UNIVERSITY OF CALIFORNIA

Los Angeles

Bayesian Network Application to Satellite Image Classification

for Stormwater Management

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Civil Engineering

by

Mi-Hyun Park

2004

UMI Number: 3164380

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 3164380

Copyright 2005 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

© Copyright by

Mi-Hyun Park

2004
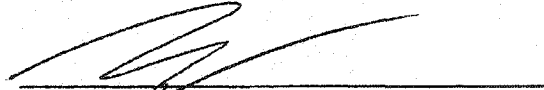
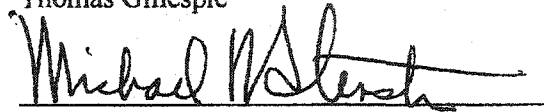The dissertation of Mi-Hyun Park is approved.

_____
Keith D. Stolzenbach

_____
Steven A. Margulis

_____
Thomas Gillespie

_____
Michael K. Stenstrom, Committee Chair

University of California, Los Angeles

2004

ii

TABLE OF CONTENTS

v

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my supervisor, Professor Michael. K. Stenstrom for his support and encouragement throughout my study at UCLA. He was truly open-minded in every aspects and conversation with him was always enjoyable. I would like to show my appreciation to the committee members, Professors Keith Stolzenbach, Steve Margulis in Department of Civil and Environmental Engineering and Thomas Gillespie in Department of Geography for their valuable advice and inspiration during my research.

Special recognition is due Dr. Doug Alsdorf in Department of Geography and now in Department of Geological Sciences at Ohio State University for his advice and help. His consistent research enthusiasm and concern about me will not be forgotten. I should thank Professor Judea Pearl in Department of Computer Science for his guidance on Bayesian networks.

My deepest thanks go to Dr. Duncan F. Gillies in Department of Computing at Imperial College for his comments and suggestion on Bayesian networks. I should thank Professors Nigel Graham and David Butler in Department of Civil and Environmental Engineering at Imperial College for their kindness and advice on Environmental Engineering. I'm also grateful to Professor Hang-Sik Shin at Korea Advanced Institute of Science and Technology for his encouragement and support.

Above all, I would like to thank my family for their belief in me. Their love and dedication made me come through this far. I am truly indebted them for everything.

ix

VITA

April 18, 1973	Born, Seoul, Korea
1996	B.E. Environmental Engineering, with Distinction Inha University Incheon, Korea
1996-1998	Korean Government Scholarship Awardee Korea Advanced Institute of Science and Technology Daejeon, Korea
1997-1998	British Government Chevening Scholarship Awardee Imperial College London, UK
1998	M.S., DIC., Civil and Environmental Engineering Imperial College London, UK
1999	M.S., Civil and Environmental Engineering Korea Advanced Institute of Science and Technology Daejeon, Korea
2001	M.S., DIC., Computing, with Distinction Imperial College London, UK
2001-2002	Department Fellow Department of Civil and Environmental Engineering University of California, Los Angeles
2001-2002	Teaching Assistant Department of Civil and Environmental Engineering University of California, Los Angeles
2002-2004	Research Assistant Department of Civil and Environmental Engineering University of California, Los Angeles
2003-2004	Dissertation Fellow Department of Civil and Environmental Engineering University of California, Los Angeles

x

PUBLICATIONS AND PRESENTATIONS

Park, M. and Stenstrom, M.K. (October, 2004) A New Classification System for Urban Stormwater Pollutant Loading – A Case Study in Santa Monica Bay Area, 8th International Conference on Diffuse/Nonpoint Pollution, Kyoto, Japan, accepted

Park, M. and Stenstrom, M.K. (October, 2004) Spatial Estimates of Stormwater Pollutant Loading using Bayesian Networks, Water Environment Federation's Technical Exhibition and Conference, New Orleans, Louisiana, accepted

Park, M. and Stenstrom, M.K. (September, 2004) Identifying Main Area Generating Stormwater Pollution from Satellite Imagery and Ancillary Data, The International Geoscience and Remote Sensing Symposium, Anchorage Alaska, accepted

Park, M. and Stenstrom, M.K., (August, 2004) Estimation of Impervious Surface Area Using Bayesian Networks, AGU 2004 Western Pacific Geophysics Meeting, August 16-20 2004 in Honolulu, Hawaii

Park, M. and Stenstrom, M.K., (August, 2003) Landuse Classification for Stormwater Modeling Using Bayesian Networks, In Proceedings of the 7th International Specialised IWA Conference, Diffuse Pollution and Basin Management, Dublin, Ireland

ABSTRACT OF THE DISSERTATION

Bayesian Network Application to Satellite Image Classification

for Stormwater Management

by

Mi-Hyun Park

Doctor of Philosophy in Civil Engineering

University of California, Los Angeles, 2004

Professor Michael K. Stenstrom, Chair

Proper management of stormwater runoff is required to protect receiving water quality since most wastewater sources have been treated to secondary standards or beyond. Urban stormwater runoff has become the primary source of many pollutants, which is caused by runoff from highly developed, impervious land use, and managing stormwater has become the primary objective of new regulatory efforts. However, monitoring and modeling is inherently difficult, and empirical methods to estimate stormwater pollution have been developed using land use data. Conventional land use data collecting methods from ground surveys are expensive and time consuming, and may not be available.

xii

This study demonstrated alternative approaches to use satellite image classification using Bayesian networks. The Bayesian network structure shows the most influential input variables for classification. The network also reveals the relationships among variables, which is useful when dealing with missing information. First, urban land use in the given area was classified and converted to corresponding stormwater pollutant loading maps based on the existing relationships between land use and the pollutant loads. Secondly, the pollutant loads for each water quality parameter were estimated directly from satellite imagery. The resulting thematic maps spatially estimated stormwater pollutant loadings, which identified areas generating high stormwater pollutant emissions. The results show that stormwater pollutants are highly correlated to impervious areas because of their high runoff coefficients, even when they had low event mean concentrations. These results are useful in developing best management strategies for stormwater pollution and in establishing total maximum daily loads in the watershed.

Chapter I

Introduction

1.1 Problem Definition and Motivation

Non-point source pollution including stormwater runoff has been recognized as the major source of many contaminants in urban areas, since most point sources have been controlled to reduce their impact on receiving waters (EPA, 1994; Arnold and Gibbons, 1996; Sleavin *et al.*, 2000; Brabec *et al.*, 2002; Ackerman and Schiff, 2003). Impervious land uses in urban areas have impaired receiving water quality by increasing stormwater runoff quantity and deteriorating the runoff quality. Therefore, recent efforts have focused on stormwater runoff in order to mitigate its adverse impact on receiving waters.

1

Environmental regulations ensure that the stormwater runoff impacts on receiving waters are acceptable. For example, urban stormwater runoff in the United States is subject to the National Pollutant Discharge Elimination Program (NPDES) permit program and should meet water quality standards (Swamikannu *et al.*, 2003). For this purpose, best management practices (BMPs) have been implemented to reduce the stormwater pollution. Clean Water Act Amendments require regulatory agencies and local governments to develop total maximum daily loads (TMDLs) to properly allocate stormwater pollution loads. However, in many regions, development and approval of TMDLs are still in progress. In order to properly develop BMPs and TMDLs, monitoring and modeling of stormwater runoff is important.

Monitoring stormwater runoff and estimating its pollutant load to receiving waters are inherently difficult due to the uncertain temporal and spatial characteristics of the domain (Wong *et al.*, 1997). The difficulty is associated with large areas, many emission points, and sampling episodic storm events. Moreover, monitoring and modeling stormwater runoff demand expertise and accumulated stormwater quality and rainfall data, but event and site specific data measurements are often not available (Vaze and Chiew, 2003).

An alternative approach to modeling stormwater pollutant loads is using land use data (Stenstrom *et al.*, 1984; Chiew and McMahon, 1999) since urban stormwater pollution is related to land use activities (Sliva and Williams, 2001; Burian *et al.* 2002; Tong and Chen, 2002). Therefore, urban land use information is important in properly managing stormwater pollution. However, conventional land use data collection methods

2

such as ground surveys can be time consuming and resource intensive. Generally, these land use collection methods are applicable only to small areas. Furthermore, the land use data are not available for many areas and not optimized for environmental purpose.

New approaches are being developed to estimate land use from satellite imagery. Satellite imagery is inexpensive compared to ground surveys and provides temporal coverage. The quality of satellite imagery has been improved with advanced technology, which makes satellite imagery more reliable. Another advantage of satellite imagery is that it is inherently digital and readily lends itself to computer methodologies (Richards and Jia, 1997). In addition, the spectral signatures of satellite images can provide meaningful environmental information.

Land use classification with satellite imagery has been explored to assist urban planning and environment management. For instance, Landsat series imagery has been extensively used for land cover/use analysis since the first satellite was launched in 1972 (Haack, 1983; Haack *et al.*, 1987; Pax-Lenney, 1997; Ridd and Liu, 1998; Stefanov, 2001; Clapham, 2003; Wilson *et al.*, 2003). Satellite imagery has been examined with various methods such as statistical methods and artificial intelligence (AI) applications (Benediktsson *et al.*, 1990; Civco, 1991; Short, 1991; Bischof *et al.*, 1992; Foody *et al.*, 1992; Heermann and Khazenie, 1992; Kanellopoulos *et al.*, 1993; Li *et al.*, 1993; Paola and Schowengerdt, 1995). The last two decades have seen the increased use of AI techniques, i.e. decision trees and neural networks, with varying degrees of success (Foody, 1995; Paola and Schowengerdt, 1995; Carpenter *et al.*, 1999; Gopal *et al.*, 1999; Stefanov *et al.*, 2001; Pal and Mather, 2003). Bayesian networks, one of the AI

3

techniques, are principled and efficient probabilistic approaches to reasoning and inference. A Bayesian network provides dependency relationship between variables with graphical representation. This helps us to easily understand those relationships between variables in a given domain. Bayesian networks are also known as excellent diagnostic and predictive tools in many areas (Charniak, 1991; Sucar and Gillies, 1994; Russell and Norvig, 1995; Lucas and Abu-Hanna, 1999; Bang and Gillies, 2002; Zweig, 2003). However, Bayesian networks have received little attention in remote sensing image classification since it is relatively new AI technique. Only recently have researchers attempted to adopt Bayesian networks for satellite image classification (Pal *et al.*, 2001; Park and Stenstrom, 2003).

1.2 Objectives

In this research, we evaluated the performance of Bayesian networks for satellite image classification. We gained an understanding of the nature of classification from the network structure by identifying inputs that are important for pixel class labels. Furthermore, we compared different network structures to find the optimal network for the given domain. Once an optimal network for classification was obtained, we examined the effect of incorporating ancillary input data.

The main goal was to generate spatial thematic maps of pollutant loads from stormwater since few attempts have been made to spatially estimate the pollutant loads. This research proposed the use of satellite imagery to estimate stormwater pollutant loads

4

in a typical urban area. Estimating stormwater pollutant loads from a given area is important to develop best management strategies. First, we classified the satellite imagery of urban land use with Bayesian networks. The urban land use information in the given area can provide related information such as runoff coefficients associated with imperviousness, and event mean concentrations (EMCs) of a particular pollutant, which are necessary for calculating pollutant loads.

After estimating pollutant loads from satellite land use classification, we explored an alternate approach that estimated stormwater pollutant loads directly from satellite imagery, which did not require land use or EMC thematic maps as intermediate processes. To facilitate this task, Bayesian networks were also used to predict stormwater pollutant loads from spectral signatures of satellite imagery.

For both approaches, we aim to identify the areas that generate high pollutant loads into receiving waters. Identifying areas contributing to high pollutant loads is useful in developing BMPs and establishing TMDLs for stormwater pollutants. Finally, a thematic map was used to develop a new classification system for stormwater management purposes.

Overall, we hope this dissertation provides improved tools for understanding and managing stormwater pollution using satellite imagery, and provides guidelines for future watershed planners and regulators.

5

1.3 Organization of the dissertation

This paper is organized in the following way. In Chapter 2, we review the general remote sensing system and existing classification algorithms that provide current state of art. Chapter 3 presents Bayesian networks as an alternative classification algorithm in detail. In Chapter 4, we review the existing stormwater pollution models. Chapter 5 describes the study area and data acquisition. In Chapter 6, urban land use classification is explored using Bayesian networks from satellite imagery. In Chapter 7, we conduct Bayesian network classification of stormwater pollutant load from satellite imagery. Finally we present conclusion and propose our future study plan in Chapter 8 and 9, respectively.

Chapter II

Remote Sensing and Image Processing

Remotely sensed data such as satellite imagery have been extensively used for many areas such as environment monitoring, global change detection, agriculture, natural resource management, weather forecasting, mapping, and military intelligence (Jensen, 1996; Sabins, 1997; Schowengerdt, 1997; Mather, 1999; Jensen, 2001). However, the use of remotely sensed data demands intensive statistical and computational knowledge for proper digital image processing.

In this chapter, we review the basic concept of the remote sensing and the existing processing methods. In section 2.1, we introduce the basic terminology related to remote sensing. In section 2.2, we explain Landsat system, one of the earliest and widely used satellite image data. In section 2.3, conventional image classification methods are reviewed. In section 2.4, incorporating ancillary data to assist classification is presented.

7

In section 2.5, we explain image transformation to enhance image interpretation and classification accuracy. In section 2.6, accuracy assessment methods are discussed.

2.1 Basic Concepts of Remote Sensing

Remote sensing process detects electromagnetic (EM) energy from an interaction between incident radiation and target surfaces. The electromagnetic spectrum ranges vary and particular EM spectrum is useful for different purpose. Visible spectrum, blue, green, and red, is ranged from approximately 0.4 to 0.7 µm. Infrared (IR) consists of two regions: reflected IR covers wavelengths from approximately 0.7 to 3.0 µm and thermal IR region covers wavelengths from approximately 3.0 to 100 µm. Microwave, ranged from about 1 mm to 1 m, covers the longest wavelengths. Figure 2.1 presents different range of wavelength used in remote sensing.

The sensor detects the EM energy using either passive or active systems (Jensen, 2000). Passive systems detect naturally occurring radiation from the target, i.e. solar radiation, where as active systems emit radiation and detect its backscattering, i.e. radar. Passive systems can detect energy only during daytime when the solar energy is available. However, thermal infrared can be detected either day or night. Conversely, active sensors can detect EM energy anytime since they provide their own energy source.

