves increasingly towards shorter visibilities (Figure 5.21). At longer visibilities, not much improvement in classification accuracy can be made because the hazy pixels already possess a high surface signal compared to haze randomness, this agrees with the corresponding SNR trend discussed in Section 5.5.1. Increasing classification accuracy can be observed from 10 to 2 km visibility. The changes in trend at about 10 to 8 km visibility are due to the transition to larger windows, needed to reduce the variation in $\beta_i^{(2)}(V)H_{i_v}$, which becomes higher as the visibility reduces.

Figure 5.22 shows classification accuracy for the average, median and Gaussian filtered data, and the unfiltered data against visibility. Comparison of classification accuracy amongst the filtering methods with hazy data indicates that Gaussian filtering is likely to be the preferable restoration method and thus will be used to reduce haze from real hazy data, as discussed in the following section. Figure 5.23 shows accuracy difference between before and after for all types of filtering, which gives a clearer picture regarding the capability of Gaussian filtering compared to average and median filtering. The accuracy difference between before and after Gaussian filtering is higher than for average and median filtering. It is noticeable that almost nothing is done by the Gaussian filter at visibilities 10 km and above, which signifies that such filtering is not needed at these visibilities.

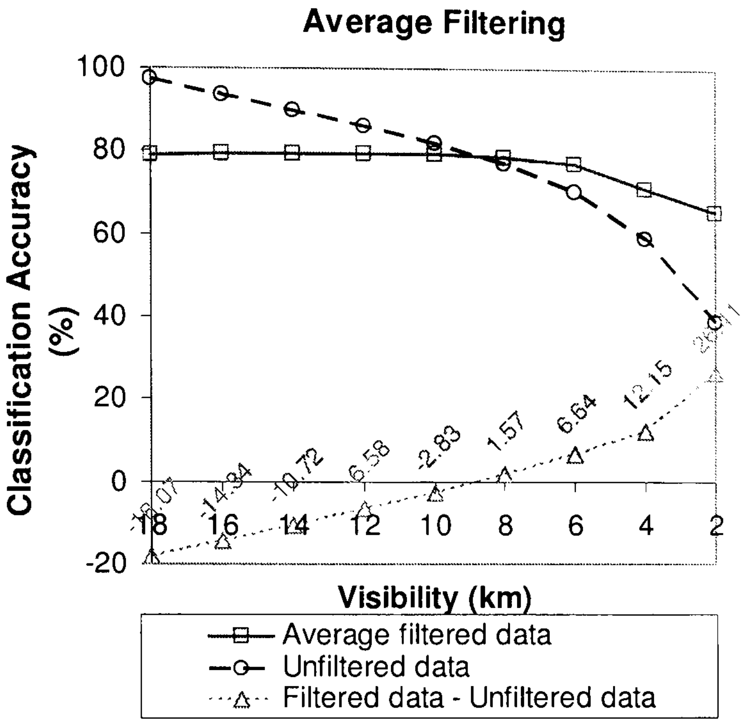


Figure 5.19: *Classification accuracy against visibility for average filtering. The accuracy difference between after and before filtering can be used to predict the improvement expected to be made for any hazy data.*

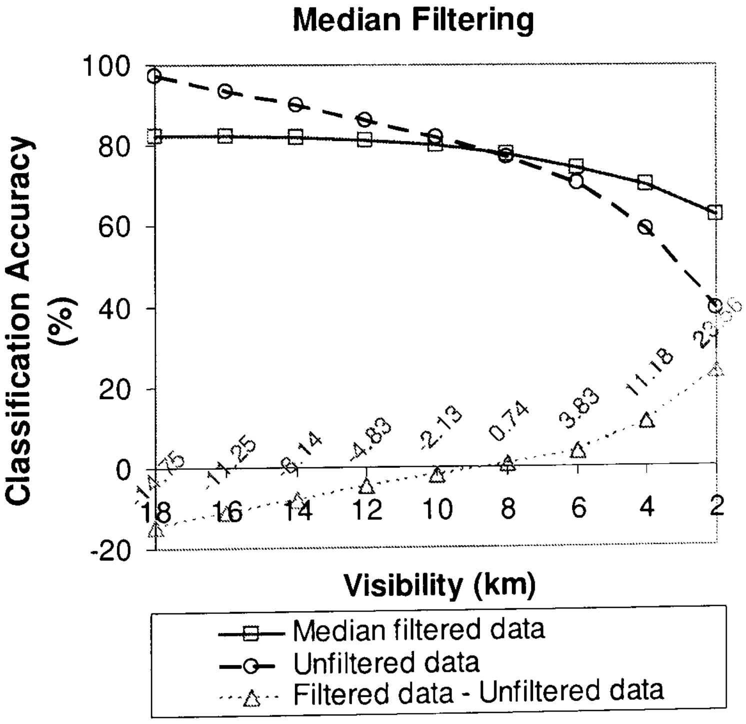


Figure 5.20: *Same as Figure 5.19 but for median filtering.*

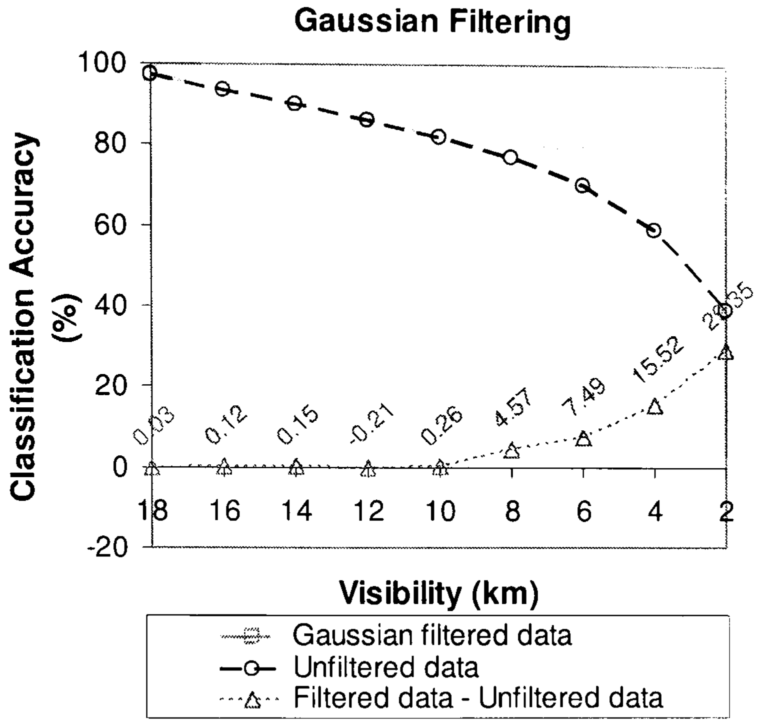


Figure 5.21: *Same as Figure 5.20 but for Gaussian filtering.*

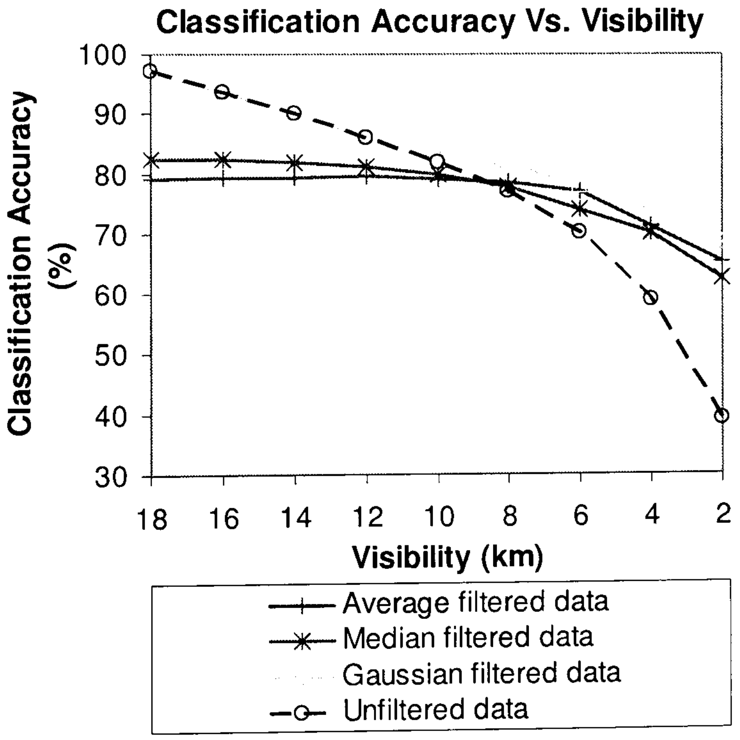


Figure 5.22: *Classification accuracy for the average, median and Gaussian filtered data, and the unfiltered data against visibility.*

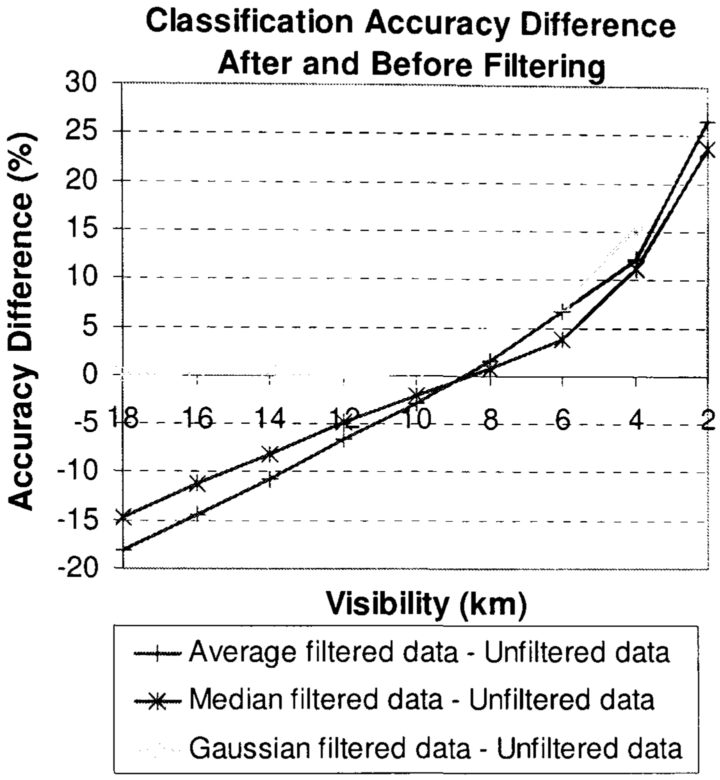


Figure 5.23: *Accuracy difference between before and after the average, median and Gaussian filtering.*

5.6 Application of Haze Removal on A Real Hazy Dataset

For testing the haze removal, we use the flowchart shown in Figure 5.25. Compared to Figure 5.4, in this flowchart, we create additional steps where we initially check whether the haze within the data is uniform or not uniform. If the haze is uniform, we straightforwardly can estimate and then subtract the weighted haze mean from the hazy data. On the other hand, if the haze is not uniform, we first need to use bands 1, 2 and 3 as input to MNF transformation. Next, the haze is segmented so that within each segment the haze is relatively uniform; then, estimation of haze mean using PIFs is performed. The later steps are similar to Figure 5.4 except in the quality assessment, we only use the classification accuracy to measure the performance of the haze removal.

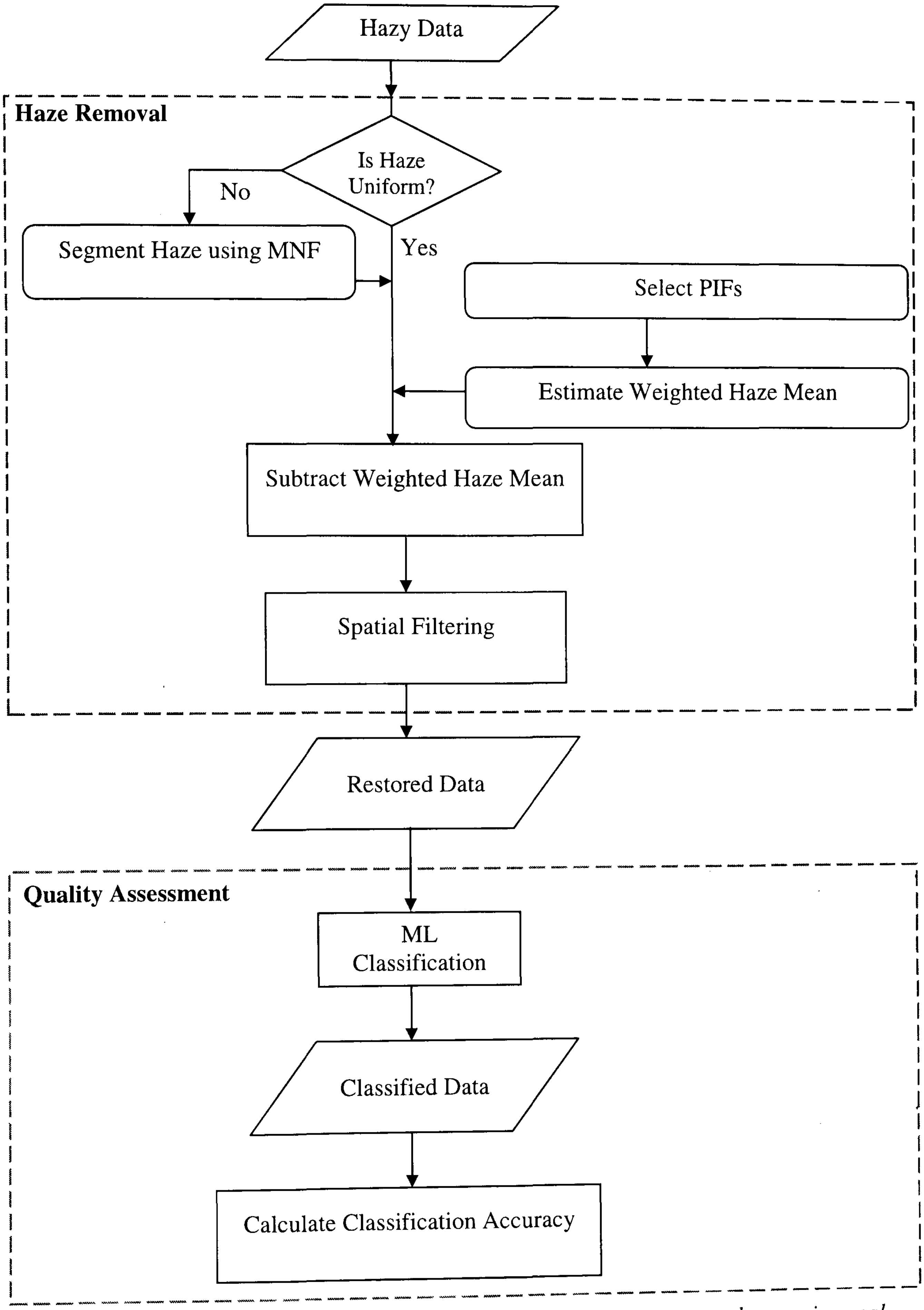


Figure 5.24: *Flowchart of haze removal and quality assessment procedures using real hazy datasets.*

In this section, we carry out haze removal on a real hazy dataset. The test site is Bukit Beruntung, located appoximately within 101° 30' and 101° 30' East and 3° 22' and 3° 27' in the district of Hulu Selangor, Selangor, Malaysia (Figure 5.25). The topography is relatively flat; with main ground cover types include rubber, urban and cleared land, which are among the classes as discussed in Chapter 3. Urban consists of mainly factory buildings, houses and recreational premises. The PM_{10} API measurements at a nearby station, Petaling Jaya on 6 August 2005 is 105, equivalent to 160 $\mu g\ m^{-3}$ of PM_{10} concentration (Department of Environment 2005) and with 6 km visibility (i.e. moderate haze) (Malaysian Meteorological Services 2005).