Each remote sensing system associates four major resolutions: spatial, spectral, radiometric, and temporal resolution (Jensen, 1996; Sabins, 1997; Mather, 1999). The spatial resolution of the sensor is the size of the smallest feature that can be detected,

8
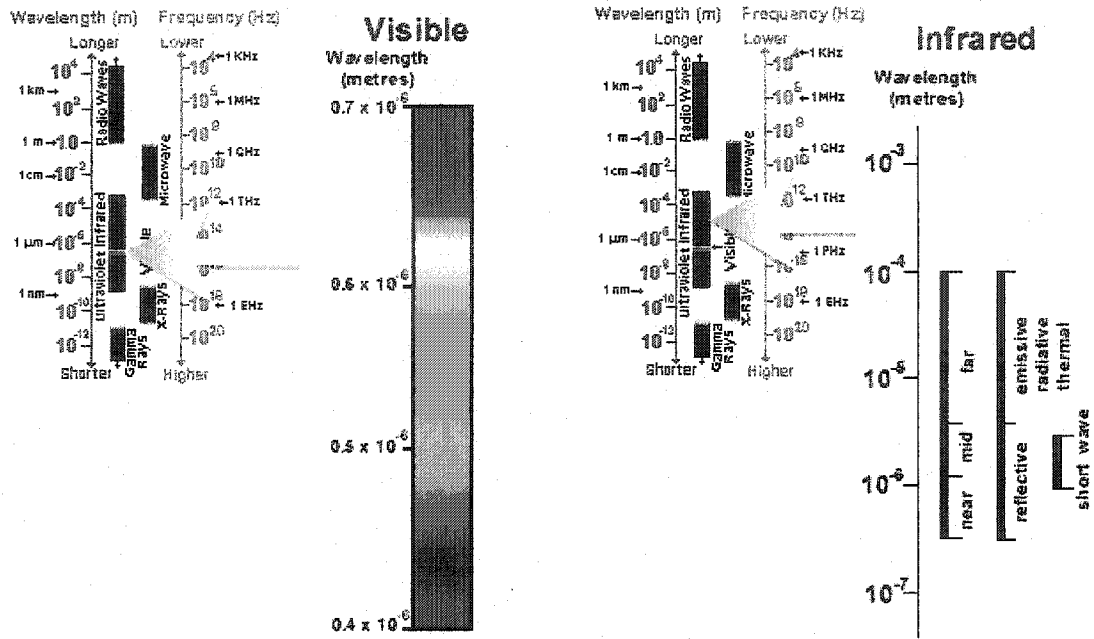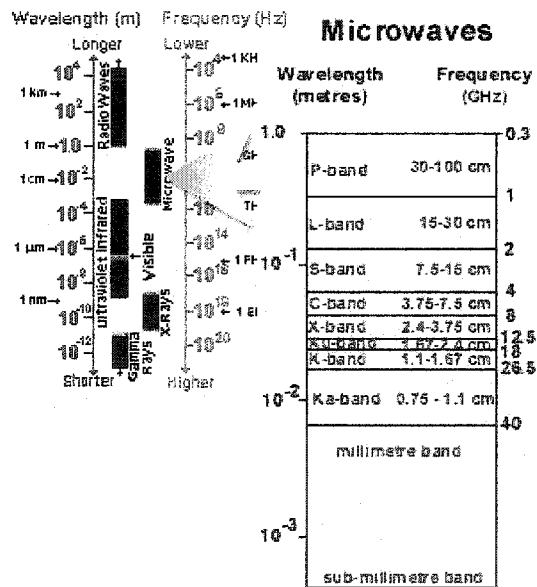
Figure 2.1 Electromagnetic spectrum (a) visible (b) infrared (c) microwave
(adopted from CCRS, http://www.ccrs.nrcan.gc.ca/ccrs/eduref/tutorial/tutore.html)

9

which depends on Instantaneous Filed of View (IFOV). The IFOV is the angular cone of visibility of the sensor and determines the area on the Earth's surface. Low spatial resolution images contain large features whereas high resolution images can detect small objects. The spectral resolution is the ability of a sensor to define wavelength intervals. Higher spectral resolution has narrower wavelength range for a particular band. The radiometric resolution is its ability to discriminate EM energy from the target surface. Higher radiometric resolution detects a more detail in EM energy. Temporal resolution is the revisiting period. The temporal resolution of a sensor depends on several factors including the satellite capabilities, the swath overlap, and latitude.

2.2 Landsat Systems

The Landsat series is the first satellite system for land observation (Jensen, 2000). The data from Landsat series have been applied to source management, mapping, environmental monitoring, and change detection (CCRS, 2002). Tables 2.1 and 2.2 show the background information of Landsat satellite series and the resolutions of each sensor system. Each Landsat system is illustrated in Figure 2.2.

The National Aeronautics and Space Administration (NASA) launched Landsat 1, 2 and 3 satellites in 1972, 1975, and 1978, respectively (Schowengerdt, 1997; Mather, 1999). These three Landsat satellites were also called Earth Resources Technology

10

Table 2.1 History of the Landsat satellite series

Satellite	Date launched	Date retired	Sensor
Landsat 1	7/23/1972	1/06/1978	MSS
Landsat 2	1/22/1975	2/25/1982	MSS
Landsat 3	3/05/1978	3/31/1983	MSS
Landsat 4	7/16/1982	standby mode	MSS, TM
Landsat 5	3/01/1984	operational	MSS, TM
Landsat 6	10/05/1993	lost at launch	MSS, ETM
Landsat 7	4/15/1999	operational	ETM$^+$

(adopted from Jensen, 2000; Lillesand *et al.*, 2004)

Table 2.2 Resolutions of Landsat sensor system

Sensor	Band	Spectral Resolution	Spatial Resolution	Temporal Resolution	Radiometric Resolution
MSS	4(green)	0.5 - 0.6 µm			
	5(red)	0.6 - 0.7 µm	79 × 79 m	18 days	7 bits
	6(near IR)	0.7 - 0.8 µm			
	7(near IR)	0.8 - 1.1 µm			6 bits
TM	1(blue)	0.45 - 0.52 µm			
	2(green)	0.52 - 0.60 µm			
	3(red)	0.63 - 0.69 µm	30 × 30 m		
	4(near IR)	0.76 - 0.90 µm		16 days	8 bits
	5(IR)	1.55 - 1.75 µm			
	6(thermal IR)	10.4 - 12.5 µm	120 × 120 m		
	7(IR)	2.08 - 2.35 µm	30 × 30 m		
ETM$^+$	1(blue)	0.45 - 0.52 µm			
	2(green)	0.52 - 0.60 µm			
	3(red)	0.63 - 0.69 µm	30 × 30 m		
	4(near IR)	0.76 - 0.90 µm		16 days	8 bits
	5(IR)	1.55 - 1.75 µm			
	6(thermal IR)	10.4 - 12.5 µm	60 × 60 m		
	7(IR)	2.08 - 2.35 µm	30 × 30 m		
	8(pan)	0.52 - 0.90 µm	15 × 15 m		

11

(a)

(b)

(c)

Figure 2.2 Landsat systems (a) Landsat 1,2,3 with MSS (b) Landsat 4,5 with TM
(c) Landsat 7 with ETM$^+$
(adopted from USGS landsat7.usgs.gov/ history.html)

Satellite (ERTS) series (Lillesand *et al*, 2004). The satellites operated in a sun-synchronous near-polar orbit at the altitude of 919 km. The sensor placed on the satellites was a multispectral scanner (MSS). The MSS collected data over a swath width of approximately 185 km that facilitates regional interpretation (Sabins, 1997). The MSS recorded the EM energy in four spectral bands with a spatial resolution of 79 m. Each pixel contains 6 bit data information with a range of digital numbers (DN) from 0 to 63 and the first three band data are rescaled to 7 bits from 0 to 127 (Sabins, 1997; Richards and Jia, 1999).

Landsat 4 and 5, launched in 1982 and 1984 respectively, are the second generation of Landsat series employing Thematic Mapper (TM) sensor systems. The TM sensor has seven bands such as visible, near infrared, two middle infrared and thermal infrared: Spatial resolution is 30 m for all bands except the thermal infrared band, which has 120 m resolution. All bands are recorded in 8 bits over a range of 256 DNs (Schowengerdt, 1997; Sabins, 1997; Richards and Jia, 1999; Mather, 1999). Improvements of the TM sensor over the MSS are longer dwell time and higher signal-to-noise ratio since TM sensor records data on both eastbound and westbound sweeps of the mirror (Sabins, 1997).

Landsat 6 was launched in 1993 but unfortunately failed to achieve orbit. Landsat 7 was launched in 1999 and also carried Enhanced Thematic Mapper Plus (ETM+) (Sabins, 1997; Goward *et al.*, 2001). ETM+ bands from 1 to 5 and 7 are identical to those of TM with the spatial resolution of 30 m but the thermal infrared band has higher spatial

13

resolution of 60 m. In addition, there is new panchromatic band that is a single broad spectral band with spatial resolution of 15 m (Richards and Jia, 1999; Mather, 1999).

2.3 Image Classification

Digital image classification assigns each pixel of the image to a particular class using spectral signature (CCRS, 2002). Classification procedures can be divided into two categories: supervised classification and unsupervised classification (Jensen, 1996; Schowengerdt, 1997; Richards and Jia, 1999; Mather, 1999). In a supervised classification, the analyst selects training data samples that represent different information classes. The selection of appropriate training data requires prior knowledge of the area of interest and each class of interest. The spectral signatures in each band are used to train the classification algorithm to recognize spectral similarity for each class and to assign a class to each pixel. In unsupervised classification, pixels with similar spectral signatures are grouped first and then labeled to a particular class by the analyst. In this case, the analyst specifies the number of classes prior to classification.

2.3.1 Supervised Classification

Supervised classification is most often used for remote sensing image data. In supervised classification, the analyst can control selected classes for a specific purpose and geographic region. Figure 2.3 shows a general process of supervised classification.

14
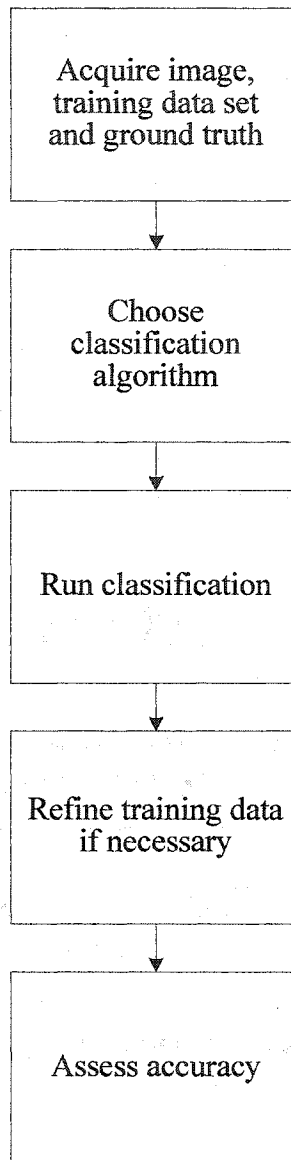
Figure 2.3 General steps of supervised classification

15

The accuracy of supervised classification depends on the size and the selection of training data. The training data size must be sufficient to properly represent each class and to provide acceptable accuracy. The size is often related to the number of bands and the statistical properties. Increase in sample size improves the classification accuracy (Dobbertin *et al.*, 1996). The training data size can be estimated from the statistical techniques. For example, Fitzpatrick-Lins (Fitzpatrick-Lins, 1981; Jensen, 1996) suggested the following binomial formula:

$$N = \frac{4 \times p(100 - p)}{\varepsilon^2} \qquad (2.1)$$

where N is a training data size, p is the expected accuracy (%) and ε is the allowable error. For example, with 85% of expected accuracy and 1% of allowable error, the required training data size becomes 5,100. However this provides only the total number of training data not the training data size for each class (Jensen, 1996). The rule of thumb that Congalton (1991) proposed is to collect a minimum of 50 data points for each class. Mather (1999) suggested the following formula as a minimum number of training data points:

$$N = 30 \times n \times c \qquad (2.2)$$

16

where n is the number of spectral bands and c is the number of classes. This equation

agrees with others (Swain, 1978; Foody *et al.*, 1995) who found that training data size for

each class should be at least 30 × n to form representative training samples. For example,

if we have 7 spectral bands and 5 classes, at least 1,050 training observations are needed.

The selection of sampling is also important for a reliable accuracy (Stehman *et*

al., 1998). A simple random sampling may be appropriate if the sample size is large

enough to represent all classes. However, random sampling may be impractical due to

undersampling small classes (Jensen 1996). It is also difficult to use randomly located

sites to assess the accuracy over a very large area (Foody, 2002). Alternative sampling,

stratified sampling, selects a minimum number of samples from each class. A

combination of random and stratified sampling can provide a balance between statistical

and practical validities. It is also important to select homogeneous samples representing

each class (Schowengerdt, 1997).

There are many supervised classification algorithms. In this section we discuss

typical supervised classification methods: parallelepiped, minimum distance, maximum

likelihood, and artificial neural networks.

Parallelepiped Classification

Parallelepiped is a very simple classification method (Jensen, 1996; Richards and

Jia, 1999; Mather, 1999). It classifies every pixel within a specified range, which is a

rectangle or parallelepiped. Figure 2.4 illustrates parallelepiped classification in a two-

17
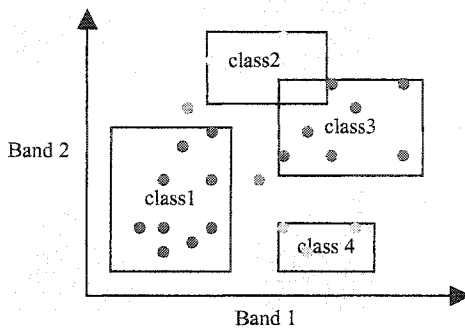
Figure 2.4 An example of parallelepipeds in two-dimensional spectral space
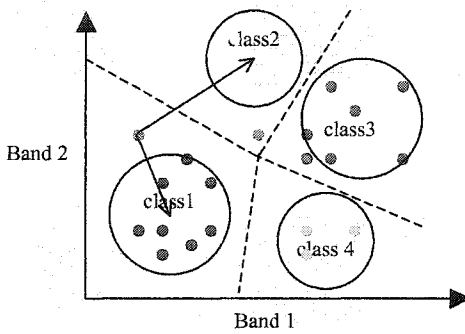
Figure 2.5 An example of minimum distance in two-dimensional spectral space
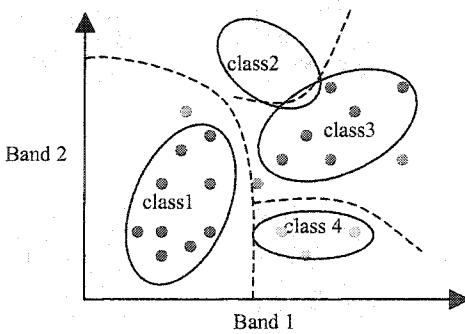
Figure 2.6 An example of maximum likelihood in two-dimensional spectral space

18

dimensional spectral space. The upper and lower bounds of the range are determined by training data statistics: the minimum and maximum DN values of each band; the mean ± standard deviations of each band; or any limits based on prior knowledge of the data and signatures. If the value of a pixel is between the bounds, the pixel is assigned to the particular class. The following equation provides the decision rule of classification:

$$\text{Decide } x \text{ is in class } \omega_i \text{ if } L_i \leq x \leq H_i \qquad (2.3)$$

where x is pixel values in each band, ω_i, $i = 1, \ldots M$, are the spectral classes and M is the total number of classes. L_i and H_i are low and high bounds respectively.