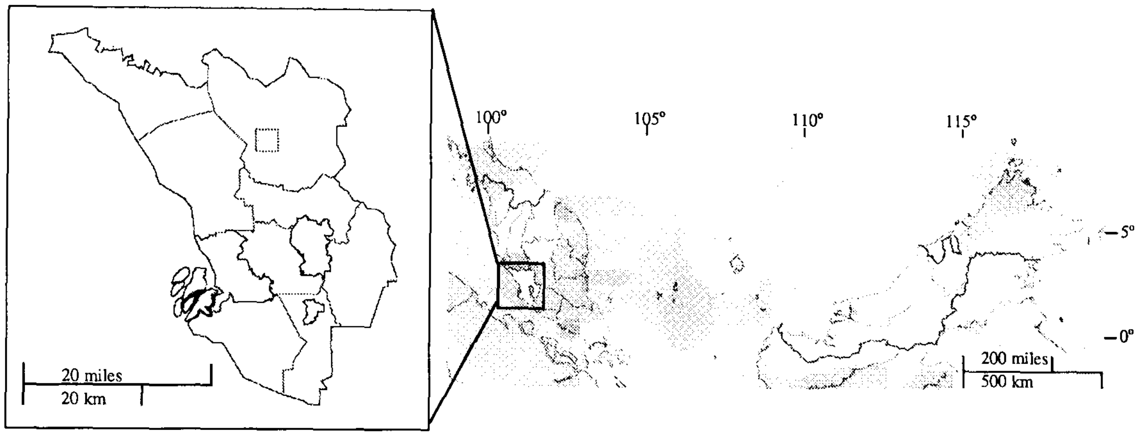


Figure 5.25: *Map of Malaysia (green), showing state of Selangor, district of Hulu Selangor (yellow) and the location of the test site, Bukit Beruntung (red box).*

A Landsat-5 TM dataset from the same date is shown in Figure 5.26(a) with bands 3, 2, and 1 assigned to the red, green and blue channels respectively. The dataset was covered by non-uniform haze caused by smoke that originated from forest fire in Sumatra, Indonesia (Mahmud 2009). Thick haze patches can be seen mainly in the centre and lower part of the image. Only a single date Landsat data is used for testing because (1) the difficulty in obtaining hazy images and their ground truth measurements and (2) the data contains important land covers which are blanketed with highly non-uniform haze; therefore, the robustness of the haze removal can be effectively assessed.

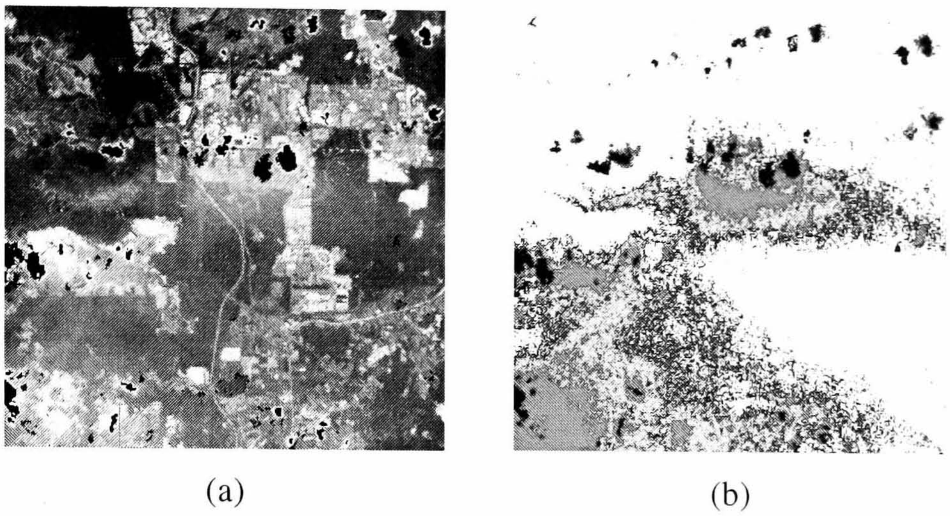
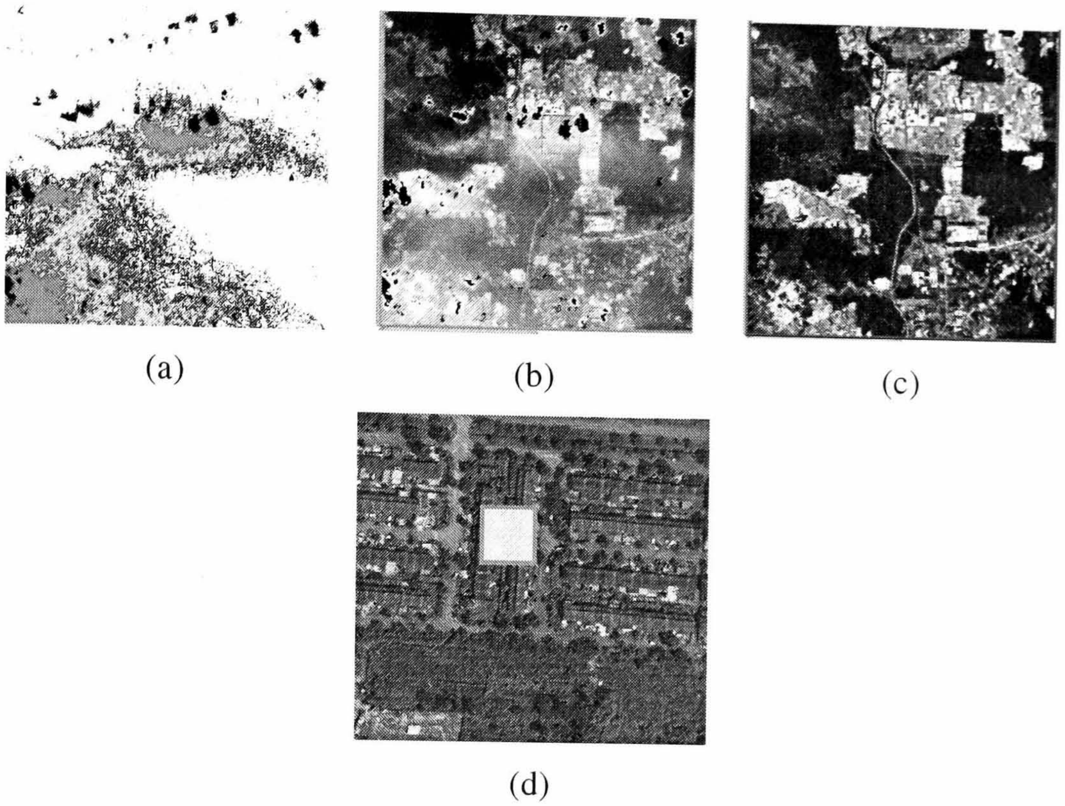

(a) (b)

Figure 5.26: *(a) Bands 3, 2 and 1 of Landsat data, for Bukit Beruntung from 6 August 2005, assigned to the red, green and blue channels respectively with cloud and its shadow masked in black and (b) Spatial distribution of haze levels: very hazy (red), hazy (green), moderate (blue) and clear (yellow).*

Initially, cloud and its shadow were masked in black (see Section 2.6). Haze is then segmented into four levels based on its severity: very hazy (red), hazy (green), moderate (blue) and clear (yellow) using MNF technique (Figure 5.26(b)) so that the haze is nearly homogenous. Ten PIF pixels were determined for each of the haze segments such as those used in the simulated datasets (see Section 5.4.1). A PIF consisting mainly terrace house rooftops is selected from the hazy data based on the knowledge of the area and aided with the Google Maps. To reduce the effects of mixed pixel, a PIF is selected among dense houses (Figure 5.27). This was carried out for the all ten PIFs within each segment.


(a) (b) (c)

(d)

Figure 5.27: *Location of some of the PIFs for (b) 6 August and (c) 22 August and (a) the haze levels: very hazy (red), hazy (green), moderate (blue) and clear (yellow); (d) close-up of a PIF observed from Google Maps.*

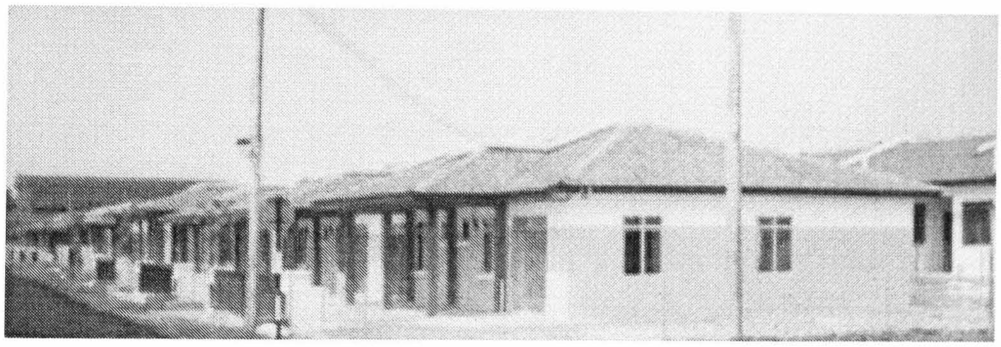


Figure 5.28: *Typical terrace houses at Bukit Beruntung.*

The PIF radiances were calculated and the haze mean radiance for band 1 was then determined based on the relationship established in Table 5.2 (see Section 5.4.1). The improved DOS method (Chavez 1988) was used to estimate the haze mean radiance for bands 2 and 3, for which the results are shown in Table 5.6.

Table 5.6: *Estimated haze mean radiances.*

Landsat Band	Haze Mean Radiance ($Wm^{-2}\ \mu m^{-1}\ sr^{-1}$)			
	Very Hazy	Hazy	Moderate	Clear
1	30	23	20	10
2	28	21	17	8
3	26	19	15	5

Because the haze has almost no effects in bands 4, 5 and 7 (we will show this later) and to avoid overcorrection in these bands, they will not undergo the haze removal procedures. The haze means for each of the haze segments were then subtracted from the hazy dataset for bands 1 to 3, and they were Gaussian filtered using a 5-by-5 window to reduce the remaining noise. The window size was based on optimal size for 8 km visibility – predicted from the classification accuracy of the hazy data itself (see Figure 5.21), in which is more representable for this hazy dataset compared to the 6 km visibility. This is because the 6 km visibility that indicate severe hazy conditions were measured from a point station, while the haze is much less severe for other parts; therefore the 6 km visibility is unrealistic to represent the entire scene. We believe this problem is due to the nature of the haze, which is highly non-uniform.

ML classification was then carried out using training pixels from the hazy dataset, which were based on a land cover map and a clear dataset of the test site. Finally, by making use of a clear dataset from 22 August 2005, the performance of the haze removal was assessed by means of visual analysis and classification accuracy. For the clear dataset, the PM_{10} API measurements in Petaling Jaya was 54 and visibility, 12 km (clear condition); the average API and visibility for the year 2005 for that station are 68 and 11 km respectively (i.e. 12 km is higher than the annual average visibility, therefore the dataset is good enough to be used as a reference dataset).

Figure 5.29 shows the individual bands before and after haze removal and the corresponding clear data for bands 1, 2 and 3; bands 4, 5 and 7 are also shown for comparison. It is noticeable that haze has much more effects on the visible bands 1, 2 and 3 than the near-infrared bands 4, 5 and 7. It seems that the near infrared bands are visually not affected by haze at all. After haze removal, it is noticeable that most of the haze patches in visible bands 1, 2 and 3 are removed. Figure 5.30 shows the results in terms of a colour composite image and ML classification. The difference in the composite image before and after haze removal is more apparent than in the individual bands. The performance of the haze removal is more evident in ML classification image. It is apparent from the middle and lower left of the image that the haze has caused some urban (red) pixels to be classified as cleared land (purple). Visual analysis of the hazy (Figure 5.30 (left)) and clear (Figure 5.30 (right)) from the enlarged image section (Figure 5.30 (bottom row)) gives a clearer picture of this. After haze removal, most of the urban pixels in the middle have been recovered.

Table 5.7 shows the confusion matrix for the hazy image with respect to the clear image in terms of pixels and percentages and the class user and producer accuracy in terms of pixels and percentages. Table 5.8 is the same as Table 5.7 but for the image after the haze removal. By comparing the confusion matrix of the data before (Table 5.7(c)) and after (Table 5.8(c)) haze removal, there is an increase of 10.7% and 2.6% in producer accuracy for urban and rubber respectively, but a drop of 4.6% and 5.8% for water and cleared land respectively. The drop is due to the remaining haze that causes an increase in reflectance (particularly in bands 1 – 3) for some of the pixels; consequently these pixels are assigned to the wrong classes by the ML. 16% (230) and 6% (2475) more pixels belonging to water and cleared land being classified as rubber and urban respectively after haze removal compared to before (Table 5.7(a, b) and Table 5.8(a, b)). The under correction is believed caused mainly by the highly non uniform condition of the haze within the scene. Nevertheless, the overall accuracy increases from 75% to 78% (i.e. 3% increase) and the kappa coefficient, from 0.62 to 0.66; these are equivalent to an increase from 8 to 9 km visibility (see Figure 5.21) and consistent with the analysis using the simulated dataset (Figure 5.21). The improvement of classification accuracy is quite small and is mainly due to the highly non-uniform haze within the scene that hampers the performance of the haze removal.

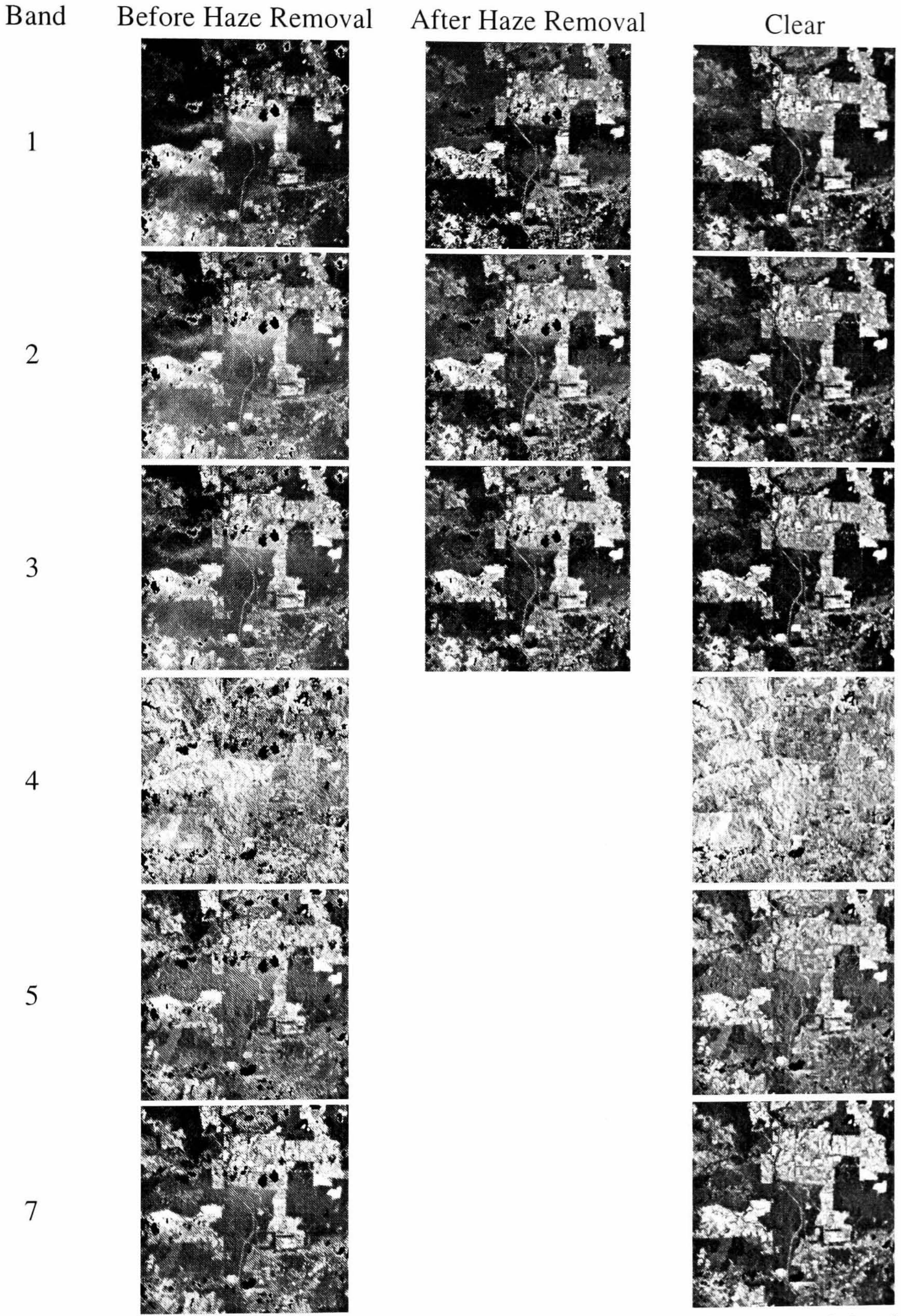


Figure 5.29: *Bands 1, 2 and 3 for the hazy dataset from 6 August 2005 before (left column) and after (middle column) haze removal, bands 4, 5 and 7 are also shown for comparison (lower left column), and the clear dataset (right column) from 22 August 2005. The black patches within the images in the first two columns are the masked clouds and cloud shadows.*

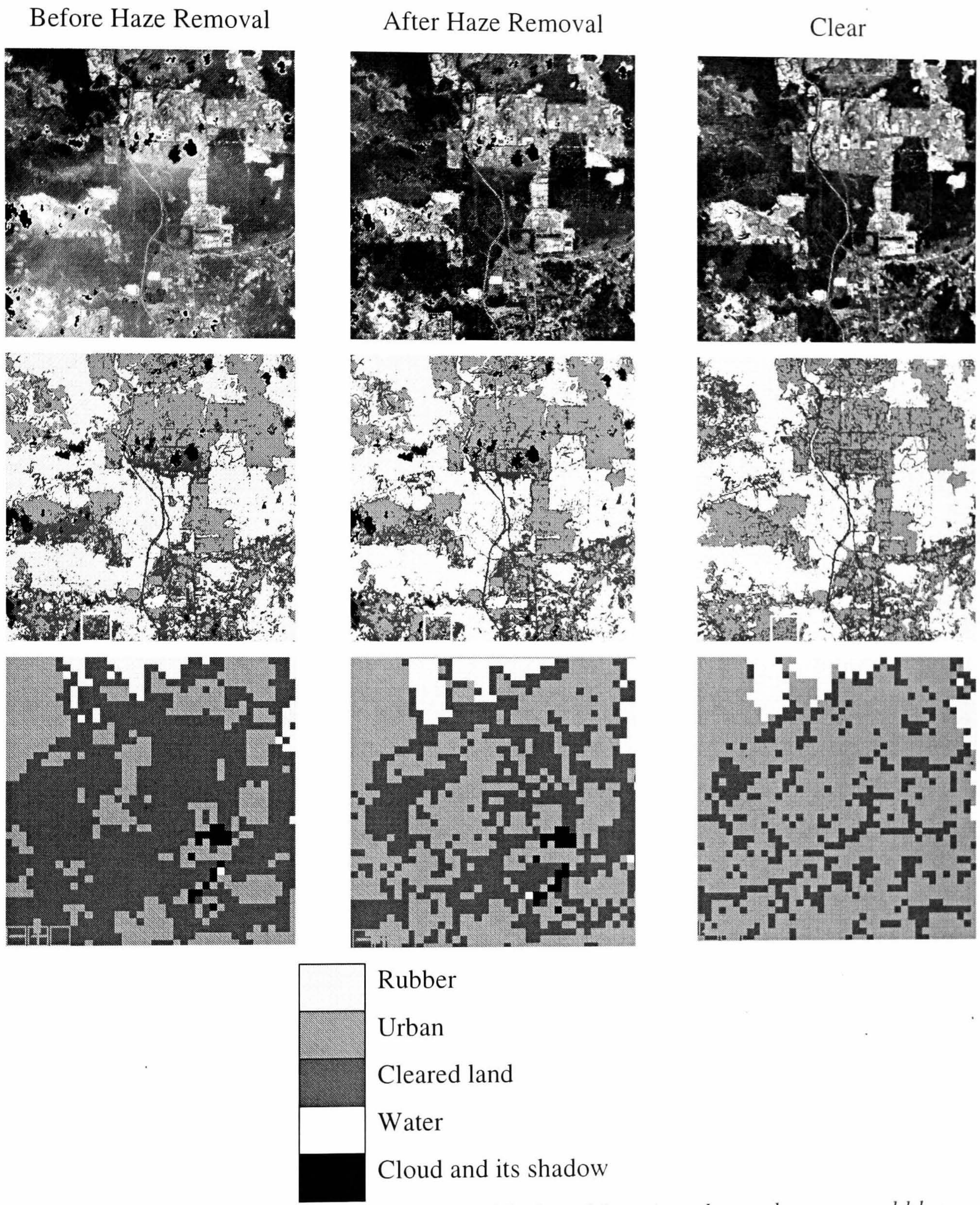


Figure 5.30: *Colour composite image of band 3, 2 and 1 assigned to red, green and blue (top row) respectively, ML classification (middle row) and the corresponding enlarged version (bottom row) before and after haze removal (left and middle column) and the clear image (right column). The enlarged version represents the area within the yellow box in the ML classification image.*

Table 5.7: *Confusion matrix of the hazy image with respect to the clear image in terms of (a) pixels and (b) percentages and (c) the class user and producer accuracy in terms of pixels and percentages.*

	Ground Truth (Pixels)				
Class	Water	Rubber	Cleared Land	Urban	Total
Water	1013	30	160	651	1854
Rubber	104	61162	7155	1472	69893
Cleared Land	112	6691	22698	13665	43166
Urban	168	391	8426	31467	40452
Total	1397	68274	38439	47255	155365

(a)

	Ground Truth (Percent)				
Class	Water	Rubber	Cleared Land	Urban	Total
Water	72.51	0.04	0.42	1.38	1.19
Rubber	7.44	89.58	18.61	3.12	44.99
Cleared Land	8.02	9.8	59.05	28.92	27.78
Urban	12.03	0.57	21.92	66.59	26.04
Total	100	100	100	100	100

(b)

Class	User Accuracy		Producer Accuracy	
	(Pixels)	(Percent)	(Pixels)	(Percent)
Water	384/1397	27.49	1013/1397	72.51
Rubber	7112/68274	10.42	61162/68274	89.58
Cleared Land	15741/38439	40.95	22698/38439	59.05
Urban	15788/47255	33.41	31467/47255	66.59

(c)

Overall Accuracy = 74.9%
Kappa Coefficient = 0.616

Table 5.8: *Confusion matrix of the image after the haze removal with respect to the clear image in terms of (a) pixels and (b) percentages and (c) the class user and producer accuracy in terms of pixels and percentages.*

	Ground Truth (Pixels)				
Class	Water	Rubber	Cleared Land	Urban	Total
Water	948	96	327	887	2258
Rubber	334	62927	6734	1548	71543
Cleared Land	61	4843	20477	8273	33654
Urban	54	408	10901	36547	47910
Total	1397	68274	38439	47255	155365

(a)

	Ground Truth (Percent)				
Class	Water	Rubber	Cleared Land	Urban	Total
Water	67.86	0.14	0.85	1.88	1.45
Rubber	23.91	92.17	17.52	3.28	46.05
Cleared Land	4.37	7.09	53.27	17.51	21.66
Urban	3.87	0.6	28.36	77.34	30.84
Total	100	100	100	100	100

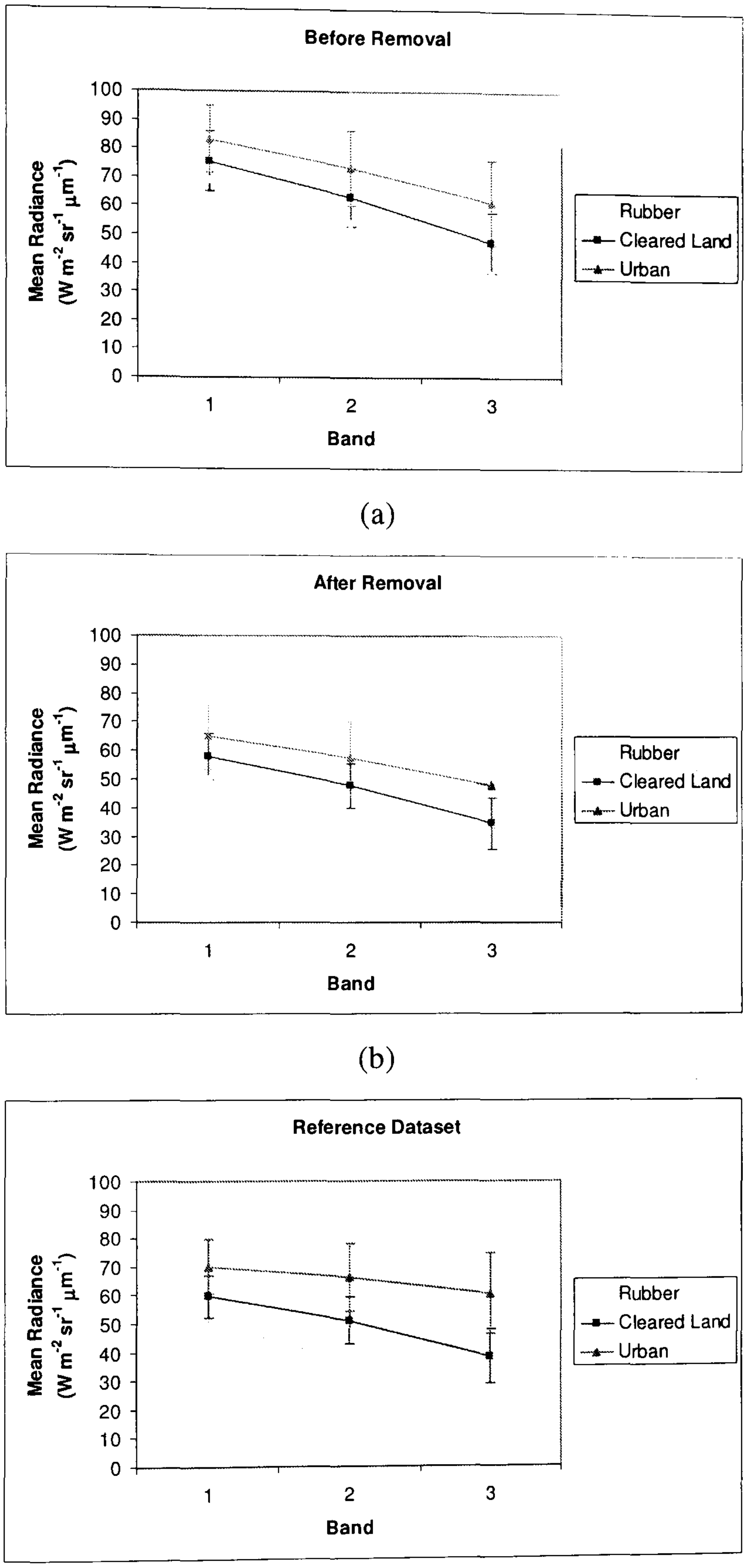
(b)