Parallelepiped classification is the fastest algorithm (Richards and Jia, 1999). However, it is not popular because of its several drawbacks. Pixels that do not belong to any parallelepiped will be unclassified. Some correlated data in the overlap of the parallelepipeds cannot be separated. In this case, the pixel is classified as the first class that passes the bounds tests, which makes the classification order dependent.

Minimum Distance Classification

Minimum distance classification assigns all pixels to the nearest class (Jensen, 1996; Richards and Jia, 1999; Mather, 1999). The means of each class are calculated from training data and determine the centers of the class. Then the Euclidean distance from a pixel to the center in each class is calculated by:

19

$$D(x, \mu_i) = \sqrt{(x_k - \mu_{ik})^2 + (x_1 + \mu_{il})^2} \qquad (2.4)$$

where x is pixel values in each band and μ_i, i = 1, ... M, are means of each classes. This gives the following decision rule of classification:

Decide x is in class ω_i if $D(x, \mu_i) < D(x, \mu_j)$ for all $j \neq i$. $\qquad$ (2.5)

This classification is slower than a parallelepiped classification (Richards and Jia, 1999). The classification uses only the mean of each class and the boundaries between each class are linear, which reduce the flexibility of the classification. The classification may not be correct with classes that show high variance in a given direction (Hubert-Moy, 2001). Figure 2.5 shows the classification in a two-dimensional spectral space.

Maximum Likelihood Classification

Maximum likelihood classification is the most commonly used supervised classification method with satellite imagery (Richards and Jia, 1999). This classification uses the training data for estimating means and covariance of the classes and then assigns all the pixels to the statistically most likely class (Swain, 1978). Figure 2.6 shows a two-dimensional classification. A maximum likelihood algorithm assumes that the datahave normal distributions (Duda *et al.*, 2001), which provides an analytical solution to the

20

decision boundaries. In fact, image training data often show normal or nearly normal distribution (Schowengerdt, 1997).

The classification is performed based on the following decision rule:

$$\text{Decide x is in class } \omega_i \text{ if } p(\omega_i \mid x) > p(\omega_j \mid x) \text{ for all } j \neq i \tag{2.6}$$

where $p(\omega_i \mid x)$, $i = 1, \ldots M$, is the probability that the correct class is for a pixel x. However, $p(\omega_i \mid x)$ is generally unknown but can be obtained based on Bayes' theorem using $p(x \mid \omega_i)$ since the likelihood can be easily estimated from sufficient training data.

$$p(\omega_i \mid x) = \frac{p(x \mid \omega_i) p(\omega_i)}{p(x)} \tag{2.7}$$

where $p(x \mid \omega_i)$ is the likelihood finding a pixel from class ω_i, $p(\omega_i)$ is a priori probability that class ω_i occurs in the image and $p(x)$ is the probability of finding a pixel from any class, which can be considered as a scale factor. Then decision rule can be expressed as:

$$\text{Decide x is in class } \omega_i \text{ if } p(x \mid \omega_i) \, p(\omega_i) > p(x \mid \omega_j) \, p(\omega_j) \text{ for all } j \neq i \tag{2.8}$$

The mathematical expression for the likelihood for each class for N bands is:

$$p(x \mid \omega_i) = \frac{\exp\{-\frac{1}{2}(x - \mu_i)^t \Sigma_i^{-1}(x - \mu_i)\}}{(2\pi)^{N/2} \mid \Sigma_i \mid^{1/2}} \tag{2.9}$$

21

where μ_i is the mean vector of class i and Σ_i is the covariance matrix of class i. this equation can be reduced by taking natural logarithm and removing $p(\omega_i)$, which is the same for all class i. Then the final decision rule is as follows:

$$\text{Decide x is in class } \omega_i \text{ if } g_i(x) > g_j(x) \text{ for all } j \neq i \qquad (2.10)$$

$$g_i(x) = -\ln|\Sigma_i| - (x-\mu_i)^t \Sigma_i^{-1} (x-\mu_i) \qquad (2.11)$$

This classification method is theoretically the most powerful classification method if provided accurate training data. If the distributions of classes are normally distributed, this method is the optimal classification (Hubert-Moy *et al.*, 2001). Otherwise, this classification might not be valid. In addition, the classification can estimate the class of pixels in overlapping areas since the boundaries are quadratics (Hubert-Moy *et al.*, 2001). However, maximum likelihood classification is slower than other classifications discussed earlier (Richeards and Jia, 1999).

Artificial Neural Networks

Neural network application to remote sensing image classification is relatively new compared to those previously discussed methods (Paola and Schowengerdt, 1995b). Artificial neural networks are developed to emulate the information process of

22

human brain and nervous systems (Russell and Norvig, 1995; Ripley, 1996; Mitchell, 1997). Thus a number of neurons are highly interconnected in the network structure.

For image classification, most artificial neural networks derive the structure from the three-layer feed forward network: input, output, and hidden layers. Typically, each input node corresponds to each band of an image and each output node to each class. One hidden layer is generally recommended for image classification (Paola and Schowengerdt, 1995a). Every single node of input and output layers is connected to all nodes of the hidden layer. Each interconnection between nodes has an associated weight that contains the distributed information of the network. When training, each processing node sums the values of its inputs and passes the sums through an arbitrary activation function to produce the node's output value. The weight controls the threshold level of the activation function (Widrow and Lehr, 1990; Yoshida and Omatu,1994; Paola and Schowengerdt, 1995a, 1995b; Ripley, 1996).

Backpropagation algorithm is the first successful and one of the most commonly used for training a network (Rumelhart *et al.*, 1986). The backpropagation algorithm learns the weights iteratively to minimize an error, E, between the original and the desired output (Paola and Schowengerdt, 1995a, 1995b; Ripley, 1996).

$$E = \tfrac{1}{2}\sum_k (d_k - o_k)^2 \qquad\qquad (2.12)$$

$$o_k = f(NET_k) = f(\sum_j (w_{kj}o_j)) \qquad\qquad (2.13)$$

23

where d_k and o_k are the desired and original outputs associated with output node k, NET_k is the sum within output node k, w_{kj} is weight between the hidden (j) and output layer (k) and o_j is the output of the hidden layer nodes. In order to adjust weights and minimize the error, gradient descent is commonly used. Thus the updated weight between the hidden (j) and output layer (k) is:

$$w_{kj}' = w_{kj} - \eta \frac{\partial E}{\partial w_{kj}} \qquad (2.14)$$

$$\Delta w_{kj} = -\eta \frac{\partial E}{\partial w_{kj}} \qquad (2.15)$$

where w_{kj}' is updated weight and η is a positive constant that controls the amount of adjustment. If we use the chain rule, the gradient descent for output node and hidden node becomes

$$\Delta w_{kj} = \eta (d_k - o_k)(1 - o_k) o_k o_j \qquad (2.16)$$

$$\Delta w_{ji} = \eta (d_j - o_j)(1 - o_j) o_j o_i \qquad (2.17)$$

Once training is complete, entire image is feed into the network to classify each pixel.

An advantage of the artificial neural networks is that the network can learn the output from a given input without prior knowledge. Moreover, artificial neural networks

24

are nonparametric therefore they are robust although the data are not normally distributed.

2.3.2 Unsupervised Classification

Unsupervised classification reverses the supervised classification process (Niblack, 1986; Jensen, 1996; Richards and Jia, 1999; Mather, 1999) Unsupervised classification requires no prior knowledge of the region and therefore, the classification may not be biased by the prior knowledge. Unsupervised classification may recognize unique classes, which may be lumped into other classes with supervised classification creating error. However labeling spectral classes is not always easy. The analyst should understand the spectral characteristics of the area of interest. Figure 2.7 describes the process of unsupervised classification. We discuss two commonly used clustering algorithms with remote sensing image: K-means clustering and ISODATA.

K-means Clustering

K-means algorithm is simple and widely used clustering methods (Schowengerdt, 1997; Mather, 1999). This method requires the number of clusters and arbitrarily locates the cluster centers. Then each pixel is assigned to the cluster of which means is the closest. The means for each cluster are recalculated and the process continues until no

25
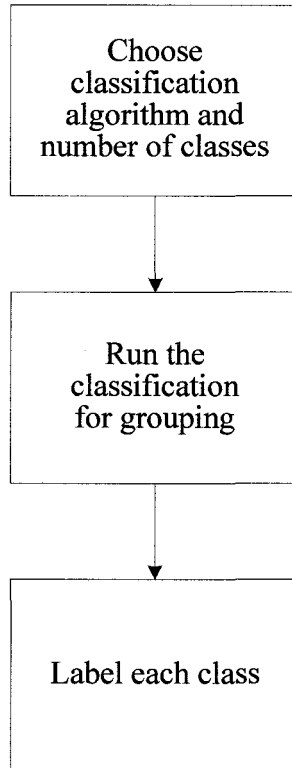
Figure 2.7 General steps of unsupervised classification

significant change in the clusters. The result of this classification is not sensitive to the initial specification of mean vectors in each cluster but the number of iterations can change. Figure 2.8 shows the application of this algorithm to a two-dimensional spectral space. Three clusters are chosen and the initial means are seeded equidistantly along the space diagonal. Diamond-shaped points are cluster means, ellipses are spectral objects and dotted lines are boundaries among each cluster.

ISODATA

ISODATA algorithm modifies K-Means clustering (Jensen, 1996; Schowengerdt, 1997; Mather, 1999) such that the algorithm includes merging clusters if the separation is below a certain threshold, and splitting a cluster into two if it is too large. ISODATA requires additional input parameters to set threshold for merging and splitting: maximum number of clusters to be merged, minimum number of pixels for each cluster, minimum distance between clusters, and maximum standard deviation. Setting a maximum number of clusters to be merged avoids overmerging. Setting a minimum number of pixels in a cluster avoids many small clusters. A threshold on minimum distance between clusters allows two close clusters to be merged. A threshold on standard deviation within each band for each class splits a cluster with too much variability.

27

initial cluster location

mean recalculation

final clusters after iteration

new clusters
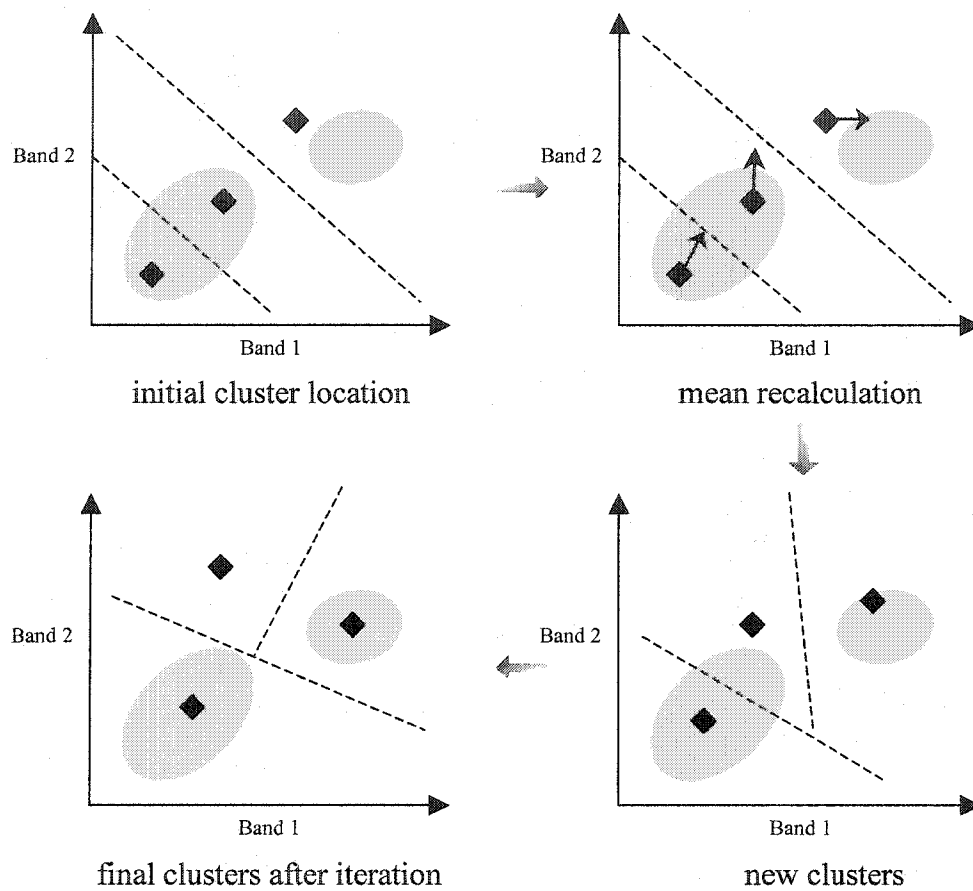
Figure 2.8 K-means algorithm in two-dimensional spectral space

28

2.3.3 Fuzzy Classification

The classifiers described in previous sections are called hard classification given that decision boundaries are well defined. However, some pixels can have mixed memberships in several classes such that the decision boundaries are fuzzy. These mixed pixels often occur at the boundary between objects. Soft classification, also known as fuzzy classification, is developed to solve the mixed pixel problem (Wang, 1990a, 1990b; Jensen, 1996; Schowengerdt, 1997). The classification employs fuzzy membership grades for each pixel as a real number from 0 to 1. If the value is close to 1, then the pixel is more likely belong to the class. The following expression is a membership grade in matrix U.

$$U = \begin{bmatrix} u_{11} & \cdots & u_{1N} \\ \vdots & & \vdots \\ u_{M1} & \cdots & u_{MN} \end{bmatrix} \tag{2.18}$$

$$0 \leq u_{mn} \leq 1$$

$$\sum_{n=1}^{N} u_{mn} > 0$$

$$\sum_{m=1}^{M} u_{mn} = 1$$

where u_{mn}, m = 1,...M and n = 1, ...N, is the membership value of pixel n in class m. M is total number of classes and N is total number of pixels.

29

Membership for each class can be defined based on the maximum likelihood classification or K-means clustering. In fuzzy C-means clustering algorithm, which is similar to K-means clustering, the membership values are updated by the normalized distance to the cluster means. In fuzzy supervised classification based on maximum likelihood, the membership values are updated by replacing mean and covariance with fuzzy mean and fuzzy covariance (Jensen, 1996; Schowengerdt, 1997).