Class	User Accuracy		Producer Accuracy	
	(Pixels)	(Percent)	(Pixels)	(Percent)
Water	449/1397	32.14	948/1397	67.86
Rubber	5347/68274	7.83	62927/68274	92.17
Cleared Land	17962/38439	46.73	20477/38439	53.27
Urban	10708/47255	22.66	36547/47255	77.34

(c)

Overall Accuracy = 77.8%
Kappa Coefficient = 0.659

Next, we examine the classes' separability in terms of classes' means. Figure 5.31 shows the mean radiance for rubber, cleared land and urban for before and after removal and the reference dataset; vertical bars indicate standard deviations. After the haze removal, a more obvious separation of means can be observed for bands 1, 2 and 3. As predicted, these separations however are less significant than the reference dataset. These separations can be better observed by plotting mean difference for urban-cleared land, cleared land-rubber and urban-rubber (Figure 5.32). It can be seen that the haze removal increases the mean separation between cleared land-rubber in bands 1, 2 and 3. Urban-rubber experiences a little increase in bands 1, 2 and 3, but almost no change is observed for urban-cleared land because haze has less effects on bright compared to dark surfaces. In overall, the haze removal increases the separability between classes. As expected, the reference dataset gives the highest separation between class means because is free from haze. There is a slight decrease in the class standard deviation after compared to before removal, due to the removal of the haze effects (Table 5.9). The reference dataset posseses the smallest standard deviation because is free from haze.

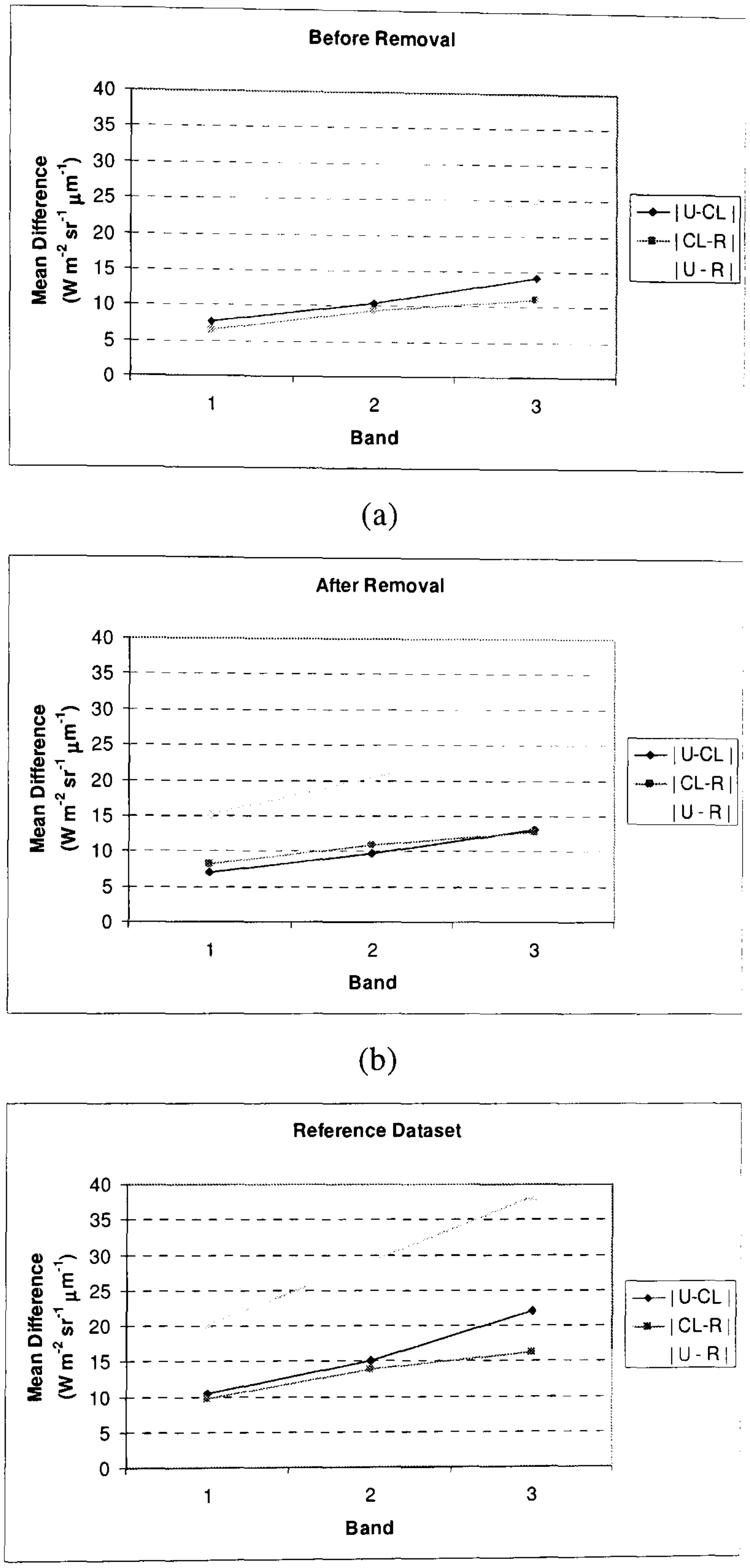


(a)

(b)

(c)

Figure 5.31: *Mean radiance for rubber, cleared land and urban: (a) before removal, (b) after removal and (c) reference dataset. Vertical bars indicate standard deviations.*



(a)

(b)

(c)

Figure 5.32: *Difference of mean radiance between urban (U) and cleared land (CL) and cleared land and rubber (R): (a) before removal, (b) after removal and (c) reference dataset.*

Table 5.9: *Mean and standard deviation for cleared land, rubber and urban: (a) before removal, (b) after removal and (c) reference dataset.*

(a) Before Removal

	Rubber		Cleared Land		Urban	
Band	Mean	Stdev	Mean	Stdev	Mean	Stdev
1	69.03	8.12	75.58	10.27	83.16	11.55
2	53.73	7.86	63.02	10.16	73.36	13.25
3	36.34	7.07	47.42	10.57	61.52	14.88
4	70.90	8.35	68.97	9.02	65.22	9.92
5	7.92	1.29	10.00	1.83	12.18	3.14
7	1.47	0.33	2.28	0.65	3.25	1.06

(b) After Removal

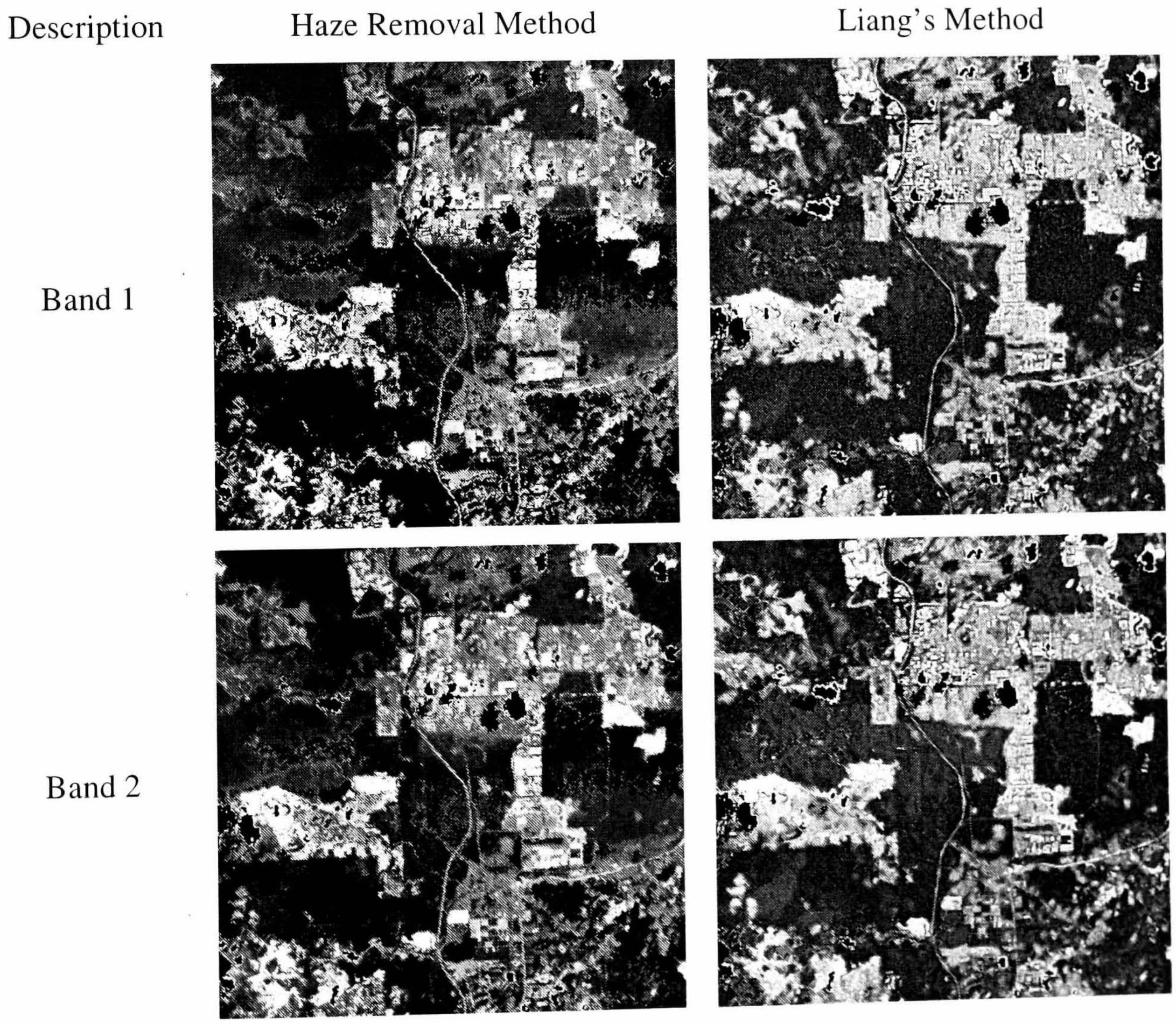
	Rubber		Cleared Land		Urban	
Band	Mean	Stdev	Mean	Stdev	Mean	Stdev
1	49.87	5.60	58.02	7.87	65.01	9.70
2	36.90	5.91	47.74	8.59	57.37	12.19
3	21.89	6.15	34.67	9.95	47.80	14.62
4	70.90	8.35	68.97	9.02	65.22	9.92
5	7.92	1.29	10.00	1.83	12.18	3.14
7	1.47	0.33	2.28	0.65	3.25	1.06

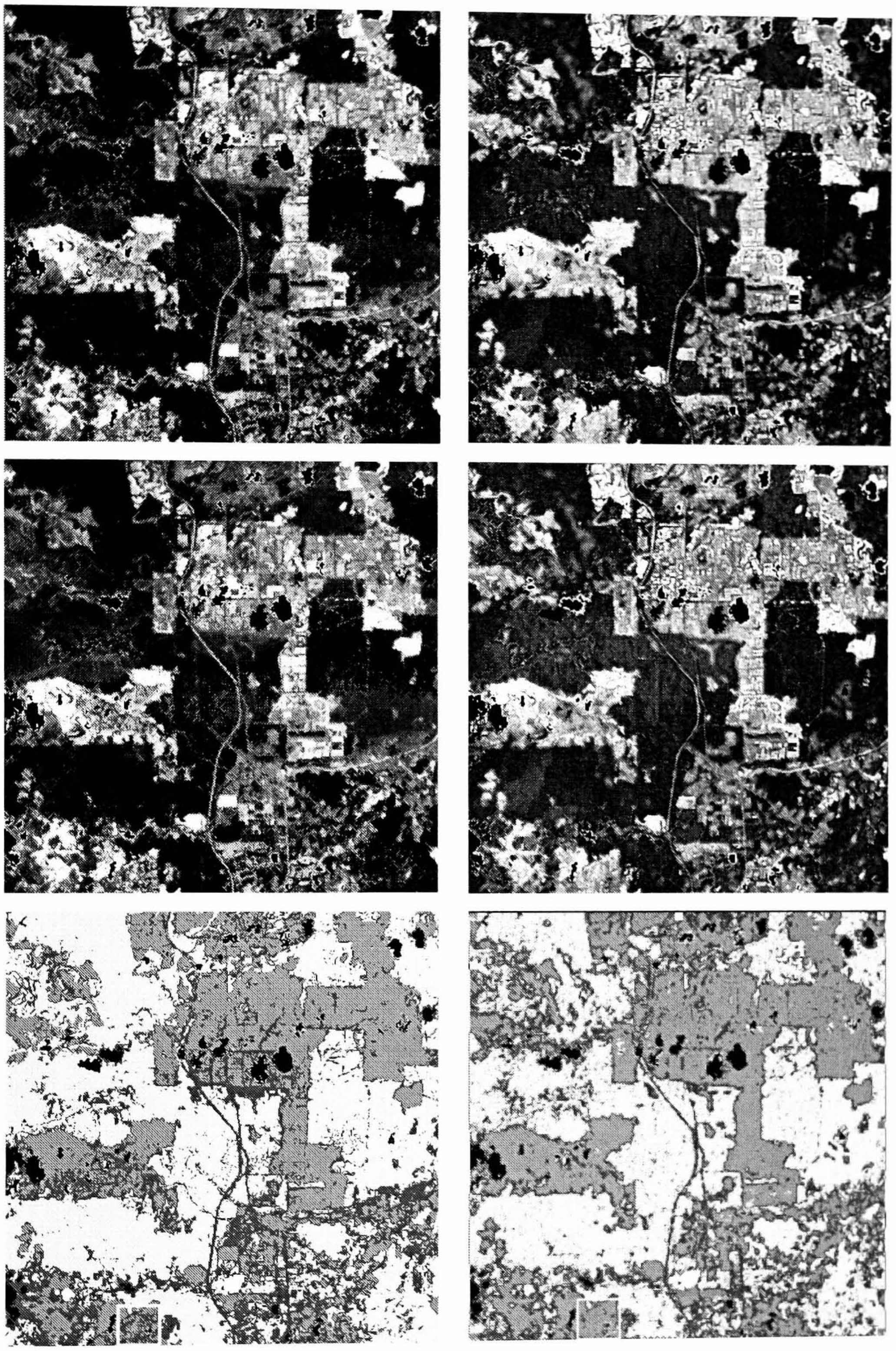
(c) Reference Dataset

	Rubber		Cleared Land		Urban	
Band	Mean	Stdev	Mean	Stdev	Mean	Stdev
1	49.88	1.84	59.64	7.50	70.13	9.40
2	37.07	2.76	50.99	8.25	66.09	11.83
3	21.99	2.21	38.24	9.55	60.24	14.18
4	71.84	11.02	69.04	10.31	61.55	10.16
5	7.39	1.24	9.94	1.65	12.75	2.78
7	1.13	0.23	2.13	0.62	3.36	1.04

Comparison with Liang's Method

Figure 5.33 shows the outcomes of the haze removal and Liang's method; rows 1 to 3 are bands 1 to 3, row 4 is the colour composite image of band 3, 2 and 1 assigned to red, green channel respectively, row 5 is the ML classification using bands 1, 2, 3, 4, 5 and 7 and row 6 is the corresponding enlarged version for the area within the yellow box. In terms of visual analysis, for bands 1, 2 and 3, both methods successfully reduced most haze within the scene. For the removal method, there seem to be slight traces of haze boundaries, particularly in the middle scene of bands 1, 2 and 3. Bands 1, 2 and 3 of Liang's method seems a little brighter than the removal method, probably due to the smaller dynamic range of pixel values when displayed under the same brightness range. For the colour composite image, the removal method seems slightly more redish than Liang's method, most likely due to the higher haze leftover effects in the red channel (i.e. band 3). For the classification image, in Liang's method most cleared land pixels (purple) that appeared within the urban (red) vanished, while this is not the case in the removal method (see the ML classification of reference data in Figure 5.30) . This is due to the effects of the haze residuals that decrease the separability between the urban and cleared land and therefore causing some cleared land pixels misclassified as urban.





Band 3

Band 3, 2, 1 assigned to R, G, B channel

ML Classification using bands 1, 2, 3, 4, 5 and 7

Enlarged version of the above

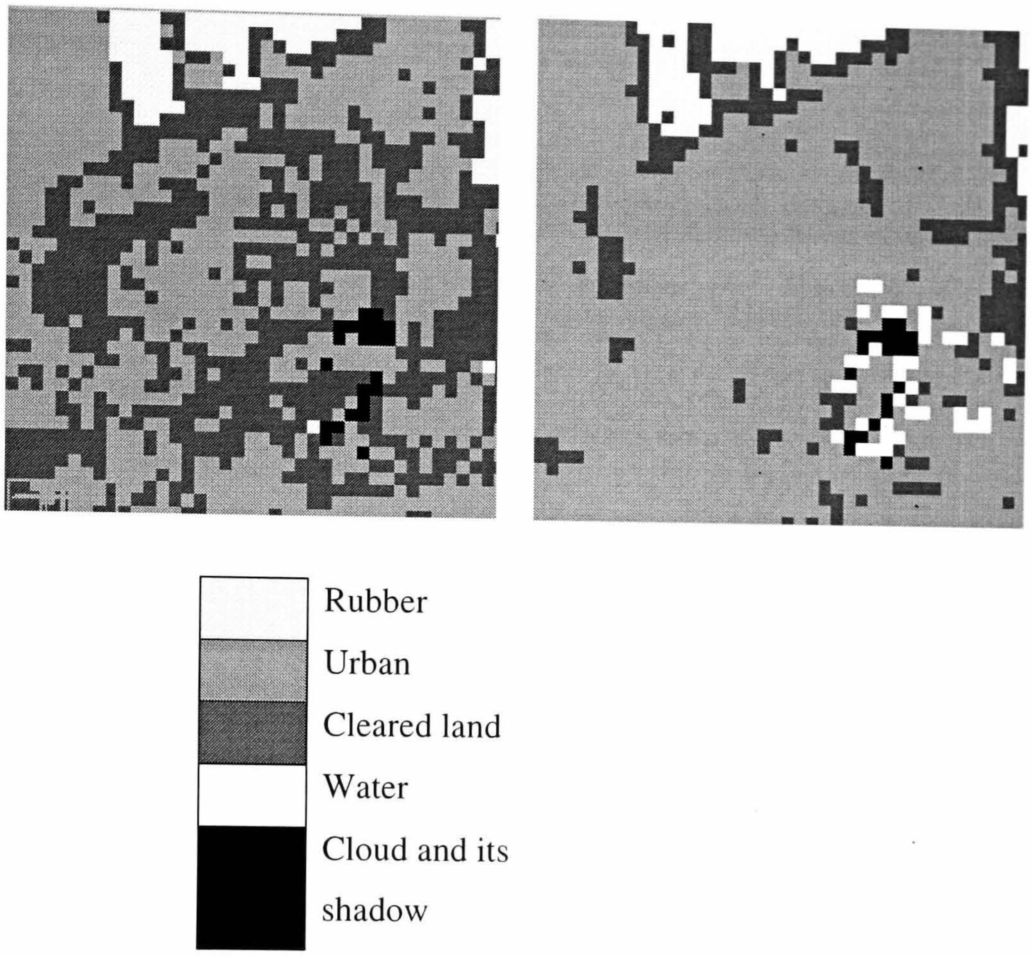


Figure 5.33: *Bands 1 to 4 (first four rows), colour composite image of band 3, 2 and 1 assigned to red, green channel respectively (fifth row), ML classification using bands 1, 2, 3, 4, 5 and 7 and the corresponding enlarged version for the area within the yellow box (sixth and seventh row respectively). The left column is the haze removal method, while Liang's method is on the right.*

Table 5.10 shows Confusion matrix of the hazy removed image using Liang's method with respect to the reference image. The overall accuracy increases only 0.1% (i.e. 75.0%) and no change for kappa coefficient (i.e. 0.616), compared to the hazy dataset. These are slightly lower than the haze removal, i.e. 77.8% for overall accuracy and 0.659 for kappa coefficient. This indicates that very small improvement in overall classification accuracy is obtained when using Liang's method compared to the haze removal method (3% increase in overall accuracy).

Table 5.10: Confusion matrix of the hazy removed image using Liang's method with respect to the reference image.

	Ground Truth (Pixels)				
Class	Water	Rubber	Cleared Land	Urban	Total
Water	1054	153	262	689	2158
Rubber	80	60859	9250	1754	71943
Cleared Land	112	5936	19235	9412	34695
Urban	151	1326	9692	35400	46569
Total	1397	68274	38439	47255	155365

(a)

	Ground Truth (Percent)				
Class	Water	Rubber	Cleared Land	Urban	Total
Water	75.45	0.22	0.68	1.46	1.39
Rubber	5.73	89.14	24.06	3.71	46.31
Cleared Land	8.02	8.69	50.04	19.92	22.33
Urban	10.81	1.94	25.21	74.91	29.97
Total	100	100	100	100	100

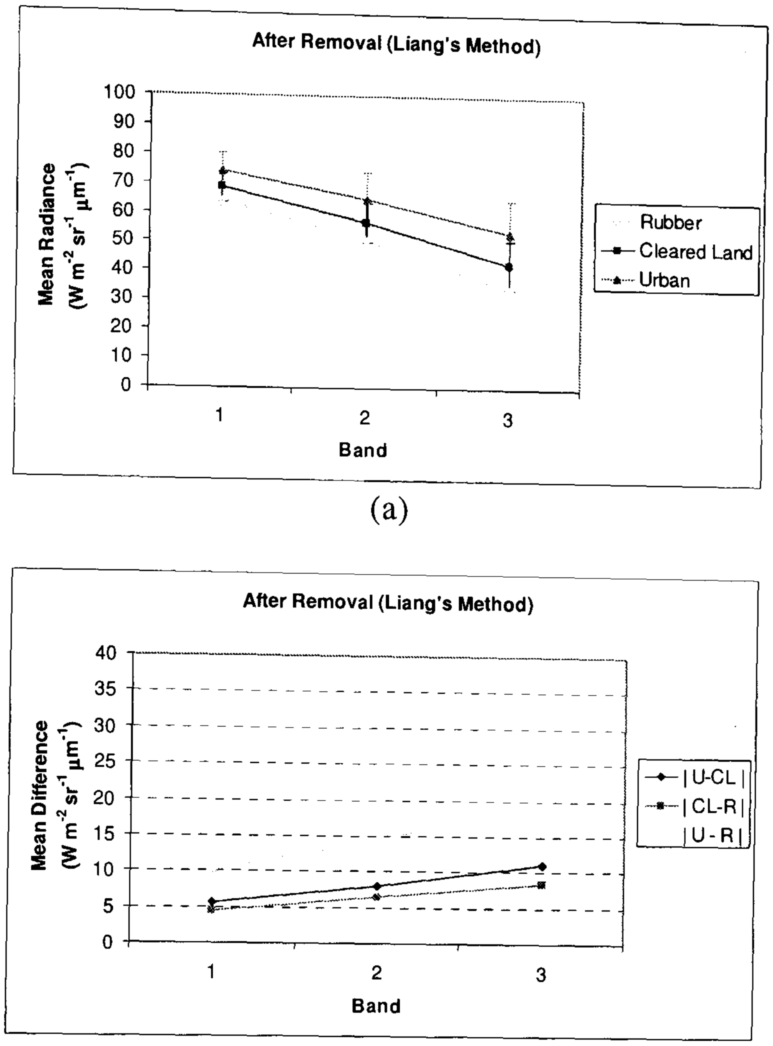
(b)

Class	Omission		Prod. Acc.	
	(Pixels)	(Percent)	(Pixels)	(Percent)
Water	343/1397	24.55	1054/1397	75.45
Rubber	7415/68274	10.86	60859/68274	89.14
Cleared Land	19204/38439	49.96	19235/38439	50.04
Urban	11855/47255	25.09	35400/47255	74.91

(c)

Overall Accuracy = 75.0%
Kappa Coefficient = 0.616

In terms of class means (Figure 5.34(a)), in Liang's method, the urban and rubber are slightly higher and lower respectively than the removal method, while the cleared land is about the same. This is because, in Liang's method, the haze correction over bright classes, e.g. urban, is much smaller than dark classes, e.g. rubber, compared to the removal method. This leads to the smaller separation between cleared land and rubber but higher separation between urban and cleared land (Figure 5.34(b)). In terms of standard deviations, Liang's method exhibits smaller values, particularly for bands 1, 2 and 3 compared to the removal method (Table 5.11). This is mainly due to the nature of correction procedures in Liang's method that initially replaces the pixels within hazy regions with the mean values of the same cluster but from clear regions.



(a)

(b)

Figure 5.34: *(a) Mean radiance againsts band for urban, cleared land and rubber, and (b) Mean difference between urban and cleared land and cleared land and rubber against band.*

Table 5.11: *Mean and standard deviation for rubber, cleared land and urban after haze removal using Liang's method.*

	Rubber		Cleared Land		Urban	
Band	Mean	Stdev	Mean	Stdev	Mean	Stdev
1	64.17	2.46	68.68	5.03	74.24	6.09
2	50.51	3.01	56.96	6.68	64.99	9.16
3	34.30	3.13	42.66	8.44	53.66	11.28
4	70.73	8.16	68.81	8.88	65.20	9.71
5	8.19	1.02	9.90	1.87	11.88	2.87
7	1.53	0.24	2.19	0.64	3.00	0.86

In overall, in relation to the purpose of this thesis, the haze removal method is preferable due to its ability to produce classification with higher accuracy compared to Liang's method. The fact that the removal method needing to use PIF to estimate haze mean radiances for all haze segments does not really matter because effort on making such task automated have been initiated (Nielsen et al. 1998; Canty et al. 2004; Schmidt et al. 2008), however this is rather beyond the scope of this thesis. When PIF is not available within the hazy scene, alternative methods, e.g. DOS, can be used with caution and by taking into account of its known weaknesses (e.g. secondary scattering effects). The main disadvantage of the removal method compared to Liang's method is the need to segment the haze into a number of locally homogenous regions using MNF method before haze radiance can be substracted.

5.7 Summary and Conclusions

In this chapter, we have developed a haze removal method based on weighted haze mean estimation and subtraction, and spatial filtering. The method was applied to simulated and real hazy datasets and its performance was measured using SNR and classification accuracy:

1. The two components that need to be dealt with in order to remove haze are weighted haze mean and haze randomness, both of which increase as visibility decreases.
2. Gaussian filtering gave the highest classification accuracy compared to average and median filtering, and its efficiency becoming more significant at shorter visibilities. For 10 km visibility and above, almost nothing is done by the Gaussian filtering, suggesting that filtering is not necessary at these visibilities.

3. Gaussian filtering gave the highest SNR and was able to improve SNR of mean subtracted data for all visibilities, but average and median filterings only improve SNR at particularly short visibilities.

4. An accurate estimation of the weighted haze mean is necessary in order to effectively subtract it from the data and so increase the SNR and the classification accuracy of the data.