2.4 Incorporating Ancillary Data

Image classification using only spectral signatures may not provide accurate results. Even high spatial resolution may not improve classification accuracy because high resolution increases spectral variation in each class (Marceau et al., 1994; Zhang, 2001). Ancillary data in addition to spectral data have been incorporated to improve the accuracy of classification (Hutchinson, 1982; Fahsi et al., 2000; Zhang, 2001). Ancillary data are any type of nonspectral information such as elevation, topology and texture (Jensen, 1996; Mather, 1999).

2.4.1 Texture

Texture is the variation of tone on an image (Jensen, 1996; Sabins, 1997). Common texture algorithms include local statistical measures and gray-level co-occurrence matrix (GLCM). Local statistics measures means, variance, standard

30

deviation and entropy. Gong *et al.* (1992) reported that the standard deviation was the best method (Jensen, 1996). GLCM represents the distance and angular spatial relationships among pixels over an image (Haralick *et al.*, 1973). GLCM measures the probability of occurrence of two gray-scale values separated by a given distance in a given direction. Several researchers found that GLCM improved classification (Peddle and Franklin, 1991; Gong *et al.*, 1992; Jensen, 1996). In addition, Zhang (2001) proposed the use of conditional variance that outperformed common texture algorithms. However, no single algorithm has been widely used for image classification.

2.4.2 Contextual Information

The context of a pixel is spatial relationship of a pixel with other pixels in the image (Jensen, 1996). Contextual classification incorporates neighboring pixel information because satellite imagery often exhibits high spatial correlation (Stuckens *et al.*, 2000; Hubert-Moy *et al.*, 2001). If the image has high spatial correlation, contextual classification will improve the accuracy. Contextual information can also correct the information class for isolated pixels.

There are two general approaches to calculate context: a moving window and image segmentation. By imposing windows, contextual information is extracted from a pixel's neighborhood and the value is assigned to the pixel. Segmentation divides the image into regions, called segments, based on spectral similarity. All pixels of a segment will belong to the same information class. Contextual data can be incorporated, either

31

during the classification (Kittler and Föglein, 1984) or as a post-classification step. Several researchers reported that contextual data assists in improving classification accuracy (Flygare, 1997; Stuckens *et al.*, 2000; Hubert-Moy *et al.*, 2001).

2.4.3 Digital Elevation Model

Elevation data such as digital elevation model (DEM) may be useful in image classification if some classes may have similar spectral values but have different elevation. Moreover, DEM can correct the effects of terrain and slope variability. Some researchers proved that DEM successfully reduced the topographic effect and, as a result, improved the classification accuracy (Justice *et al.*, 1981; Fashi *et al.*, 2000). However, DEM accuracy is critical because inaccurate DEM may introduce additional errors in classification.

2.4.4 Knowledge-based system

Expert systems, especially knowledge-based systems, have been adopted to incorporate both spectral and ancillary data (Jensen, 1996; Richards and Jia, 1999). A knowledge-based system simulates the human reasoning process for problem solving (Jackson, 1990). The most commonly used system is a rule-based system that represents knowledge as a set of rules. The following form is the well known rule system.

32

IF premise THEN consequence

where premise is logical expression that is true or false and if it is true the consequence part is justified.

When handling multisource data, either a single knowledge base or a set of knowledge base can be adopted. A single knowledge base contains all the rules in the premise and determines a particular class. In contrast, a set of knowledge base contains one rule for each premise that determines its subclass and combines the individual results into a module (Richards and Jia, 1999).

Many researchers have adopted rule-based systems. Mason *et al.* (1988) developed a rule-based system in conjunction with topographic map information. Their result showed improvement of classification. Bolstad and Lillesand (1992), Kartikeyan *et al.* (1995) and Stefanov *et al.* (2001) could achieve accuracy improvement using rule-based system with ancillary data. They emphasized the strength of the system in flexibility with multisources.

2.5 Image Transformations

Image transformation generates a new pixel value from arithmetic operation of pixel values of multispectral bands (Mather, 1999; Rees, 2001). Vegetation indices are the simplest image transformation that is used for vegetation feature extraction (Schowengerdt, 1997). Normalized difference vegetation index (NDVI) is one of most

widely used among more than 20 vegetation indices, which is based on two spectral

bands. Tasseled cap transformation is based on six spectral bands of Lansat TM and

ETM$^+$. Another commonly used technique is principle component analysis (PCA), which

removes spectral redundancy due to highly correlated multispectral bands (Ready and

Wintz, 1973; Schowengerdt, 1997)

2.5.1 Normalized Difference Vegetation Index

Simple ratio is the first vegetation index based on the ratios of the near IR band to

the red band for each pixel (Birth and McVey, 1968; Jensen, 2000).

$$SR = \frac{B_{IR}}{B_R} \qquad (2.19)$$

where B_{IR} and B_R are DN values of near IR and red band, respectively. This index

reflects the fact that chlorophyll of vegetation absorbs red wavelength such that lower

values of red band indicate higher chlorophyll content, and that leaves reflect near IR

wavelengths such that higher values of near IR band indicate more vigorous vegetation

growth (Sabins, 1997).

NDVI is a more complex version of vegetation index (Rouse *et al.*, 1974; Jensen,

2000)

$$NDVI = \frac{B_{IR} - B_R}{B_{IR} + B_R} \qquad\qquad (2.20)$$

The range of NDVI is between -1 and 1 and higher values indicate higher vegetation whereas lower values indicate less vegetation. NDVI is favorable over simple ratio since the value is not influenced by the absolute pixel values due to atmospheric conditions (Mather, 1999).

2.5.2 Tasseled Cap Transformation

Kauth and Thomas (1976) developed Tassled Cap transformation, which is a more sophisticated approach derived from linear transformation in four dimensional space (Schowengerdt, 1997; Lillesand *et al.*, 2004). The transformation rotates the Landsat MSS band data and generates new axes: brightness, greenness, yellowness, and nonsuch. Brightness is an axis associated with variations in soil background reflectance. Greenness is orthogonal to the first axis, which is associated with variations in green vegetation. Yellowness is associated with the yellowing of senescent variation. Nonesuch is associated with atmospheric conditions. The transformation is as follows:

$$\begin{bmatrix} brightness \\ greenness \\ yellowness \\ nonesuch \end{bmatrix} = \begin{bmatrix} 0.433 & 0.632 & 0.586 & 0.264 \\ -0.290 & -0.562 & 0.600 & 0.491 \\ -0.829 & 0.522 & -0.039 & 0.194 \\ 0.223 & 0.012 & -0.543 & 0.810 \end{bmatrix} \begin{bmatrix} DN_{B4} \\ DN_{B5} \\ DN_{B6} \\ DN_{B7} \end{bmatrix} \qquad (2.21)$$

35

Crist (1985) extended Tassled Cap transformation for Landsat TM imagery (Schowengerdt, 1997; Mather, 1999; Jensen, 2000). In this system, three axes are considered important: brightness, greenness, and wetness. Brightness is a weighted average of the TM bands, greenness is a contrast between visible and near IR, and wetness is a contrast between bands 5 and 7, and bands 3 and 4 (Mather, 1999).

$$
\begin{bmatrix} \text{bright} \\ \text{green} \\ \text{wet} \end{bmatrix} = \begin{bmatrix} 0.3037 & 0.2793 & 0.4343 & 0.5585 & 0.5082 & 0.1863 \\ -0.2848 & -0.2435 & -0.5436 & 0.7243 & 0.0840 & -0.1800 \\ 0.1509 & 0.1793 & 0.3299 & 0.3406 & -0.7112 & -0.4572 \end{bmatrix} \begin{bmatrix} DN_{B1} \\ DN_{B2} \\ DN_{B3} \\ DN_{B4} \\ DN_{B5} \\ DN_{B7} \end{bmatrix} \quad (2.22)
$$

Both transformations are derived from empirical data of small part of the North America. Therefore, the application of the transformations to other area may not be successful (Mather, 1999).

2.5.3 Principle Component Analysis

Principle component (PC) analysis is a linear transformation of feature space to reduce the correlation of spectral bands since multispectral bands are often highly correlated, which results in redundancy (Ready and Wintz, 1973; Schowengerdt, 1997). New pixel values in PC are linear combination of the pixel values in the original axis and

36

the transformation coefficients, i.e. eigenvectors. Each component is orthogonal to the previous component.

$$DN_{PC_p} = \sum_{k=1}^{n} e_{kp} DN_k \qquad (2.23)$$

where DN_{PC_p} is new pixel DN value of p component in PC image, e_{kp} is eigenvector, and DN_k is pixel DN values in band k of original data.

Eigenvectors and eigenvalues of the covariance matrix are computed from the following equation.

$$\Sigma_{PC} = E \ \Sigma \ E^T = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \\ \lambda_7 \end{bmatrix} I \qquad (2.24)$$

where E is eigenvector matrix for each bands (n×n), Σ_{PC} and Σ are covariance matrix of PC image and the original data (n×n), respectively, λ_i is eigenvalue, and I is identity matrix. The eigenvalues and eigenvectors represent the length and the direction of principle components (Mather, 1999). Then the percentage of total variance of each principle component is calculated by

37

$$\%_{PC} = \frac{\lambda_p}{\sum\limits_{p=1}^{n} \lambda_p} \times 100 \qquad\qquad (2.25)$$

where λ_p is eigenvalue for each principle component. The first component has the largest variance and the higher component has less variance. In general, the first three components contain most percentage of the total image variance, which results in dimensionality reduction (Rees, 2001). This is useful since the PCA can reduce computational cost. Therefore, PCA transformation can be effectively employed as a preprocessing method prior to image classification (Lillesand *et al.*, 2004).

2.6 Accuracy Assessment

Accuracy assessment is not always easy for satellite image classification. Overall accuracy alone is not meaningful because it does not provide accuracy of individual classes (Foody, 2002). In order to provide the accuracy of individual classes, confusion matrix is most widely used (Jensen, 1996; Richards and Jia, 1999; Foody, 2002). A confusion matrix, also called error matrix, is a square table of the resulting map class labels as rows against the original class in the ground truth as columns. This matrix assumes that every pixel belongs to one of the classes that are exhaustive and mutually exclusive (Foody, 2002).

There are two different error measures: omission and commission errors. Omission error corresponds to those pixels actually belonging to the class of interest but

38

failed to be recognized as a member of that class whereas commission error corresponds to pixels from other classes that has incorrectly labeled as the class of interest (Congalton, 1991; Janssen and van der Wel, 1994).

$$\text{Omission error} = 1 - \frac{n_{ij}}{\sum_i n_{ij}} \qquad (2.26)$$

$$\text{Commission error} = 1 - \frac{n_{ij}}{\sum_j n_{ij}} \qquad (2.27)$$

$$\text{overall accuracy} = \frac{\sum_k n_{kk}}{N} \qquad (2.28)$$

where n_{ij} is the number of pixels row i and column j, and N is the total number of test pixels in the confusion matrix. However, the confusion matrix has a problem in that some pixels may have been correctly allocated merely by chance. To resolve this problem, kappa coefficient has been used (Congalton 1991, Foody, 2002).

$$\kappa = \frac{N \sum_k n_{kk} - \sum_k (\sum_j n_{kj} \sum_i n_{ik})}{N^2 - \sum_k (\sum_j n_{kj} \sum_i n_{ik})} \qquad (2.29)$$

The accuracy of ground truth data should be ensured for proper accuracy assessment. Errors in the ground truth data can significantly influence the classification accuracy (Zhu *et al.*, 2000, Foody, 2002).

39

Chapter III

Bayesian Networks

Bayesian networks, also called probabilistic networks, belief networks, or causal networks, are powerful probabilistic approaches to knowledge representation and handling problems under uncertainty. Bayesian networks have been successfully applied to many areas such as pattern recognition, language understanding, computer vision, medical informatics, and decision-making (Heckerman, 1990; Goldman, 1990; Charniak, 1991; Sucar and Gillies, 1994; Lucas and Abu-Hanna, 1999; Bang and Gillies, 2002). Bayesian networks have been also adopted in environmental areas such as risk assessment, water quality management and wastewater treatment system (Varis, 1995; Chong and Wally, 1996; Sangüesa and Burrell, 2000; Borsuk and Stow, 2000; Sahely and Bagley, 2001; Borsuk et al., 2004). Bayesian networks have been compared with other

40

AI techniques such as rule-based systems and neural networks, which have been widely used in the environmental engineering area (Ozgur and Stenstrom, 1994; Zhang and Stanley, 1997; Brion and Lingireddy, 1999; Baeza, *et al.*, 2002; El-Din and Smith, 2002; Comas *et al.*, 2003). They were found to outperform rule-based systems for diagnosis in wastewater treatment systems (Chong and Wally, 1996; Sangüesa and Burrell, 2000) and an alternative to neural networks in wastewater treatment modeling (Hiirsalmi, 2000).

In this chapter, we review Bayesian networks, especially naïve Bayesian classifiers, and Bayesian networks based on maximum weight spanning tree (MWST) algorithm. In section 3.1, we present the fundamental terminology of Bayesian networks. In section 3.2, we explain the discretization methods for Bayesian networks. In section 3.3, Bayesian network structures are compared. In section 3.4, propagation process in a network is explained. In section 3.5, Bayesian networks are compared with other AI methods, i.e. decision trees and neural networks.

3.1 Basic Concepts of Bayesian Networks

A Bayesian network is a probabilistic approach that graphically represents relationships between variables in a given domain (Pearl, 1988; Neapolitan, 1990). The structure of the network provides cause and effect relationships among variables and an assessment of uncertainty. The network is graphically represented as a directed acyclic graph (DAG) consisting of nodes and arcs where the nodes stand for random variables and the arcs show direct link between variables. In a Bayesian network, the relationship

41

between variables is quantified with conditional probabilities, and any node is conditionally independent of it non-descendent variables given its parents. The following concepts and terminology are used in Bayesian network (Pearl, 1988; Neapolitan, 1990).

Definition 3.1 Conditional probability

Let X and Y be events and $P(Y) \neq 0$. The conditional probability, $P(X|Y)$ is defined as

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}.$$

Definition 3.2 Conditionally independency

Two events X and Y are conditionally independent given Z if $P(Z) \neq 0$ and it satisfies one of the followings.