5. In filtering, the window size needs to be increased in order to reduce the higher haze randomness as haze becomes severe, but an oversized window may worsen the data quality.

6. From visual analysis, the haze removal method successfully removed most haze from the real hazy dataset (i.e. Bukit Beruntung, Selangor, Malaysia). This was more evident in bands with shorter wavelengths.

7. From accuracy analysis, the haze removal only led to 3% improvement in classification accuracy in real data, mainly due to the highly non-uniform haze within the hazy dataset that hampered the performance of the haze removal process; nevertheless, this was close to that predicted using the simulated hazy dataset.

8. When compared with Liang's method, the haze removal method produced higher overall classification accuracy and kappa coefficient, therefore is more preferred; however, the main disadvantage is the need to segment the haze into a number of locally homogenous regions using MNF method and to find PIFs in each segment.

Chapter 6

Summary and Conclusions

The work presented in this thesis comprises results from four main parts: analysis of cloud detection and masking techniques for Malaysia, determining a suitable classification scheme for the study area, investigation of the effects of haze on land cover classification and development of haze removal methods. The main conclusions drawn from this thesis are as follows:

1. Spectral analysis based on MODIS scheme is suitable for cloud masking over Malaysia and consistent with climatological information mainly due to involving optimal use of MODIS rich bands.

   The MODIS analysis can be used to develop cloud mask for Landsat data with reasonably good accuracy. The developed cloud mask and cloud shadow mask give a high agreement when compared with the ACCA scheme and Luo et al. (2008) respectively when used onto two scenes of Landsat data.

   Since very thick haze has a standard deviation and reflectance similar to cloud; cloud can be used to simulate haze in remote sensing data when suitable hazy data is not available.

2. ML classification is suitable for Malaysian land covers due to its simplicity, objectivity and ability in classifying land covers with acceptable accuracy.

   Classification accuracy assessed using overall accuracy and producer accuracy (Congalton 1991) is most suitable due to its robustness and simplicity in assessing the quality of land cover classifications; the result is consistence with the accuracy measures introduced by Kaokoulas and Blackburn (2001).

The ability of ML to position class means at the different side of most of the decision boundaries appeared to be one of the key factors that enabled ML to discriminate effectively between classes.

3. Classification accuracy and producer accuracy decreases faster with visibility when the training pixels are drawn from the clear dataset rather from the hazy dataset itself.

   The investigation on the effects of haze on land classification shows that the haze becomes intolerable (i.e. below 85% overall classification accuracy) at visibilities less than about 11 km and 12 km for ML classification that use training pixels from the hazy dataset and the one that used training pixels from the clear dataset respectively.

   As haze gets very thick, spectral signatures curves of land covers become very close to each other, approximating the pure haze spectral signature.

   Statistical parameters that are most affected by haze are the class mean and standard deviation; the increase in class mean and standard deviation as haze increases are particularly significant for less reflective classes because the dark class has a low radiance, therefore the radiance scattered by the haze directly to the satellite's field of view dominates the apparent radiance observed by the satellite.

4. The analysis of haze removal shows that Gaussian filtering gives the best performance compared to average and median filtering in terms of SNR and classification accuracy, and the filter performance is better for shorter than longer visibilities.

   Analysis of the haze removal at a test site in Selangor, Malaysia, shows that after the haze removal, most of the haze was removed from the data; nevertheless, the improvement in accuracy was small, mainly due to the highly non-uniform haze within the hazy dataset that hampered the haze removal process.

The haze removal method produced higher overall classification accuracy and kappa coefficient than Liang's method, therefore is more preferred; however, the main disadvantages are the need to segment the haze into a number of locally homogenous regions using MNF method and to find PIFs in each of the segments so that the haze radiance can be estimated.

Summary of key findings

In studying the effects of atmosphere on land cover classification, the thesis revealed that haze becomes intolerable at visibilities less than about 11 km when ML classifications use training pixels from the hazy dataset and 12 km for those use training pixels from the clear dataset. To correct the effects of haze on land cover classification, haze radiance needs to be subtracted from the hazy data using PIF method and then the remaining noise needs to be filtered using Gaussian filtering., The key drawbacks of the removal method are that the haze needs to be segmented into smaller homogenous regions using MNF method and to find PIFs in each of the segments in order to determine the haze radiance.

Suggestions for Future Work

1. The MODIS cloud analysis need to be tested on more data with different locations and cloud conditions, and it is best to have ground truth data of cloud during satellite over passes. In order to really know the performance of the cloud analysis when ground data is not available, it is necessary to develop an objective accuracy measure.

2. Since the distribution of real haze tends to be non-uniform, further work need to be carried out to improve the haze simulation by improving the spatial correlation of haze and for a wider range of land covers.

3. To truly reveal the performance of the haze removal, it needs to be tested on a wider range of scenes with different haze conditions and land covers.

References

1. Ackerman, S. A., Strabala, K. I., Menzel, W. P., Frey, R. A., Moeller, C. C., & Gumley, L. E. (1998). Discriminating clear-sky from clouds with MODIS. *Journal of Geophysical Research*, 103(D24), 32141 – 32157.

2. Ackerman, Strabala, K., S., Menzel, P., Frey, R., Moeller, C., Gumley, L., Baum, B., Seemann, S. W. & Zhang, H., 2006. *Discriminating clear-sky from cloud with MODIS - Algorithm theoretical basis document. Products: MOD35. ATBD Version 5.0*, Madison: MODIS Cloud Mask Team.

3. Ackerman, S., Frey, R., Strabala, K., Liu, Y., Gumley, L., Baum, B. & Menzel, P., 2010. *Discriminating clear-sky from cloud with MODIS - Algorithm theoretical basis document. Products: MOD35. ATBD Version 6.1*, Madison: MODIS Cloud Mask Team.

4. Ahmad, A. & Hashim, M., 2002. Determination of haze using NOAA-14 AVHRR satellite data. *Proceedings on Asian Conference on Remote Sensing (ACRS 2002).* Available at: http://www.gisdevelopment.net/aars/acrs/2002/czm/050.pdf [Accessed: 7 September 2009].

5. Ahern, F. J., Erdle, T., Maclean, D. A. & Kneppeck, I. D., 1991. Quantitative relationship between forest growth rates and Thematic Mapper reflectance measurements. *International Journal of Remote Sensing*, 12(3), 387 – 400.

6. Anderson, G. P., Kneizys, F.X., Chetwynd, J. H., Clough, S. A. & Sheltle, E. P., 1986. *AFGL Atmospheric Constituent Profiles (0-120km).* Hanscom AFB, Ma : Air Force Geophysics Laboratory.

7. Anderson, J.R., Hardy, E.E., Roach, J.T., & Witmer, R. E., 1976. *A Land Use and Land Cover Classification System for Use with Remote Sensor Data.* Washington D.C. : United States Government Printing Office.

8. Andreae, M., Browell, E., Garstang, M., Gregory, G., Harriss, R., Hill, G., Jacob, D., Pereira, M., Sachse, G., Setzer, A., Silva Dias, D., Talbot, R., Torres, A. & Wofsey, S., 1988. Biomass-burning emissions and associated haze layers over Amazonia. *Journal of Geophysical Research*, 93(D2), 1509 –1527.

9. ARSM, 2005. *Malaysian Remote Sensing Agency Annual report 2005*. Malaysia : Malaysian Remote Sensing Agency.

10. ARSM, 2011. *Malaysian Remote Sensing Agency Website* (Updated 31 August 2011). Available at: http://www.remotesensing.gov.my [Accessed: 10 September 2011].

11. Baban, S. M. & Yusof, K. W., 2001. Mapping land use/cover distribution on a mountainous tropical island using remote sensing and GIS. *International Journal of Remote Sensing*, 22(10), p.1909 – 1918.

12. Bartholome E. & Belward A.S., 2005. GLC2000: a new approach to global land cover mapping from Earth observation data. *International Journal of Remote Sensing*, 26 (9), p.1959 – 1977.

13. Baum, B. A. & Trepte, Q., 1999. A grouped threshold approach for scene identification in AVHRR imagery. *Journal of Atmospheric and Oceanic Technology*, 16, p.793 – 800.

14. BBC 2011. *BBC Weather Website* (Updated 1 June 2011). Available at: http://www.bbc.co.uk/weather/weatherwise/factfiles/basics/clouds_types.shtml [Accessed: 11 July 2011].

15. Bendix, J., Rollenbeck, R. & Palacios W. E., 2004. Cloud detection in the Tropics – a suitable tool for climate – Ecological studies in the high mountains of Equador. *International Journal of Remote Sensing*, 25(21), p.4521 – 4540.

16. Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I. & Heynen, M., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GISready information. *ISPRS Journal of Photogrammetry & Remote Sensing*, 58, p.239 – 258.

17. Berk, A., Bernstein, L. S., Anderson, G. P., Acharya, P. K., Robertson, D. C., Chetwynd, J. H. & Adler-Golden, S. M., 1998. MODTRAN cloud and multiple scattering upgrades with application to AVIRIS. *Remote Sensing of Environment*, 65(3), p.367 – 375.

18. Bicheron, P., Amberg, V., Bourg, L., Petit, D., Huc, M., Miras, B., Brockmann, C., Hagolle, O., Delwart, S., Ranera, F., Leroy, M. & Arino, O., 2011. Geolocation Assessment of MERIS GlobCover Orthorectified Products. *IEEE Transactions on Geoscience and Remote Sensing*, 49(8), p. 2972 – 2982.

19. Bird, R. E. & Hulstrom, R. L., 1981. *A simplified clear sky model for direct and diffuse insolation on horizontal surfaces*. Colorado, USA : Solar Energy Research Institute (SERI), US Dept. of Energy

20. Buriez, J. C., Vanbauce, C., Parol, F., Goloub, P., Herman, M., Bonnel, B., Fouquart, Y., Couvert, P. & Seze, G., 1997. Cloud detection and derivation of cloud properties from POLDER. *International Journal of Remote Sensing*, 18(13), p.2785 – 2813.

21. Campbell, J. B., 2002. *Introduction to remote sensing*. London: Taylor & Francis

22. Canty M. J., Nielsen A. A. & Schmidt, M., 2004. Automatic radiometric normalization of multi-spectral imagery. *Remote Sensing of Environment*, 91(34), p. 441 – 451.

23. Canty, M. J., 2006. *Image analysis, classification and change detection in remote sensing: with algorithms for ENVI/IDL.* London : Taylor & Francis.

24. Chander, G., Markham, B. L. & Helder, D. L., 2009. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+ and EO-1 ALI sensors. *Remote Sensing of Environment*, 133, p.893 – 903.

25. Chandrasekhar, S., 1960. *Radiative transfer.* New York: Dover

26. Chavez, Jr. P. S., 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote Sensing of Environment.* 24, p.459 – 479.

27. Chen P. Y.; Srinivasan R.; Fedosejevs G.; Narasimhan B., 2002. An automated cloud detection method for daily NOAA-14 AVHRR data for Texas, USA. *International Journal of Remote Sensing*, Volume 23(22), p.2939 – 2950.

28. Chiang, C., Chen, W., Liang, W., Das, S. K. & Nee, J., 2007. Optical properties of tropospheric aerosols based on measurements of lidar, sun-photometer, and visibility at Chung-Li (251N, 1211E). *Atmospheric Environment*, 41, p.4128 – 4137.

29. Choi, Y.-S. & Ho C.-H., 2009. Validation of the cloud property retrievals from the MTSAT-1R imagery using MODIS observations. *International Journal of Remote Sensing*, 30, p.5935 – 5958.

30. Coakley, J. A., Jr., & Bretherton, F. P., 1982. Cloud Cover From High-Resolution Scanner Data: Detecting and Allowing for Partially Filled Fields of View. *Journal of Geophysical Research*, 87(C7), p.4917 – 4932.

31. Congalton, R. G., 1991. A review of assessing the accuracy of classification of remotely sensed data. *Remote Sensing of Environment*, 37, p.35 – 46.

32. Couvert, P. & Seze, G., 1997. Cloud detection and derivation of cloud properties from POLDER. *International Journal of Remote Sensing*, 18(13), p.2785 – 2813.

33. DeFries, R. S. & Townshend, J. R. G., 1994. NDVI-derived land cover classification at global scales. *International Journal of Remote Sensing*, 15, p.3567 – 3586.

34. DeFries, R. S., Hansen, M. C., Townshend, J. R. G. & Sohlberg, R. S., 1998. Global land cover classifications at 8 km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers. *International Journal of Remote Sensing,* 19, p.3141 – 3168.

35. Department of Environment, 1997. *A guide to air pollution index (API) in Malaysia*. Malaysia : Department of Environment.

36. Department of Environment, 2004. *Annual report 2004*. Malaysia : Department of Environment.

37. Department of Environment, 2010. *Air quality monitoring* (Updated 2008). Available at: http://www.doe.gov.my/en/content/air-quality-monitoring [Accessed: 7 September 2010].

38. Di Vittorio, A. V. & Emery, W. J., 2002. An automated, dynamic threshold cloud-masking Algorithm for daytime AVHRR images over land. *IEEE Transactions on Geoscience and Remote Sensing*, 40(8), p.1682–1694.

39. Du, Y., Guindon, B. & Cihlar, J., 2002. Haze detection and removal in high resolution satellite image with wavelet analysis, *IEEE Transaction on Geoscience and Remote Sensing*, 40(1), p.210 – 217.

40. Dybbroe, A., Karlsson, K.-G. & Thoss, A., 2005. SAFNWC AVHRR cloud detection and analysis usingdynamic thresholds and radiative transfer modeling. Part I: Algorithm description. *Journal of Applied Meteorology*, 44(1), p.39 – 54.

41. Dzubay, T. G. S., Steven, R. K., Lewis, C. W., Hem, D. H., Courtney, W. J., Tesch, J. W. & Mason, M. A., 1982. *Environmental Scence and. Technology.*, 16, p.514 – 525.

42. Eckhardt, J. P., Verdin, D. W. & Lyford, G. R., 1990. Automated update of an irrigated lands GIS using SPOT HRV imagery, *Photogrammetric Engineering and Remote Sensing*, 56, p.1515 – 1522.