- $P(X|Y, Z) = P(X|Z)$ and $P(X|Z) \neq 0$, $P(Y|Z) \neq 0$

- $P(X|Z) = 0$, $P(Y|Z) = 0$

Definition 3.3 Bayes' Theorem

Let X and Y be events such that $P(X) \neq 0$ and $P(Y) \neq 0$. Then

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}.$$

42

Furthermore, If there are n mutually exclusive and exhaustive events $X_1, X_2, ..., X_n$ such that $P(X_i) \neq 0$ for all i where $1 \leq i \leq n$ then,

$$P(X_i \mid Y) = \frac{P(Y \mid X_i)P(X_i)}{\sum_{i=1}^{n} P(Y \mid X_i)P(X_i)}.$$

The Bayes' theorem can be rearranged by replacing $\alpha = \dfrac{1}{P(Y)}$

$$P(X \mid Y) = \alpha P(X)P(Y \mid X).$$

where P(X) is a prior probability and P(Y|X) is likelihood information. Prior probability represents our belief about the given domain and can be taken from statistics or in a subjective way. Likelihood represents uncertainty and should be obtained from data.

Definition 3.4 Joint Probability Distribution

Let $V = \{X_1, X_2, ..., X_n\}$ be a set of n random variables and the domain of x_i be the same as the domain of X_i for $1 \leq i \leq n$. A function assigning a real number $P(X_1=x_1, X_2=x_2, ..., X_n=x_n)$ to every combination of values of the x_i's is a joint probability distribution of the variables in V if it holds that

- For every combination of values of the x_i's

43

$$0 \leq P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) \leq 1$$

$$\sum_{x_1, x_2, \ldots x_n} P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) = 1.$$

The joint probability distribution in a Bayesian network is equal to the product of the conditional probability distributions of all nodes given values of their parents whenever these conditional probabilities exist.

$$P(\{X\}) = \prod P(x_i \mid PA_{x_i})$$

Definition 3.5 Markov Condition

Let P be a joint probability distribution of variables in some set V and a DAG G=(V, E) where V is a finite, nonempty set whose elements are nodes and E is a set of ordered pairs of distinct elements of V, whose elements are arcs. (G, P) is in the Markov condition if for each variable $X \in V$, {X} is conditionally independent of the set of all its non-descendents given the set of all its parents. Let PA_X be the set of parents and ND_X non-descendents of X then,

$$I_P(\{X\}, ND_X \mid PA_X).$$

If X is a class node, its parent PA_X is an empty set so that the Markov condition means {X} is independent of ND_X, which is denoted $I_P(\{X\}, ND_X)$.

44

Definition 3.6 D-separation

Let X, Y and Z be disjoint subsets of nodes in a DAG, then Z is said to d-separate X from Y, denoted as

$$<X \mid Z \mid Y>_D$$

if there is no path between a node in X and a node in Y along which the following condition satisfies

- Every node with converging arcs is in Z or has a descendant in Z.

- Every other node is outside Z.

D-separation is used for testing independence in a Bayesian network.

Definition 3.7 I-map

A DAG, G, is said to be an I-map if every d-separation condition displayed in G corresponds to a valid conditional independence relationship, I, i.e.

$$<X \mid Z \mid Y>_G, \Rightarrow I(X, Z, Y)_M$$

45

This means that a DAG is an I-map if all the independencies in the DAG are present in a probability distribution. A DAG is a minimal I-map if it is an I-map with the minimum number of links, that is, if none of its arcs can be deleted without destroying its independence.

Definition 3.8 Bayesian Network

Let P be a joint probability distribution of the random variables in some set V, and G = (V, E), then (G, P) is a Bayesian network if (G, P) satisfies the Markov condition or if and only if G is a minimal I-map of P. In other words, a Bayesian network is a graph with the minimum number of links that faithfully represents all the probabilistic independencies for variables.

3.2 Discretization of Data

Bayesian networks normally require discrete data if the variables are not normally distributed. Therefore, continuous data should be converted into discrete data. In general, discrete data offer many advantages compared with continuous data. First, discrete data provide faster and more accurate learning (Dougherty *et al.*, 1995; Liu *et al.*, 2002; Yang and Webb, 2003). Secondly, discretization makes data easier to understand and more representative by simplifying the data (Liu *et al.*, 2002). Thirdly, discretization requires no assumption of the probability distribution and therefore can avoid any inaccuracy related to the distribution (Yang and Webb, 2003). For example, many classification

46

algorithms, such as maximum likelihood classifiers, assume that the data are normally distributed, and if not, this assumption might bias the analysis. Finally, discretization particularly improves the Bayesian network application to those domains violating the independence assumption (Pazzani, 1995).

Many researchers have developed and categorized discretization methods in various ways: supervised vs. unsupervised (Dougherty *et al.*, 1995), dynamic vs. static (Liu *et al.*, 2002), global vs. local (Chmielewski and Grzymala-Busse, 1994; Dougherty *et al.*, 1995), split vs. merge (Kerber, 1992; Yang and Webb, 2003), and direct vs. incremental (Cerquides and Mantaras, 1997; Liu *et al.*, 2002). Supervised methods use class information of the training data for partitions whereas unsupervised methods do not use class information. Dynamic methods conduct discretization during classification whereas static methods discretize continuous values before classification. Global methods discretize the entire data space whereas local methods can be applied to localized space. Therefore, each attribute in the data set can be differently partitioned in local approach. Split methods start with the entire range with zero interval and partition it into subintervals whereas merge methods start with continuous values and merges adjacent values. Both methods continue the process until certain predefined threshold is met. Direct methods partition the range of intervals simultaneously, which requires the information of the number of intervals, whereas incremental methods starts with a simple discretization and improves, which requires information of stop point of discretization. Table 3.1 summarizes each discretization method.

Table 3.1 Examples of discretization methods

	Supervised/ Unsupervised	Dynamic/ Static	Global/ Local	Split/ Merge	Direct/ Incremental
Equal width interval	U	S	G	S	D
Equal frequency interval	U	S	G	S	D
1R (Holte '93)	S	S	G	S	D
ID3 (Quinlan '86)	S	D	L	S	I
C4.5 (Quinlan '93)		D		S	
Hierarchical maximum entropy (Chiu et al. '90)	S		L		
D2 (Catlett, '91)	S	S	G	S	I
MDLP (Fayyad et al. '93)	S	S	L	S	I
Mantaras distance (Cerquides et al. '97)	S	S	L	S	I
Zeta (Ho et al. '97)	S	S	G	S	D
Adaptive quantizer (Chan et al., '91)	S, U	S	G	S	D
MCC (Van de Merckt, '93)	S, U		G		
Predictive value maximization (Weiss et al., '90)	S		G		
Vector quantisation (Kohonen, '89)	S		L		
ChiMerge (Kerber, '92)	S	S	G	M	I
Chi2 (Liu et al. '95)	S	S	G	M	I
StatDisc (Richeldi et al. '95)				M	
k-means clustering (Maass, '94)	U		L		
Fuzzy learning (Kononenko, '92)		S	G		
Interative improvement (Pazzani, '95)	S	S	G	S, M	

(adopted from Dougherty et al., 1995; Liu et al., 2002; Yang and Webb, 2003)

48

Dougherty *et al.* (1995) and Liu *et al.* (2002) reviewed numerous discretization and concluded that entropy-based discretization such as minimum description length principle (MDLP, Fayyad and Irani, 1996) was the best method. Dougherty *et al.* also noted that simple methods such as equal width interval method could be successful under a normal distribution. Yang and Webb (2003) proposed proportional k-interval discretization and equal size discretization for naïve Bayesian classifiers. However, there exists no consensus on discretization methods.

3.2.1 Binning Methods

The simplest discretization method is collecting bins (Liu *et al.*, 2002). The examples are equal width interval and equal frequency interval. Equal width interval method breaks the range of the observed data values with k equally sized intervals (Dougherty *et al.*, 1995; Liu *et al.*, 2002; Yang and Webb, 2003). The number of intervals k can be selected subjectively. Each interval width is defined as follows:

$$\text{interval width} = \frac{x_{max} - x_{min}}{k} \tag{3.1}$$

where x_{min} and x_{max} are the minimum and the maximum value of a variable x respectively. This method can be applied to each variable independently of the others without considering the class node information.

49

Equal frequency interval method divides the range of the data values with k unequally sized intervals. Each interval contains the same number of data values:

$$\text{data numbers in an interval} = \frac{n}{k} \qquad (3.2)$$

where n and k are the total number of data values and intervals respectively (Liu *et al.*, 2002). In this case, those training data with identical values should belong to the same interval and therefore each interval may not always have the same number of data. These two methods have been widely used in many machine learning algorithms since they are simple and reasonably good in performance (Hsu *et al.*, 2000). However, these methods are sensitive to the interval, k (Dougherty *et al.*, 1995).

1R algorithm is a different binning method from above methods since it requires class information (Holte, 1993; Dougherty *et al.*, 1995; Liu *et al.*, 2002). The method breaks the range of the data values and adjusts the discretization boundaries based on the class associated with the data values. If each interval contains minimum number of values that is prespecified, the discretization will be terminated.

3.2.2 Entropy-based Methods

Another discretization method is using entropy measure to find a discretization boundary. The following equation represents Shannon entropy, which is one of the mostly widely used (Shannon and Weaver, 1949; Thornton, 1992; Liu *et al.*, 2002):

50

$$H(X) = -\sum_{x} p(x) \log p(x) \qquad\qquad (3.3)$$

where p(x) is probability of value x of X.

Hierarchical discretization method (Chiu *et al.*, 1990) discretizes continuous variables maximizing the Shannon entropy. This method employs hill-climbing search to find discretization boundaries (Dougherty *et al.*, 1995).

ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993) binarizes a range of values using entropy measure for decision trees (Liu *et al.*, 2002). In these algorithms, a discretization boundary is selected where the entropy is the lowest. ID3 uses greedy search to find discretization boundaries.

D2 algorithm (Catlett, 1991) is similar to ID3 except it discretizes before classification (Liu *et al.*, 2002). This method is used for decision trees and increases the speed of induction. The discretization completes when the minimum number of samples in one partition, the maximum number of partitions, or the minimum information gain is obtained.

Fayyad and Irani (1993) selected a discretization boundary to minimize the class entropy. The method uses MDLP to complete discretization (Dougherty *et al.*, 1995; Liu *et al.*, 2002; Yang and Webb, 2003) In this method, the class entropy is calculated as follows (Fayyad and Irani, 1993; Dougherty *et al.*, 1995):

51

$$E(A,T;S) = \frac{|S_1|}{|S|} ENT(S_1) + \frac{|S_2|}{|S|} ENT(S_2) \qquad (3.4)$$

where $E(A, T; S)$ is the class entropy, S is set of instances, A is a continuous variable and T is a discretization boundary. The method is repeatedly applied and completes if the following condition is satisfied:

$$Gain(A,T;S) < \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N} \qquad (3.5)$$

$$Gain(A,T;S) = Ent(S) - E(A,T;S) \qquad (3.6)$$

$$\Delta(A,T;S) = \log_2(3^k - 2) - (k \cdot Ent(S) - k_1 \cdot Ent(S_1) - k_2 \cdot Ent(S_2)) \qquad (3.7)$$

where N is the total number of instances of S and k_i is the number of class labels in S_i.

3.2.3 Accuracy Measure

Zeta is the maximum accuracy achievable of a value assigned to different partition class (Ho and Scott, 1997). A discretization boundary is selected where the zeta value is the highest (Liu *et al.*, 2002).

Adaptive quantizer (Chan *et al.*, 1991) starts with simple binning methods and tests classification rules for accuracy improvement. The splitting continues for the interval with the lowest accuracy. This method can compromise the limit of binning methods but is computationally expensive (Dougherty *et al.*, 1995).

52

3.2.4 Contrast-based Method

Monothetic Contrast Criterion (MCC) approach (Van de Merckt, 1993) uses an unsupervised clustering algorithm to find the most contrasted discretization boundary (Dougherty et al., 1995). However, this approach cannot distinguish two close cut-points with similar contrast nor evaluate the correlation between variables and class information (Bang, 2002).

Mixed supervised/unsupervised MCC incorporates entropy measures into the previous MCC to overcome the limitation of the MCC (Dougherty et al., 1995). This approach can recognize the correlation between variables and their class information.

3.2.5 Merge Method

ChiMerge method (Kerber, 1992) considers the relationship between the feature and the class (Liu et al., 2002). The method tests χ^2 values of the adjacent intervals and merges them if they have the least χ^2 values, which is defined by:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \qquad (3.8)$$

$$E_{ij} = \frac{\sum_{j=1}^{k} A_{ij} \times \sum_{i=1}^{m} A_{ij}}{\sum_{j=1}^{k} \sum_{i=1}^{m} A_{ij}} \qquad (3.9)$$

53

where k is the number of classes, m is the number of interval, A_{ij} is the number of values in the interval, i and class, j, and E_{ij} is expected frequency of A_{ij}.

Chi2 (Liu and Setiono, 1995) modifies chiMerge using inconsistency measure to set a proper threshold. This method has an advantage in that it can remove noisy attributes (Liu *et al.*, 2002).

3.3 Structures of Bayesian Networks

Naïve Bayesian classifiers are the simplest Bayesian network structure. Modified structures of naïve Bayesian classifiers have been developed to overcome the limit of strong conditional independence assumption (Clark and Niblett, 1987; Kononenko, 1991; Langley *et al.*, 1992; Langley and Sage, 1994; Pazzani, 1995). Another approach to construct networks is adopting a tree structure using MWST algorithm.

3.3.1 Naïve Bayesian Classifiers

Naïve Bayesian classifiers, also called Bayesian classifiers or simple Bayesian classifiers, are the simplest Bayesian networks, which have only one class node. An example is shown in Figure 3.1. In this figure, C stands for the class node and the other nodes such as X, Y, and Z are all its child nodes.
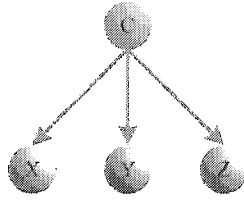
Figure 3.1 An example of naïve Bayesian classifiers
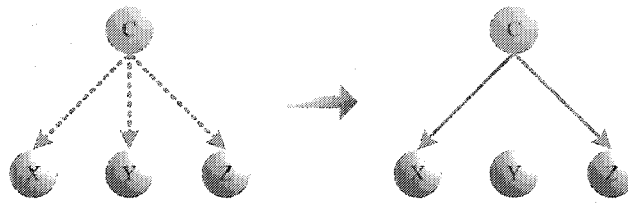
Figure 3.2 An example of selective naïve Bayesian classifiers
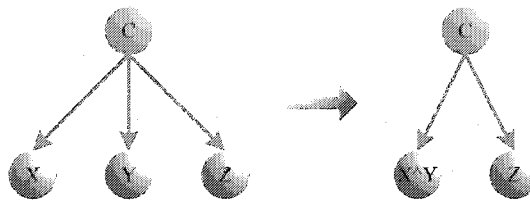
Figure 3.3 An example of joint naïve Bayesian classifiers
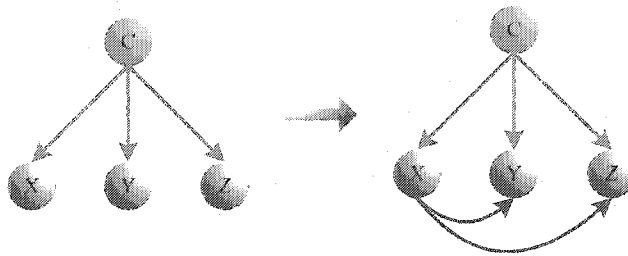
Figure 3.4 An example of tree-augmented naïve Bayesian classifiers

55

The basic assumption is that computation in the naïve Bayesian classifiers is under a strong conditional independence assumption between variables. This independence assumption reduces the number of dependency checks needed to connect the network. On the other hand, it has been pointed out as the limitation of the naïve Bayesian classifiers since its performance is doubtful to handle highly dependent variables. In a real world situation, there are few domains that absolutely satisfy the conditional independence assumption. But many researchers found that naïve Bayesian classifiers perform as well as or better than other more complicated networks (Clark and Niblett, 1989; Langley *et al.*, 1992; Michie *et al.*, 1994; Friedman *et al.*, 1997; Webb and Pazzani, 1998).

One of the advantages of the naïve Bayesian classifiers is that they are easier to construct. Another advantage also stems from their computational simplicity so that the networks require less time and memory. Besides, their inference is much simpler given that there is only one propagation step between the class node and any child node. The propagation does not rely on the order in which the evidence arrives. Another useful feature of naïve Bayesian classifiers is that additional information can be easily added into the existing network. This means there is no need to reconstruct the network whenever incorporating new information.

3.3.2 Modified Naïve Bayesian Classifiers

In order to overcome any violation of conditional independence in naïve Bayesian classifiers, some researchers developed modified naive Bayesian classifiers: selective Bayesian classifiers, joint naïve Bayesian classifiers, and tree-augmented naïve (TAN) Bayesian classifiers. Selective Bayesian networks modify naïve Bayesian classifiers by eliminating a redundant variable that lowers accuracy (Langley and Sage, 1994). This method starts with an empty set and then adds a variable if the variable increases the accuracy of the outcome. This process continues until all variables have been tested. Figure 3.2 shows an example. Some researchers reported successful performance of selective Bayesian classifiers compared with naïve Bayesian classifiers or decision trees (Langley and Sage; 1994; Michie and Al Attar; 1991).

Joint naïve classifiers are extended version of naïve Bayesian classifiers by joining correlated variables (Pazzani, 1995). This approach is useful for a domain that has strong dependency between variables. Figure 3.3 shows an example of joint Naïve Bayesian classifiers. The networks have a potential problem that the joint variable from two correlated variables might require a complex search. Therefore, the joining algorithm can be limited to pairs of variables (Pazzani, 1995). Pazzani found that joint Bayesian classifiers outperformed naïve Bayesian classifiers.

TAN Bayesian classifiers represent the correlation between variables as opposed to eliminating or modifying the correlated variables (Friedman *et al.*, 1997). The network structure has additional arcs between child nodes as shown in Figure 3.4. For example of

57

three variables, X, Y, Z, if X is correlated to Y and Z then the network connects X to both

Y and Z to represent the dependence between them. Friedman *et al.* (1997) reported that

this approach was better than both naïve Bayesian classifiers and decision trees based on

C4.5. The approach is useful is for small domain problems and domains with many

classes (Friedman and Goldszmidt, 1996).

3.3.3 Tree Structured Bayesian Networks

Another Bayesian network structure is based on MWST algorithm. The network

structure can be constructed in a subjective way from expert knowledge or in an objective

way from data. In order to establish a network from data, the spanning tree algorithm is

widely used. Chow and Liu (1968) presented an approach to build an optimal dependence

tree from data. Their intention was to minimize the number of measuring mutual

information of variables to estimate an underlying n-dimensional probability distribution.

Mutual information provides measurement of dependency. It is zero for completely

independent variables but increases as the variables become dependent.

$$MI(X_i, X_j) = \sum P(x_i, x_j) \log(\frac{P(x_i, x_j)}{P(x_i)P(x_j)}) \qquad (3.10)$$

For constructing a network, it is needed to compute the joint probabilities of the

variables and their mutual information as dependency measure. The procedure continues

58

by adding arc with the largest mutual information, also called weight, into the network. When the construction is completed, the sum of weight in the network is maximized (Chow and Liu, 1968). The resulting network does not contain the directions of the arcs. Finally the network would be completed by adding causal directions that represent cause and effect relationships between variables. Selecting causal directions as well as choosing link for the same mutual information pairs and the class node would require the intervention of expert's subjective knowledge. However, we can apply this algorithm only to tree structures. Figures 3.5 and 3.6 present an example of the network and the process of construction, respectively. In Figure 3.6, the numbers stand for the order of magnitude of dependency between variables. As shown in the figure, the arc with dependency values of 3 is not selected since this would form a loop. Instead, the arc of the next dependency is selected, which is arc with the value of 4. As a result, we can obtain a directed acyclic graph representing cause and effect relationships between variables.

In complex problems, it would be difficult to elicit experts' knowledge since the expert might not completely understand all the dependencies of the problems. In this case, the causal direction can be objectively obtained. If we select one class node, the class node is presumably points all the arrows away from it. Similarly, for cases of two class nodes, we can set direction of all arrows from the class nodes. This would create a singly connected network as shown in Figure 3.7. On the other hand, if there is no information of the causal relationship, we can obtain the causal direction from the

59
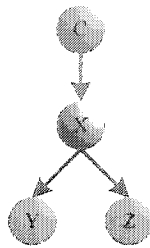
Figure 3.5 An example of tree structured Bayesian networks
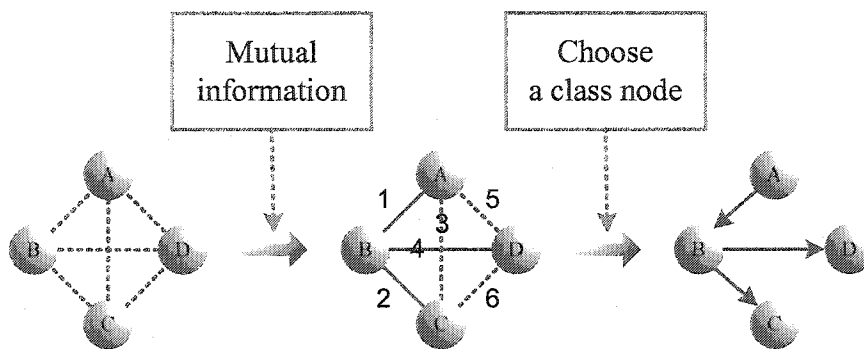
Figure 3.6 Construction of a network based on spanning tree algorithm
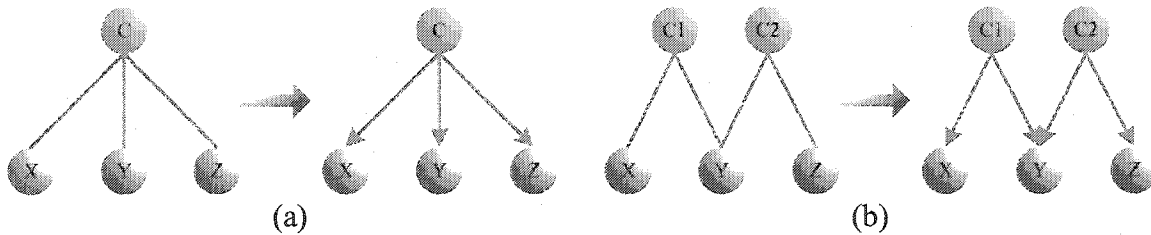
Figure 3.7 Adding causal directions to a spanning tree
(a) with one root node (b) with two root node

60

conditional independence. For example, if there are three nodes in the network, i.e. A, B, C, and node A and node C are independent but dependent given node B, i.e. P(A|C)=P(A) and P(C|A) = P(C), provided P(A|C,B)≠P(A|C) or P(A|B)=0, P(C|B)=0, the direction of arrows will be from A and from C to B. For another example, if node A and node C are conditionally independent given node B, i.e. P(C|A, B) = P(C|B), then the arrow directs from A to B and from B to C. These are illustrated in Figure 3.8 but are not always very effective.

MWSTs offer advantages compared with naïve Bayesian classifiers. The networks can explicitly provide information of the most important variables to determine target variable values. This allows us to minimize the problem without considering all possible variables. Moreover, the trees can be constructed from data, which does not require any other knowledge except mutual information between variables.

3.4 Probability Propagation

Probabilistic reasoning in Bayesian networks comprises instantiating the input variables and propagating their influences through the networks to update the probability of the variables given the evidence. To facilitate the message passing, Pearl proposed a method for inference in tree-structured networks in which every node has only one parent except the class node (Pearl, 1982). Afterwards the message passing for singly connected networks was developed in which a node can have more than one parent with a single path between any two nodes (Kim and Pearl, 1983). Spiegelhalter (1986) and Lauritzen *et*

61

(a)

(b)

Figure 3.8 Causal directed networks (a) P(A|C)=P(A), P(C|A)=P(C) (b) P(C|A, B)=(C|B)
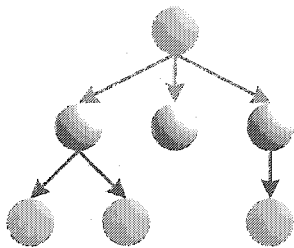
(a)
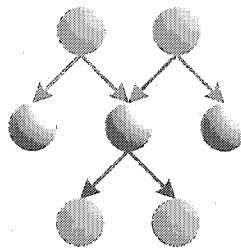
(b)
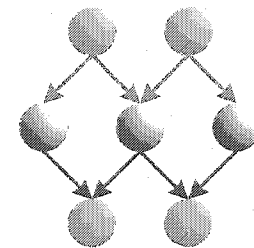
(c)

Figure 3.9 Three different network structures (a) tree (b) singly connected networks and (c) multiply connected networks
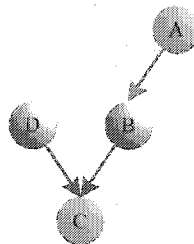
Figure 3.10 A tree network for message passing

62

al. (1986) presented different approaches for multiply connected networks in which there can be more than one path between nodes. Figure 3.9 shows examples of these three different networks.

In each node of a network, there are generally six different data stored: the link matrix, π evidence, λ evidence, π message, λ message and the posterior probability. The link matrix is a matrix form of the conditional probability of a node given its parent node. The π evidence is the prior evidence of a node derived from its ancestors where as the λ evidence is likelihood evidence accumulated from its child nodes and derived from measurement from data. A π message passes from a parent node to a child node while a λ message passes from a child node to a parent node, in the opposite direction to π message. When receiving and combining the evidences, a posterior probability distribution of a node is obtained. We will restrict our focus on the algorithm for singly connected networks since it provides a general idea. The algorithms are based on following five operating equations associated with the network structure in Figure 3.10.

Operating Equation 1: The λ messages

$$\lambda_B(a_i) = \sum_{j=1}^{m} P(b_j \mid a_i)\, \lambda(b_j)$$

Operating Equation 2: The π messages

    - If A is instantiated for a_i: $\pi_B(a_i) = 1$

    - If A is instantiated, but not for a_i: $\pi_B(a_i) = 0$

- If A is not instantiated:
$$\pi_B(a_i) = \frac{P'(a_i)}{\lambda_B(a_i)}$$

Operating Equation 3: The λ evidence

- If B is instantiated for b_j: $\qquad\qquad \lambda(b_j) = 1$

- If B is not instantiated, but not for b_j: $\qquad \lambda(b_j) = 0$

- If B is not instantiated: $\qquad\qquad \lambda(b_j) = \prod_i \lambda_{C_i}(b_j)$

Operating Equation 4: The π evidence

- For one parent: $\qquad \pi(b_j) = \sum_{i=1}^{m} P(b_j \mid a_i)\pi_B(a_i)$

- For two parents: $\qquad \pi(c_i) = \sum_{j=1}^{n} \sum_{k=1}^{m} P(c_i \mid b_i \ \& \ d_k)\pi_C(b_j)\pi_C(d_k)$

Operating Equation 5: Posterior Probability

$$P'(b_j) = \alpha \ \lambda(b_j) \ \pi(b_j)$$

where α is normalization factor.

Initially, the information of the network contains only the prior probabilities of the class nodes and the link matrices of conditional probability. This initialization process is performed as follows.

- Set all λ messages and λ values to 1.

64

- Set all π messages and π to 1.

- Set π values to the prior probabilities for class nodes.

- Post the π messages from the class nodes.

There are two ways of massage passing:upward propagation and downward propagation. In upward propagation, evidences are posted up the network whereas evidences are passed down the network in downward propagation. If B receives a λ message from a child node and B is not instantiated, upward propagation will be performed. If B is instantiated, the evidence from its child nodes does not affect. The upward propagation steps are as follows:

- Compute a new $\lambda(B)$ value based operating equation 3.

- Compute a new posterior probability $P'(B)$ based on operating equation 5.

- Post a λ message to all B's parent nodes

- Post a π message to B's other child nodes

If B receives a π message from a parent and B is not instantiated, the downward propagation will be performed as follows:

- Compute a new $\pi(B)$ value based on operating equation 4.

- Compute a new posterior probability of the node based on operating equation 5.

65

- Post a π message to each child of B

- Post a λ message to the other parents if there is λ evidence in C.

When new information arrives to the network, a variable is instantiated. Instantiation of a network requires recalculation of the posterior probabilities of each node in the entire network. Evidence changes of a node should be informed to its neighbors to update their local parameters. The following steps are instantiation in the networks.

- Set posterior probability of the instantiated state of a variable to 1 and that of all the other states of the variable to 0.

- Compute λ evidence of the variable based on operating equation 3.

- Post a λ message to each parent of the variable.

- Post a π message to each child of the variable.

In a singly connected network, the network will be in a steady state when there is no more variable needs updating. However, the message passing of a multiply connected network will not be terminated due to its topological characteristic of loop. Figure 3.11 shows the method of message passing in a singly connected network with one class node.