43. ENVI, 2006. *User's guide*. Research systems.

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

45. Fischer J., Preusker, R. & Schueller, L., 1997. *ATBD cloud top pressure. Algorithm Theoretical Basis Document POTN-MEL-GS-0006*. European Space Agency.

46. Franca, G. B. & Cracknell, A. P., 1995. A simple cloud masking approach using NOAA AVHRR daytime data for tropical areas. *International Journal of Remote Sensing*, 9, p.1697 – 1705.

47. Fraser, R. S., Bahethi, O. P. & Al-Abbas, A. H., 1977. The effect of the atmosphere on classification of satellite observations to identify surface features, *Remote Sensing Environtment.*, 6, p.229 – 249.

48. Fuller, R. M., Groom, G. B. & Jones, A. R., 1994. The Land Cover Map of Great Britain: An automated classification of Landsat Thematic Mapper Data. *Photogrammetric Engineering & Remote Sensing*, 60(5), p.553 – 562.

49. Fuller, R.M., Cox, R., Clarke, R.T., Rothery, P., Hill, R.A., Smith, G.M., Thomson, A.G., Brown, N.J., Howard, D.C. & Stott, A.P., 2005. The UK land cover map 2000: Planning, construction and calibration of a remotely sensed, user-oriented map of broad habitats. *International Journal of Applied Earth Observation and Geoinformation*, 7(3), p.202 – 216.

50. Fuller, R. M., Smith, G. M. & Devereux, B. J., 2003. The characterisation and measurement of land cover change through remote sensing: problems in operational applications. *International Journal of Applied Earth Observation and Geoinformation*, 4(3), p.243 – 253.

51. Gallego, F. J., 2004. Remote sensing and land cover area estimation. International *Journal of Remote Sensing*, 25, p.3019 – 3047.

52. Gao, B.-C., Heidebrecht, K. B. & Goetz, A. F. H., 1993. Derivation of scaled surface reflectances from AVIRIS data. *Remote Sensing of Environment*, 44, p.165 – 178.

53. Gemmell, F. M., 1995. Effects of forest cover, terrain, and scale on timber volume estimation with Thematic Mapper data in a rocky mountain site. *Remote Sensing of Environment*, 51(2), p.291 – 305.

54. Godish, T., 1991. *Air Quality*. 2nd. ed. London : Lewish Publisher.

55. Guerschman, J. P., Paruelo, J. M., Di Bella, C., Giallorenzi, M. C. & Pacin, F., 2003. Land cover classification in the Argentine Pampas using multi-temporal Landsat TM data. *International Journal of Remote Sensing*, 24(17), p.3381 – 3402.

56. Guild, L. S., Kauffman, J. B., Cohen, W. B., Hlavka, C. A. & Ward, D. E., 2004. Modeling Biomass Burning Emissions for Amazon Forest and Pastures in Rondônia, Brazil. *Ecological Applications*, 14(4), p S232 – S246.

57. Gupta, A., 1996. Erosion and sediment yield in Southeast Asia: a regional perspective. *Erosion and Sediment Yield: Global and Regional Perspectives (Proceedings of the Exeter Symposium, July 1996)*. 236, p.215 – 222.

58. Green, A.A., Berman, M., Switzer, P., & Craig, M. D., 1988. A Transform for Ordering Multispectral Data in terms of Image Quality with Implications for Noise Removal, *IEEE Transactions Geoscience and Remote Sensing*, 26(1), p.65 – 74.

59. Hadjimitsis, D. G., Clayton, C. R. I., & Retalis, A., 2009. The use of selected pseudo-invariant targets for the application of atmospheric correction in multi-temporal studies using satellite remotely sensed imagery. *International Journal of Applied Earth Observation and Geoinformation*, 11(3), p.192 – 200.

60. Hahn, C. J. & Warren, S. G., 1999. Extended Edited Cloud Reports from Ships and Land Stations over the Globe, 1952-1996. *Numerical Data package NDP-*

026C, Carbon Dioxide Information Analysis Center (CDIAC). Oak Ridge, Tennessee, USA : Department of Energy.

61. Hansen, M. C., DeFries, R. S., Townshend, J. R. G. & Sohlberg, R., 2003. Global land cover classification at 1km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, 21, p.1331 – 1364.

62. Hashim, M., Kanniah, K. D., Ahmad, A., Rasib, A. W. & Ibrahim, A. L., 2004. The use of AVHRR data to determine the concentration of visible and invisible tropospheric pollutants originating from a 1997 forest fire in Southeast Asia. *International Journal on Remote Sensing*, 25(21), p.4781 – 4794.

63. Hashim, M, Ibrahim, A. L., Mohd, M.I. S. & Ahmad, S., 2004. Remote Sensing Education At University Of Technology Malaysia For Supporting Local Related Industries In *Attaining Sustainable Natural Resource And Environmental Managements. In: XXth ISPRS Congress , 12-23, July 2004, Istanbul, Turkey.*

64. Haywood, J. M., Pelon, J., Formenti, P., Bharmal, N., Brooks, M. E., Capes, P., Chou, C., Christopher, S. A., Coe, H., Cuesta, J., Derimian, Y., Desboeufs, K., Greed, G., Harrison, M. A. J., Heese, B., Highwood, E. J., Johnson, b. T., Mallet, T. M., Marticorena, B., Marsham, J., Milton, S. F., Myhre, G., Osborne, S., Parker, D. J., Rajot, J.-L., Schulz, M., Slingo, A., Tanré, D., Tulet, P., 2008. Overview of the dust and biomass burning experiment and african monsoon multidisciplinary analysis special observing period. *Journal of Geophysical Research*, 113, p. D00C17.

65. Henipavirus Ecology Collaborative Research Group (HERG), 2010. *Malaysia's smoke haze, 8th September 2005* (Updated 2005). Available at: http://www.henipavirus.org/features/malaysia_smoke_haze/malaysia_smoke_haze.htm [Accessed: 7 September 2010].

66. Heil, A., Langmann, B.. & Aldrian, E., 2007 Indonesian peat and vegetation fire emissions: Study on factors influencing large-scale smoke haze pollution using a regional atmospheric chemistry model. *Mitigation and Adaptation Strategies for Global Change*, 12(1), p.113 – 133.

67. Heil, A. & Goldammer, J. G., 2001. Smoke-haze pollution: a review of the 1997episode in Southeast Asia. *Regional Environmental Change*, 2, p.24 – 37.

68. Horvath, H., 1971. On the applicability of the Koschmieder visibility formula. *Atmospheric Environment*, 5, p.177 – 184.

69. Inoue, T., 1987. A cloud type classification with NOAA 7 split-window measurements. *Journal of Geophysical Research*, 92, p.3991 – 4000.

70. Ismail, M. H., Jusoff, K., 2008. Satellite data classification accuracy assessment based from reference dataset. *International Journal of Computer and Information Science and Engineering*, 2(2), 96 – 102.

71. Irish, R. R., Barker, J. L., Goward, S. N., & Arvidson, T., 2000. Characterization of the Landsat-7 ETM Automated Cloud-Cover Assessment (ACCA) Algorithm. *Photogrammetric Engineering & Remote Sensing*, 72(10), p.1179 – 1188.

72. Jacobson, N. P. & Gupta, M. R., 2005. Design goals and solutions for display of hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 43(11), p.2684 – 2692.

73. Janssen, L.F. & Molenaar, M., 1995. Terrain objects, their dynamics and their monitoring by integration of GIS and remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 33, p.749 – 758.

74. Jensen, J.R., 1996. *Introductory Digital Image Processing: A Remote Sensing Perspective.* 2nd Ed. New Jersey, USA: Pearson Prentice Hall.

75. Ji, C. Y., 2008. Haze reduction from the visible bands of LANDSAT TM and ETM+ images over a shallow water reef environment. *Remote Sensing of Environment*, 112, p.1773 – 1783.

76. Jia, X. & Richards, J. A., 1994. Efficient maximum likelihood classification for imaging spectrometer data sets. *IEEE Transactions on Geoscience and Remote Sensing*, 32(2), p.274 – 281.

77. Jose, A. T. A, Francisco G. R., Mercedes P. L. & Canton, M., 2003. An automatic cloud-masking system using backpro neural nets for AVHRR scenes. *IEEE Transactions on Geoscience and Remote Sensing*, 41(4), p.826 – 831.

78. Justice, C., J., Kendall, P., Dowty & Scholes, R. J., 1996. Satellite remote sensing of fires during SAFARI using the NOAA advanced very high resolution radiometer *Journal of Geophysical Research*, 101, p.23,851– 23,863.

79. Karlsson, K.-G., 1989. Development of an operational cloud classification model. *International Journal of Remote Sensing*, 10, p.687 – 693.

80. Kaufman, Y. J. & Fraser, R. S., 1983. Different atmospheric effects in remote sensing of uniform and nonuniform surfaces. *Advances in Space Research,* 2(5), p.147 – 155.

81. Kaufman, Y. J. & Fraser, R. S., 1984. Atmospheric effects on classification of finite fields. *Remote Sensing of Environment,* 15, p.95 – 118.

82. Kaufman, Y. J. & Sendra, C., 1988. Algorithm for automatic atmospheric corrections to visible and near-IR satellite imagery. *International Journal on Remote Sensing*, 9(8), p.1357 – 1381.

83. Kaufman, Y. J., 1987. Satellite Sensing of Aerosol Absorption. *Journal of Geophysical Research.* 92 (D4), p.4307 – 4317.

84. Kaufman, Y. J., Tanre, D., Gordon, H. R., Nakajima, T., Lenoble, J., Frouins R., Grassl, H., Herman, B. M., King M. D. and Teillet P. M., 1997. Passive remote sensing of tropospheric aerosol and atmospheric correction for the aerosol effect. *Journal of Geophysical Research*, 102(D14), p.16815 – 16830.

85. Kaufman, Y. J., and D. Tanre, 1996. Direct and indirect methods for correcting the aerosol effect on remote sensing. *Remote Sensing of Environment*, 55, p.65 – 79.

86. King, M., Kaufman, Y., Menzel, W. P. & Tanre, D., 1992. Remote sensing of cloud, aerosol and water vapor properties from the moderate resolution imaging spectrometer (MODIS). IEEE Trans. *IEEE Transactions on Geoscience and Remote Sensing*, 30(1), p.2 – 27.

87. Kneizys, F. X., Shettle, E. P., Abreu, L. W., Chetwynd, J. H., Anderson, G. P., Gallery, W. O., Selby, J. E. A. & Clough, S. A., 1988. *Users Guide to LOWTRAN 7, AFGL-TR-88-0177, (NTIS AD A206773).*

88. Kotchenova, S. Y. & Vermote, E. F., 2007. Validation of a vector version of the 6S radiative transfer code for atmospheric correction of satellite data. Part II: Homogeneous Lambertian and anisotropic surfaces, *Applied Optics*, 46(20), p.4455 – 4464.

89. Kotchenova, S. Y., Vermote, E. F., Matarrese, R. & Klemm Jr., F. J., 2006. Validation of a vector version of the 6S radiative transfer code for atmospheric correction of satellite data. Part I: Path Radiance, *Applied Optics*, 45(26), p.6726 – 6774.

90. Koukoulas, S. & Blackburn, G. A., 2001. Introducing new indices for accuracy evaluation of classified images representing semi-nat ural woodland environments. *Photogrammetric Engineering & Remote Sensing*, 67(4), p.499 – 510.

91. Kriebel, K. T., 1976. On the variability of the reflected radiation field due to differing distributions of the irradiation. *Remote Sensing of Environment*, 4, p.257 – 264.

92. Lo, C. P. & Choi, J., 2004. A hybrid approach to urban land use/cover mapping using Landsat 7 Enhanced Thematic Mapper Plus (ETM+) images. *International Journal of Remote Sensing*, 25(14), 2687 – 2700.

93. Loveland, T.R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L. & Merchant, J. W., 2000. Development of global land cover characteristics data base and IGBP DISCover from 1km AVHRR data. *International Journal of Remote Sensing*, 21, p.1303 – 1330.

94. Liang, S., Fang, H. & Chen, M., 2001. Atmospheric correction of Landsat ETM+ land surface imagery: Part I: Methods. *IEEE Transactions on Geoscience and Remote Sensing*, 39(11), p.2490 – 2498.

95. Liang, S., Fang, H. Hongliang Fang, Jeffrey, T., Chen, M. M., Shuey, C. J., Walthall, C. L. & Daughtry, C .S. T., 2002. Atmospheric correction of Landsat ETM+ land surface imagery: Part II: Validation and Applications. *IEEE Transactions on Geoscience and Remote Sensing*, 40(12), p.2736 – 2746.

96. Li, J., Menzel, W. P., Yang, Z., Frey, R. A. & Ackerman, S. A., 2003. High-spatial-resolution surface and cloud-type classification from MODIS multispectral band measurements. *Journal of Applied Meteorology*, 42, p.204 – 226.

97. Liu, X., Li, X., & Zhang, X., 2010. Determining class proportions within a pixel using a new mixed-label analysis method. *IEEE Transactions on Geoscience and Remote Sensing*, 48(4), p.1882 – 1891.

98. Liu, K., Shi, W & Zhang, H., 2011. A fuzzy topology-based maximum likelihood classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(1), 103 – 114.

99. Lillesand, T. M., Kiefer, R. W. & Chipman, J. W., 2004. *Remote Sensing and Image Interpretation.* 5th Ed. Hoboken, NJ, USA: John Wiley & Sons.

100. Logar, A. M., Lloyd, D. E., Corwin, E. M, Penaloza, M. L., Feind, R. E., Berendes, T. A., Kuo, K. & Welch, R. M., 1998. The ASTER polar cloud mask. *IEEE Transactions On Geoscience And Remote Sensing*, 36(4), p.1302 – 1312.

101. Longshore, R., Raimondi, P. & Lumpkin, M., 1976. Selection of detector peak wavelength for optimum infrared system performance. *Infrared Physics*, 16(6), p.639 – 647.