66
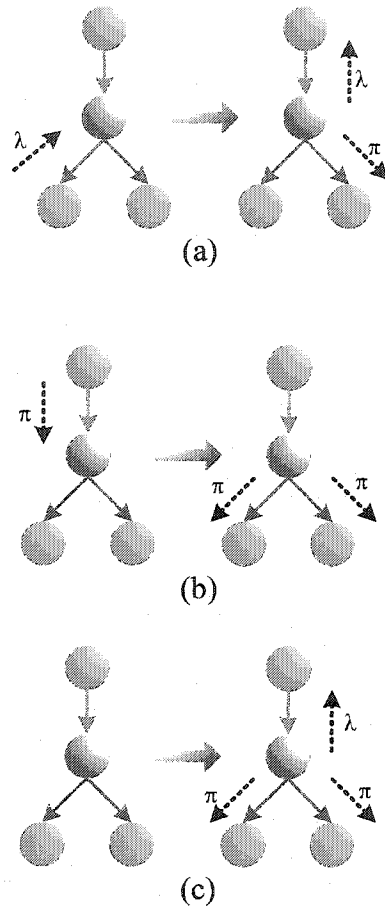
Figure 3.11 Schematic diagram of massage propagation in a singly connected network (a) upward propagation (b) downward propagation (c) instantiation

67

3.5 Comparison of Artificial Intelligence Approaches

In recent years, several AI methods have been introduced into many areas. We compare some important AI techniques such as decision trees with rules, artificial neural networks and Bayesian networks. Figure 3.12 presents simple structure of each system. The decision tree is an inference diagram corresponding to the rules. The example of neural networks is a three-layered network consisting of an input layer, a hidden layer and an output layer. Bayesian networks show cause and effect relationships among variables.

3.5.1 Decision Trees with Rules

Decision trees are simple and yet widely used since they are very easy to comprehend and implement (Russell, 1995; Mitchell, 1997; Lucas and Abu-Hanna, 1999). A decision is made by passing information from a class node down to some leaf nodes. These methods are suitable for problems with discrete values rather than continuous values. Decision trees may require disjunction of values of each node. However, decision trees can handle data containing error or missing values (Mitchell, 1997; Ripley, 1996). Decision trees have been popular in medical diagnosis, engineering and other classification problems (Ripley 1996) and have been explored in pattern recognition and image processing (Pal and Mather, 2003).

68
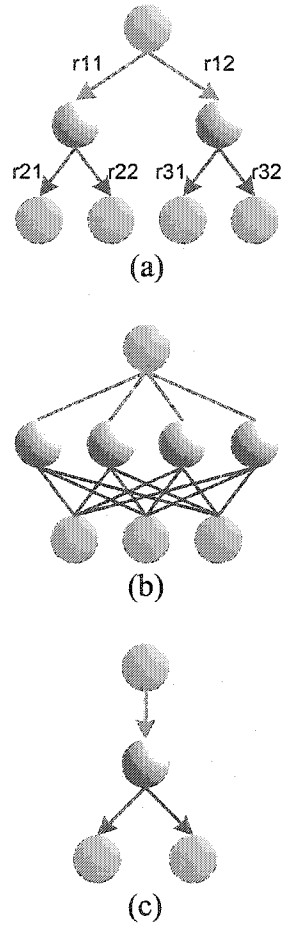
Figure 3.12 A schematic diagrams of AI structures
(a) decision trees (b) neural networks and (c) Bayesian networks
Note that a blue node is a class node (in a decision tree and a Bayesian network) or an output node (in a neural network) and a green node is a leaf node (in a decision tree) or an input node (in a neural network)

Decision trees are often represented as sets of rules. Rule-based approaches are suitable for problems with very strong dependencies. The requirement of the rule-based system is to satisfy detachment and locality, which implies that if there is a rule, then whenever the premise is true, consequence is certainly true regardless of how it was derived, i.e. detachment, and regardless of all other data in the database, i.e. locality. On the other hand, we cannot infer from the previous rule that if consequence is true then premise is true, which reveals the fact that inference in this system is unidirectional (Chong and Wally, 1996). Therefore, it is complicated to properly handle bidirectional reasoning in the system since adding the opposite direction of a rule leads cycle between premise and consequence and eventually an infinite loop (Chong and Wally, 1996; Sangüesa and Burrell, 2000). Rule based approach is computationally efficient but it may not represent some problems under uncertainties because we cannot apply detachment and locality in this situation (Lucas and Gaag, 1991). The system is also unable to express correct probability values. The system becomes more complex as the number of evidences in the problem increases especially dealing with a large number of combinations of premises for each consequence.

3.5.2 Artificial Neural Networks

Artificial neural networks are practical and robust methods in that they can learn real-valued as well as discrete-valued problems (Mitchell, 1997; Ripley, 1996). Artificial neural networks are able to estimate the relationship between input and output without a

70

mathematical formulation or statistical information. These networks are suitable for dealing with incomplete, inconsistent or noisy data. Artificial neural networks are known to be very effective and therefore have been applied to many practical problems including pattern recognition, classification, function approximation, process control, optimization, and prediction (Jain *et al.*, 1996; Ripley, 1996).

Despite many applications of artificial neural networks, the relationships captured by the networks are not easy to understand. This black-box characteristic of the networks makes them less attractive to many researchers. The outputs of the networks can be represented with only either probabilities or variables but not both of them simultaneously. Artificial neural networks incorporate only objective probabilities, therefore, there is no possibility to include an expert's belief.

3.5.3 Bayesian Networks

Bayesian networks offer several merits compared with other AI algorithms. First, Bayesian networks create an understanding of the dependent and independent relationships among variables from their topological structure (Friedman and Koller, 2003). Secondly, Bayesian networks can combine prior knowledge and data (Heckerman, 1995) such that they are able to adopt subjective information elicited from expert's knowledge as well as objective information from data. Moreover, Bayesian network can be constructed automatically from the relationships among variables (Pearl, 1999) and generalize from data despite the limitation of conditional independence. In fact, their

71

assumption, conditional independence, reduces the complexity of learning (Mitchell, 1997). Thirdly, Bayesian networks are flexible in many ways. The networks can readily include additional information into the existing network when updating. Additional data measured on different scales can also be easily adopted. Any subset of the variables in the networks can be treated as inputs and other as outputs (Russell and Norvig, 1995). The output can be either values or probability of each value, or both, to select the best result (Russell and Norvig, 1995). Fourthly, Bayesian networks can handle problem under uncertainty such as missing data (Heckerman, 1995). Even if some inputs are not observed, the networks may be able to accurately predict the target node value from the relationships of other variables. Finally, Bayesian networks are able to provide not only a classification tool but also higher level of explanation, for example, which inputs are informative to reach the conclusion (Ripley, 1996).

However, Bayesian networks have been restricted to problems having variables with discrete values (Mitchell 1997) like decision trees. Although they can handle continuous values in normal distribution, the algorithm for dealing with continuous values has not extensively developed (Russell, 1995). Some researchers pointed out that Bayesian networks might be slow and NP-hard in the worst case (Cheng *et al.*, 2002; Salmerón *et al.*, 2000).

Chapter IV

Stormwater Pollution Modeling

Stormwater runoff, one of the major contaminants to receiving waters, is related to land use activities in urban areas. For example, the stormwater monitoring program in Los Angeles reported that the stormwater quality was significantly different among different urban land uses (LADPW, 2000). Stormwater from residential area is generally better quality than that from other land uses and might contain nutrients, fertilizers, pesticides from grass yards and vegetation (Pitt, 1999; Asaf *et al.*, 2004) and heavy metals from roof materials. Stormwater quality from industrial area varies and case specific, and the pollutant concentration is generally high. Stormwater quality from light industrial area might be similar to stormwater quality from commercial area (Mikkelsen *et al.*, 1994; Asaf *et al.*, 2004). Transportation area such as streets was identified as the highest pollutant loading land use (Bannerman *et al.*, 1993; Arnold and Gibbons, 1996).

Many researchers reported that roof surfaces are also important stormwater pollution source since they are half of the impervious surfaces in urban areas (Förster, 1996). Therefore, urban land use information is important to manage stormwater runoff pollution. In recent years, land use information has been used for stormwater modeling since many land use data sets become more available and accessible (Burian *et al.*, 2002).

In this chapter, we review impact of imperviousness associated with urban land use on stormwater runoff and existing stormwater quantity and quality modeling. In section 4.1, we present the impact of impervious surface on receiving waters, the relationship between impervious surfaces and land use, and the estimates of imperviousness using remotely sensed data. In section 4.2, we review the stormwater runoff models associated with imperviousness. In section 4.3, we explain the concept of event mean concentration (EMC) of stormwater pollutants and its relationship with land use. In section 4.4 we present existing stormwater pollutant loading models related to runoff coefficient and EMCs, and other approaches.

4.1 Imperviousness

Many approaches to monitoring and modeling stormwater runoff have involved estimating impervious surfaces since impervious surfaces increase stormwater runoff (Morgan *et al.*, 1993) and the runoff conveys contaminants into a receiving water body (Arnold *et al.*, 1996). Therefore, impervious surfaces have been an indicator of urbanization and its impact on the receiving water (Arnold *et al.*, 1996; Brabec, 2002).

74

Arnold *et al.* (1996) addressed that impervious surfaces degrade the receiving water body (1) by contributing to the hydrologic changes; (2) by containing land use that generate pollution; (3) by preventing infiltration into ground; and (4) by transporting contaminants into the receiving water. Figure 4.1 shows stormwater runoff associated with impervious surfaces and Figure 4.2 shows an existing relationship between impervious surfaces and the stream quality (Schueler, 1994; Arnold *et al.*, 1996). Stream degradation begins at 10% of imperviousness and becomes severe at 30% of imperviousness. The stream environment is classified into three categories as "protected", "impacted", and "degraded" (Arnold *et al.*, 1996). Brabec (2002) summarized different degradation measures and reported that impervious threshold for degradation was ranged from 4% to 50% depending on the associated type of impact measurement as shown in Table 4.1.

Impervious surfaces can be measured in several ways with different cost and accuracy such as ground surveys, aerial photographs, digital maps, or satellite remote sensing (Stoker, 1998; Sleavin *et al.*, 2000). Other methods employ demographic parameters such as population density, number of households, etc., which are not always applicable to all urban areas (Brabec, 2002). Ground surveys are accurate but time-consuming and labor intensive and therefore they are impractical especially in large study areas. Aerial photographs are accurate but still expensive compared to satellite images. Satellite images have advantages especially for large areas due to their inexpensive data and large coverage. Despite the current limit of pixel resolution for detail features, the accuracy is improving with advances in higher resolution and classification methods.

75
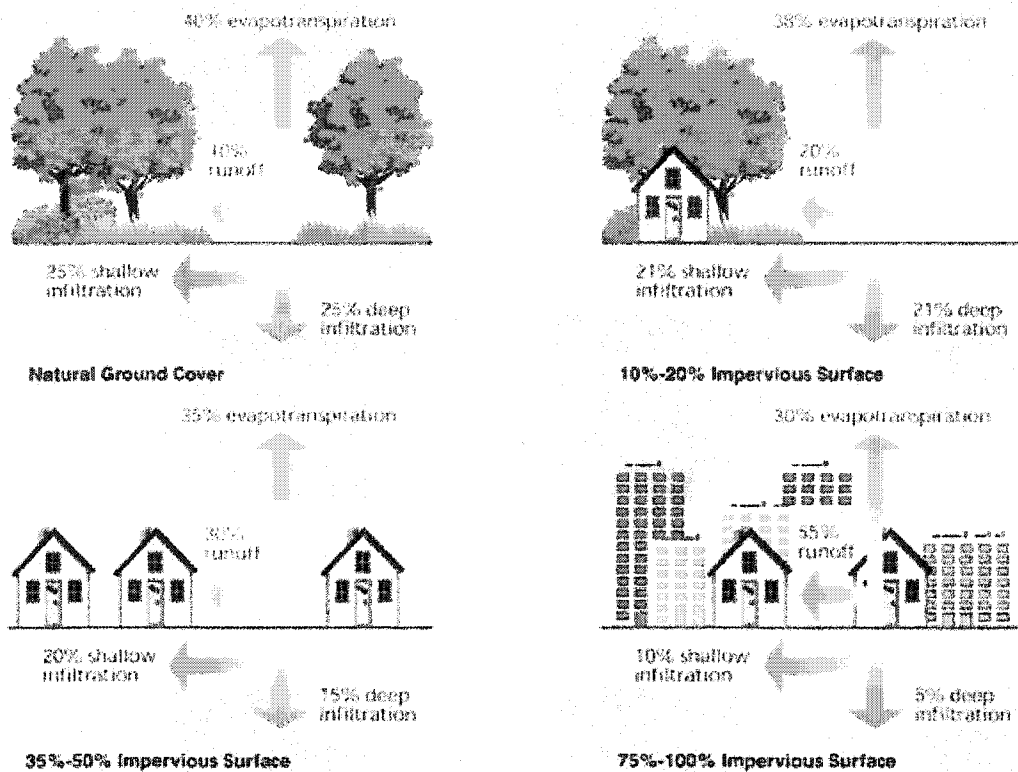
Figure 4.1 Stormwater runoff depending on impervious surface
(adopted from EPA, http://www.epa.gov/owow/nps/nps_edu/urbanx3.htm)
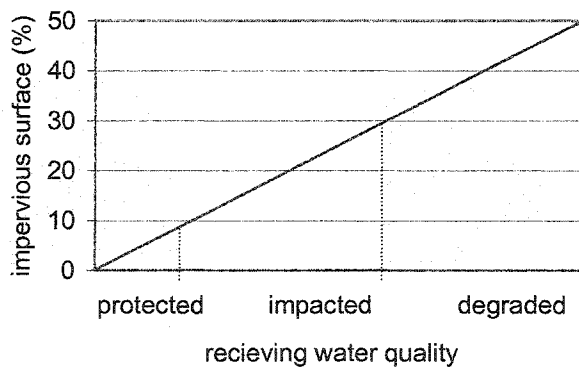
Figure 4.2 A relationship between receiving water quality and percentage of impervious surfaces
(adopted from Arnold *et al.*, 1996)

76

Table 4.1 Selected degradataion measures associated imperviousness

Impact measurement parameter		Impervious threshold (%)
Biotic	Fish diversity	3.6 - 12
Physical	Base flow	45
	Stream flow	21
	Peak flow	4.6
	Sediment	20 - 50
Chemical	Oxygen	7.5 - 43
	Nutrients	42
	Phosphorus	45
	Metals	50
	Zinc	40

(modified from Brabec, 2002)

77

Satellite imagery has been extensively used for impervious surface measurement (Jackson, 1975; Jackson and McCuen, 1979; Morgan et al., 1993; Deguchi, and Sugio, 1994; Civco and Hurd, 1997; Slonecker et al., 2001; Wu and Murray, 2003). A general approach to estimate impervious surfaces is classification from its land use (Jackson, 1975; Cermak et al., 1979; Plunk et al., 1990; Slonecker, 2001; Brabec et al., 2002) because the impervious surface varies with land use (Arnold and Gibbons, 1996). Impervious surfaces mainly consist of rooftop and transportation areas such as highway, road, driveway and parking lot (Schueler, 1994; Arnold and Gibbons, 1996). These impervious surfaces consist of each land use with varied degrees. Jackson (1975) estimated impervious areas from Landsat MSS image using land use averaging and spectral mixing techniques and found that land use mixing technique including band ratio, parallelepiped classification, and principle components analysis was accurate. Cermak et al. (1979) estimated impervious surfaces from land use classification with Landsat MSS image using unsupervised clustering algorithm. Plunk et al. (1990) extracted impervious surfaces from land cover using Landsat TM bands 2, 3, and 4. Table 4.2 displays existing finings on the relationship between imperviousness and land use. This table shows that commercial, industrial and transportation land uses contains high percentage of imperviousness although imperviousness can vary even within a land use type (Brabec et al., 2002).