102. Lu, D., 2007. Detection and substitution of clouds/hazes and their cast shadows on IKONOS images. *International Journal of Remote Sensing*, 28(18), p.4027 – 4035.

103. Lu, D. & Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5), 823 – 870.

104. Lu, D., Mausel, P., Brondízio, E. & Moran, E., 2004. Relationships between forest stand parameters and Landsat TM spectral responses in the Brazilian Amazon Basin. *Forest Ecology and Management*, 198(1–3), p.149 – 1 67.

105. Lu, D., Mausel, P.., Brondizio, E. & Moran, E., 2002. Assessment of atmospheric correction methods for Landsat TM data applicable to Amazon basin LBA research. *International Journal on Remote Sensing*, 23(13), 2651 – 2671.

106. Lu, D., Moran, E. & Batistella, M., 2003. Linear mixture model applied to Amazonian vegetation classification. *Remote Sensing of Environment*, 87, p.456 – 469.

107. Luo, Y, Trishchenko, A.P. & Khlopenkov, K. V., 2008. Developing clear-sky, cloud and cloud shadow mask for producing clear-sky composites at 250-meter spatial resolution for the seven MODIS land bands over Canada and North America. *Remote Sensing Of Environment*, 112(12), p.4167 – 4185.

108. Malaysian Meteorological Services, 1997. *Monthly Abstract of Meteorological Observations, September 1997.* Kuala Lumpur, Malaysia: Malaysian Meteorological Services.

109. Mahmood, N.N., Loh, K.F. & Ahmad, S., 1997. Remote sensing research in Malaysia. *Proceedings of the IEEE International Geoscience and Remote Sensing, 1997. IGARSS '97. Remote Sensing - A Scientific Vision for Sustainable Development.*, 3, p.1418 – 1420.

110. Malaysian Prime Minister Office 2010. *Malaysian Prime Minister Office Website* (Updated 2010). Available at: http://www.pmo.gov.my [Accessed: 29 December 2010].

111. Mahmud, M., 2009. Mesoscale equatorial wind prediction in Southeast Asia during a haze episode of 2005. *Geofizika*, 26(1), p.67 – 84.

112. Markham, B. L., Storey, J. C., Williams, D. L. & Irons, J. R. 2004. Landsat sensor performance: history and current status. *IEEE Transaction on Geoscience and Remote Sensing.* 42 (12), p.2691 – 2694.

113. Mather, P. M., 2004. *Computer Processing of Remotely-sensed Images: An Introduction.* 3rd Ed. West Sussex, England: John Wiley & Sons.

114. Martins, J. V., Tanré, D., Remer, L., Kaufman, Y., Mattoo, S & Levy, R., 2002. MODIS Cloud screening for remote sensing of aerosols over oceans using spatial variability. *Geophysics Research Letter*, 29(12), p.8009 – 8012.

115. Mccormick, C. M., 1999. Mapping exotic vegetation in the everglades from large-scale aerial photographs. *Photogrammetric Engineering and Remote Sensing*, 65, p.179 – 184.

116. Meng, Q., Borders, B. E., Cieszewski, C. J. & Madden, M., 2009. Closest Spectral Fit for Removing Clouds and Cloud Shadows. *Photogrammetric Engineering & Remote Sensing*. 75(5), p.569 – 576.

117. Ministry of Forestry Indonesia 2010. *Ministry of Forestry Indonesia Website* (Updated September 2010). Available at: http://www.dephut.go.id/index.php?q=en [Accessed: 29 December 2010].

118. MODIS Characterization Support Team, 2006. *MODIS Level 1B Product User's Guide: For Level 1B Version 5.0.6 (Terra) and Version 5.0.7 (Aqua).* Greenbelt, MD, USA: NASA/Goddard Space Flight Center.

119. MODIS, 2007. *Components of MODIS* (Updated March 2007). Available at: http://modis.gsfc.nasa.gov/about/specifications.php [Accessed: 31 August 2007].

120. Isa, M. H. M., Zhao, X., & Yoshino, H., 2010. Preliminary study of passive cooling strategy using a combination of PCM and copper foam to increase thermal heat storage in building facade. *Sustainability*, 2(8), p.2365 – 2381.

121. Molenar, V., Malm, W.C. & Johnson, C. E., 1994. Visual air quality simulation techniques. *Atmospheric Environment*, 28, p.1055 – 1063.

122. Moro, G. D. & Halounova , L, 2007. Haze removal for high-resolution satellite data: a case study. *International Journal on Remote Sensing*, 28(10), p.2187 – 2205.

123. Morris, W., 1975. *The Heritage Illustrated Dictionary of English Language of The English Language*. New York : American Heritage Publishing Co.

124. Mountrakis, G., Im, J. & Ogole, C., 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247 – 259.

125. NASA, 2007. *Level 1 and Atmosphere Archive and Distribution System (LAADS)* (Updated October 2007). Available at: http://ladsweb.nascom.nasa.gov [Accessed: 16 January 2008].

126. Oakley, J. P. & Satherley, B. L., 1998. Improving image quality in poor visibility conditions using a physical model for contrast degradation, *IEEE Transactions on Image Processing*, 7(2), p.167 – 179.

127. Otukei, J. R. & Blaschke, T., 2010. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms.

International Journal of Applied Earth Observation and Geoinformation, 12(1), S27 – S31.

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

129. Radojevic, M., 2003. Chemistry of Forest Fires and Regional Haze with Emphasis on Southeast Asia. *Pure and Applied Geophysics*, 160(1), 157 – 187.

130. Rashed, T., Weeks, J. R., Gadalla, M. S. & Hill, A. G., 2001, Revealing the anatomy of cities through spectral mixture analysis of multispectral satellite imagery: a case study of the Greater Cairo region, Egypt. *Geocarto International*, 16, p.5 – 15.

131. Remer, L. A., Kaufman, Y. J. & Holben, B. N., 1996. The size distribution of ambient aerosol particles: Smoke versus urban/industrial aerosol, *Global Biomass Burning*. Cambridge, MA : MIT Press.

132. Richards, J. A., 1999. *Remote sensing digital image analysis: An introduction.* Berlin, Germany: Springer-Verlag.

133. Richter, R., 2008. Classification Metrics for Improved Atmospheric Correction of Multispectral VNIR Imagery. *Sensors*, 8, p.6999 – 7011.

134. Rossow, W.B. & Schiffer, R.A., 1999. Advances in understanding clouds from ISCCP. *Bulletin of American Meteorological Societies*, 80, p.2261 – 2288.

135. Rossow, W. B. & Gardner, L. C., 1993. Cloud detection using satellite measurement of infrared and visible radiances for ISCCP. *Journal of Climate*, 6, p.2341 – 2369.

136. Rossow, W. B., Walker, A.W. & Gardner, L. C., 1993. Comparison of ISCCP and other cloud amounts. *Journal of Climate*, 6, p.2394 – 2418.

137. Rozenstein, O., & Karnieli, A., 2011. Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. *Applied Geography*, 31(2), p.533 – 544.

138. San Miguel-Ayanz, J. & Biging, G. S., 1996. An iterative classification approach for mapping natural resources from satellite imagery. *International Journal of Remote Sensing*, 17, p.957 – 982.

139. Saunders, R. W. & Kriebel, K. T., 1988. An improved method for detecting clear sky and cloudy radiances from AVHRR data. *International Journal of Remote Sensing*, 9, p.123 –150.

140. Saunders, R.W., 1986. An automated scheme for the removal of cloud contamination form AVHRR radiances over Western Europe. *International Journal of Remote Sensing*, 7, p.867 – 886.

141. Scepan, J., 1999. Thematic validation of high-resolution global land-cover data sets. *Photogrammetric Engineering and Remote Sensing*, 65, p.1051 – 1060.

142. Schott, C. S., Salvaggio, C. & Volchok, W., 1988. Radiometric scene normalization using pseudoinvariant features. *Remote Sensing of Environment*, 26, p.1 – 16.

143. Song, C., Woodcock, C. E., Seto, K. C., Lenney, M. P., & Comber, S. A., 2001. Classification and change detection using Landsat TM data: When and how to correct atmospheric effects?. *Remote Sensing of Environment*, 75, p.230 – 244.

144. Stehman, S. V., 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62, p.77 – 89.

145. Stehman, S. V., 2000. Practical Implications of Design-Based Sampling Inference for Thematic Map Accuracy Assessment. *Remote Sensing of Environment*, 72,(1), p.35 – 45.

146. Stehman, S. V., Arora, M. K., Kasetkasem, T. & Varshney, P. K., 2007. Estimation of Fuzzy Error Matrix Accuracy Measures Under Stratified Random Sampling. *Photogrammetric Engineering & Remote Sensing*, 73(2), p.165 – 173.

147. Stowe, L. L., Davis, P. A. & McClain, E. P., 1999. Scientific basis and initial evaluation of CLAVRI global clear/cloud classification algorithm for the advanced very high resolution radiometer. *Journal of Atmospheric and Oceanic Technologies*, 16, p.656 – 681.

148. Swain, P. H. & Davis, S. M., 1978. *Remote sensing: The quantitative approach.* New York: McGraw-Hill.

149. Swap, R. J., et al. (2002), The Southern African Regional Science Initiative (SAFARI 2000) dry-season field campaign: An overview. S. Afr. J. Sci.,98, 125–130.

150. Simeh, A. & Ahmad, T. M A. T., 2001. The case study on the Malaysian palm oil, *Regional Workshop On Commodity Export Diversification And Poverty Reduction In South And South-East Asia(Bangkok, 3-5 April, 2001) Organized By UNCTAD In Cooperation With ESCAP And MARDI.*

151. Smith, G. M. & Fuller, R. M., 2001. An integrated approach to land cover classification: an example in the Island of Jersey. *International Journal of Remote Sensing*, 22, p.3123 – 3142.

152. Stehman, S. V., 1999. Basic probability sampling designs for thematic map accuracy assessment. *International Journal of Remote Sensing*, 20(12), p.2423 – 2441.

153. Story, M. & Congalton, R., 1986. Accuracy assessment: a user's perspective. *Photogrammetric Engineering and Remote Sensing,* 52, p.397 – 399.

154. Suzuki, K., 2005. *Plantation Technology in Tropical Forest Science*. Tokyo : Springer-verlag.

155. Thenkabail, P.S., Schull, M., Turral, H., 2005. Ganges and Indus river basin Land Use/Land Cover (LULC) and irrigated area mapping using continuous streams of MODIS data. *Remote Sensing of Environment* 95, 317–341.

156. The Star Online, 2005. *Open burning in Klang Valley banned* (Updated 9 August 2005). Available at: http://www.thestar.com.my/news/story.asp?file=/2005/8/9/nation/11717460&sec=nation [Accessed: 13 September 2010].

157. Thomlinson, J. R., Bolstad, P. V. & Cohen, W. B., 1999. Coordinating methodologies for scaling landcover classifications from site-specific to global: steps toward validating global map products. *Remote Sensing of Environment*, 70, p.16 – 28.

158. Thomson, G., Fuller, R. M. & Eastwood, J. A., 1998. Supervised versus unsupervised methods for classification of coasts and river corridors from

airborne remote sensing. *International Journal of Remote Sensing*, 19(17), 3423 - 3431.

159. USGS, 2010. *USGS Website* (Updated December 2009). Available at: http://landsat.usgs.gov/about_landsat5.php [Accessed: 7 September 2010].

160. Vallero, D., 2008. *Fundamentals of air pollution*. USA : Elsevier.

161. Vermote, E. F., Tanre, D., Deuze, Herman, M. and Morcrette, J., 1997. Second simulation of the satellite signal in the solar spectrum, 6S: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 35, p.675 – 686.

162. Wang, L., Sousa, W. P., Gong, P. & Biging, G. S., 2004. Comparison of IKONOS and QuickBird images for mapping mangrove species on the Caribbean coast of Panama. *Remote Sensing of Environment*, 91, p.432 – 440.

163. Wang, Z., Jensen, J. R. & Jungho, I. M., 2010. An automatic region-based image segmentation algorithm for remote sensing applications. *Environmental Modelling & Software*, 25, 1149 – 1165.

164. Warren, S. G., & Hahn, C. J., 2002: Cloud climatology. *Encyclopedia of Atmospheric Sciences (Eds. Holton J. R., Curry, J. A. & Pyle, J. A.)*, p. 476 – 483. London: Academic Press.

165. Weather Forecast Office, 2011. *National Weather Service Website* (Updated March 2011). Available at: http://www.crh.noaa.gov/lmk/?n=cloud_classification [Accessed: 30 July 2011].

166. Wilkinson, G. G., 2005. Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3), p.433 – 440.

167. Wulder, M. A., Franklin, S. E., White, J. C., Linke, J. & Magnussen, S.. 2006. An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data. *International Journal of Remote Sensing*, 27, p.663 – 683.

168. Wylie, D., Menzel, W. P., & Strabala, K. I., 1994, Four years of global cirrus cloud statistics using HIRS. *Journal of Climate*, 7, p.1972 – 1986.

169. Yoshida, T. & Omatu, S., 1994. Neural network approach to land cover mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5), p.1103 – 1109.

170. Yusoff, N. M., Ahmad, S. A. & Othman, M., 2002. Utilization of TiungSAT-1 data for environmental assessment and monitoring. *Canadian International Development Agency (CIDA)-Canadian Space Agency (CSA) Conference, Space Applications for Sustainable Development*, Hull, Canada, May 21-22, 2002.

171. Zhang, J. & Foody G. M., 2001. Fully-fuzzy supervised classification of sub-urban land cover from remotely sensed imagery: statistical and artificial neural network approaches. *International Journal of Remote Sensing*, 22(4), 615 – 628.

172. Zhang, G. P., 2000. Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, And Cybernetics—Part C: Applications And Reviews*, 30(4), p. 451 – 462.

173. Zhang, J. & Kirby R. P., 1999. Alternative criteria for defining fuzzy boundaries based on fuzzy classification of aerial photographs and satellite images. *Photogrammetric Engineering and Remote Sensing*, 65(12), 1379 – 1387.

174. Zhang, Y., Guindon, B. & Cihlar, J., 2002. An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images. *Remote Sensing of Environment*, 82, p.173 – 187.