Recently, a new approach has emerged using sub-pixel analysis for imperviousness measures (Ji and Jensen, 1999; Smith, 2000; Flanagan and Civco, 2001;

78

Table 4.2 Average percentage of imperviousness depending on land use

		Percentage of imperviousness (%)							
		NRCS (1975)	Sullivan et al. (1978)	Alley & Veenhuis (1983)	Stenstrom et al. (1993)	Arnold & Gibbons (1996)	Giannotti et al. (1998)	Cappiella & Brown, (2001)	Bannerman, (2001)
Single-residential (acre lots)	1	20	30		42	20	36		
	0.5	25		15		25			
	0.33	30		26		30		33	41
	0.25	38		39		38			
Multiple- residential		65		66	68	65	85	44	49
Commercial		85	81	88	92 / 80[*1]	85 / 95[*2]	85	72 / 34[*1]	83 / 39[*1] / 92[*2] / 96[*3]
Industrial		72	40	60	91	75		53	69
Other urban					80		50[*4] / 100[*5]		
Open			5		0				

[*1]: public

[*2]: shopping center

[*3]: downtown commercial

[*4]: major roads

[*5]: impervious surfaces, roof and paved

Slonecker *et al.*, 2001; Hung and Ridd, 2002; Yang *et al.*, 2003). This approach resolves mixed pixel problems with the percent imperviousness of a pixel, which characterizes mixed pixels at higher resolution than their pixel resolution (Slonecker *et al.*, 2001).

4.2 Stormwater runoff

As a watershed becomes impervious or urbanized, stormwater runoff volume increases (Riordan *et al.*, 1978; Corbett *et al.*, 1997). One of the main component to determine annual average storm runoff volume is the runoff coefficient, which is defined as the average ratio of runoff to rainfall (Wong *et al.*, 1997). The runoff coefficient represents the fraction of rainfall that actually reaches the receiving water. The runoff coefficient is highly correlated to imperviousness of the area. The following equation is an example of this relationship (Wong *et al.*, 1997):

$$RC = 0.7 \times I + 0.1 \qquad (4.1)$$

where RC is runoff coefficient, I is impervious fraction. Morgan *et al.* (1993) suggested an empirical relationship for the north Texas area:

$$RC_p = 0.3 \times I + 0.2 \qquad (4.2)$$

$$RC_{imp} = I - 0.15 \qquad (4.3)$$

80

where RC_p is runoff coefficient for the pervious areas (I < 50%) and RC_{imp} is runoff coefficient for the impervious areas (I > 50%).

Then the annual average storm runoff volume can be calculated as follows:

$$RV = RC \times A \times CF \times RF \qquad (4.4)$$

where RV is annual storm runoff (m³/yr), A is drainage area (m²), CF is conversion factor, and RF is annual storm rainfall (mm). The following equation is also proposed for calculating annual runoff considering runoff from both impervious and pervious areas (Chiew and McMahon, 1999):

$$RV = RC_{imp} \times A_{imp} \times CF \times RF + RC_p \times (1-A_{imp}) \times CF \times (RF + W) \qquad (4.5)$$

where A_{imp} is fraction of effective impervious areas and W is outdoor water use. In this equation, effective impervious areas are a main factor in calculating runoff volume.

As impervious surfaces can be estimated from land use types in Table 4.2, runoff coefficient can be also estimated from land use of the given area. Corbett *et al.* (1997) found that stormwater runoff volume from urban area was approximately 5.5 times higher than runoff from forest, and runoff coefficient of urban area was 15% higher than runoff coefficient of forest. Table 4.3 presents existing relationship among imperviousness, runoff coefficient and land use in the Santa Monica Bay Watershed.

81

4.3 Event Mean Concentrations

In order to estimate stormwater runoff pollutant loads, EMCs of pollutants have been widely used (Wong *et al.*, 1997) since the concentration of the pollutant considerably varies during an entire storm event. The EMC is an average pollutant concentration during the storm event and defined as the total pollutant mass divided by total runoff volumes as follows (Huber, 1993):

$$EMC = \frac{M}{V} = \frac{\sum C(t)Q(t)}{\sum Q(t)} \qquad (4.6)$$

where M is total mass of pollutant during the storm event (kg), V is total stormwater runoff volume (m^3), C(t) is pollutant concentration varied over time (mg/L), Q(t) is flow over time (L/min), and T is total duration of runoff (min).

EMCs might be dependent on sites and storm events (Smullen *et al.*, 1999). Generally, EMCs are characterized mainly by the land uses in the watershed (Wong *et al.*, 1997). Table 4.4 shows the EMCs related to the land use in the Santa Monica Bay watershed.

82

Table 4.3 Runoff coefficients and imperviousness based on urban land use

Land use	Imperviousness	Runoff Coefficient
single family residential	0.42	0.39
multiple family residential	0.68	0.58
commercial	0.95	0.74
public	0.80	0.66
industrial	0.91	0.74
transportation	0.80	0.66
open	0.00	0.1

(adopted from Stenstrom and Strecker, 1993; Wong *et al.*, 1997)

Table 4.4 Water quality characteristics (EMCs) based on urban land use

Land use	COD	BOD_5	TSS	TKN	$NO_{2\&3}$	TP	Cu	Pb	Zn	O&G
single	140	17	290	4.3	1.85	0.85	0.095	0.350	0.350	3
multiple	130	15	210	2.4	1.00	0.62	0.100	0.440	0.380	22
commercial	90	14	180	2.0	1.20	0.43	0.072	0.225	0.694	22
public	90	14	180	2.0	1.20	0.43	0.072	0.225	0.694	22
industrial	90	14	180	2.0	1.20	0.43	0.072	0.225	0.694	22
transportation	130	15	210	2.4	1.00	0.62	0.100	0.440	0.380	22
open	95	2	490	2.8	1.45	0.52	0.055	0.140	0.440	0

(adopted from Stenstrom and Strecker, 1993; Wong *et al.*, 1997)

4.4 Pollutant loading

Stormwater runoff carries various pollutants including suspended solids, organic matters such as BOD and COD, inorganic materials, nutrients, heavy metals, and fecal coliforms (Corbett, 1997). TSS, BOD_5 and COD have been used to estimate urban stormwater runoff quality (Shinya *et al.*, 2003). Tong and Chen (2002) reported that nitrogen and phosphorus loads were generated mainly from agriculture and impervious urban areas than other land uses. Asaf *et al.* (2004) reported that volatile organic compound (VOC) concentrations generated from the industrial area are higher than those form residential area. Heavy metals are also one of the main pollutants in stormwater runoff from urban area (Macdonald *et al.*, 1997; Yuan *et al.*, 2001). Heavy metals are mostly generated from transportation area such as highway, street and parking lots (Morrison *et al.*, 1984; Yuan *et al.*, 2001) and roof materials (Gromaire *et al.*, 2001). For example, major sources of lead are roofing and paint materials (Davis *et al.*, 2001; Asaf *et al.*, 2004) and vehicular emission since lead has been used for a vehicle fuel additive (Yuan et al., 2001). Major sources of cupper and zinc are reported to be roofing materials and drain water systems due to the corrosion (Förster, 1996; Gromaire *et al.*, 2001; He *et al.*, 2001). Fecal coliforms are mainly transported from urban land uses despite seasonal variation (Gannon and Busse, 1989; Kelsey *et al.*, 2004).

Many researchers have tried to develop empirical models for stormwater runoff pollutant loads based on land use data (Stenstrom *et al.*, 1984; Stenstrom and Strecker, 1993; Wong *et al.*, 1997; Burian *et al.*, 2002; Ackerman and Schiff, 2003). If we have

84

information of runoff volume and EMCs for each pollutant, we can obtain annual pollutant load from the following equation (Wong et al., 1997):

$$PL_i = RV \times EMC_i \qquad (4.7)$$

where PL_i and EMC_i are annual pollutant loading and the EMCs for pollutant i, respectively. If we assume that the rainfall and annual number of storms are equal for all pixels, the following equation represents pollutant loads per unit pixel and unit rainfall:

$$PL_i = \alpha \times RC \times EMC_i \qquad (4.8)$$

where α is a normalization factor that depends on units and conversion factors.

When considering dry weather pollution, the following equation can be used to calculate stormwater pollutant loads (Chiew and McMahon, 1999):

$$PL_i = RV \times EMC_i + BF \times DWC \qquad (4.9)$$

where BF is baseflow and DWC is dry weather concentration. Dry weather flow pollutant load was reported to be significant with lower precipitation (McPherson et al., 2002).

Some researchers suggested simple empirical models of stormwater pollutant loads (Chiew and McMahon, 1999; Vaze and Chiew, 2003):

$$PL = \alpha \times RV^{\beta} \tag{4.10}$$

$$PL = \alpha \times RF^{\beta} \tag{4.11}$$

$$PL = \alpha \times \sum (1 - RI_{min})^{\beta} \tag{4.12}$$

$$PL = \alpha \times \sum (1 - RR_{min})^{\beta} \tag{4.13}$$

where α and β are optimizing parameters, RI_{min} is minimum rainfall intensity, and RR_{min} is minimum runoff rate. Some researchers reported that stormwater pollutant loads are positively related to the antecedent dry period (Brezonic and Stadelmann, 2002; Vaze and Chiew, 2002), which confirms first flush phenomenon.

86

Chapter V

Study Area and Data

This chapter describes our methodology to classify satellite image for urban land use and stormwater pollutant loading. In section 5.1, we describe our study area, Marina del Rey in the Santa Monica Bay Watershed. In section 5.2, we explain satellite imagery used in this study: Landsat TM and ETM$^+$ images. The ground truth data are also given for both image classification. In section 5.3 ancillary data such as locational data, digital elevation model (DEM), and vegetation indices such as normalized difference vegetation index (NDVI) and Tasseled Cap Transformation are presented.

87

5.1 Study Area

The study area was focused on Marina del Rey and its vicinity (latitudes 33° 56' 42"-33° 59' 45", longitudes 118°24' 42"-118°27' 34") as shown in Figure 5.1. This area includes Ballona Wetlands located in the southern part of the Santa Monica Bay watershed. Santa Monica Bay is popular recreational resources and important ecological resources of natural habitat (Dojiri *et al.*, 2003). Santa Monica Bay watershed covers total area of 1,465 km^2 (EPA) from Point Dume on the north to Palos Verdes point on the south (Wong *et al.*, 1997) as shown in Figure 5.2. It receives various contaminants including municipal and industrial wastewater and stormwater runoff from the City of Los Angeles that impair the water quality. Improved sewage treatment plants in the watershed have reduced the pollutant emissions from wastewater discharges (Bay *et al.*, 2003). Therefore, emissions from stormwater runoff are receiving more concern since stormwater runoff constitutes the dominant source of contaminants (Bay *et al.*, 2003).

As shown in Figure 5.2, there are two main drainages to Santa Monica Bay: Ballona Creek and Malibu Creek. The Malibu Creek watershed is located in the northern part of the Bay (Bay *et al.*, 2003). Majority of the watershed is open area (Dojiri *et al.*, 2003) i.e. 9% of residential, 1% of commercial and industrial respectively, and 88% of open area (Bay *et al.*, 1999). Conversely, the Ballona Creek watershed is located in the south central part of the Bay and most of the watershed is urban area, i.e. 83% of the watershed is urbanized (Bay *et al.*, 1999; Dojiri *et al.*, 2003). The watershed consists of 64% of residential, 8% of commercial, 4% of industrial, and 17% of open areas (Bay *et*

88

Figure 5.1 Study area of Marina del Rey and its vicinity
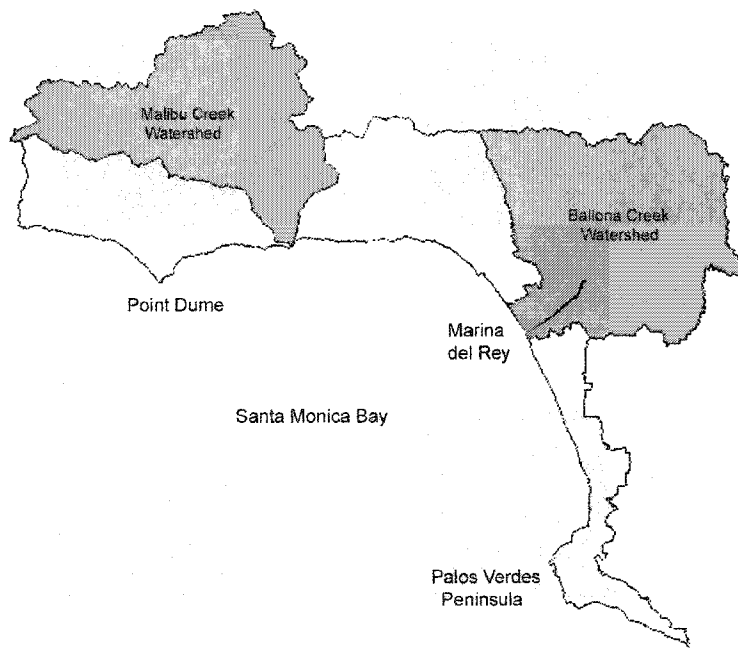
Figure 5.2 Location of Marina del Rey and Santa Monica Bay watershed.
Note that the red box area is the study area

89

al., 1999). Therefore stormwater runoff from Ballona Creek watershed more impairs the receiving water quality than that from Malibu Creek watershed (Bay *et al.*, 2003a; Dojiri *et al.*, 2003). Moreover the runoff from Ballona Creek watershed has been increased during the last century due to the population growth and an increase in impervious surfaces. Accordingly, pollutant emissions from the watershed have been reported to increase (Dojiri *et al.*, 2003).

The south of Ballona Creek is receiving growing concern due to the importance of the Ballona Wetlands (Ballona Wetlands Foundation). The Ballona Wetlands have been degraded by urban growth especially the development of Marina del Rey and its vicinity. Its freshwater marsh is important for wildlife habitat and water quality. In order to preserve the marsh, it is important to properly monitor and manage stormwater quality and urban features since the freshwater marsh is fed mainly by the stormwater runoff from urban area.

The size of the study area is approximately 25 km^2 that is including Ballona Wetland with area of 8 km^2. The surrounding urban area is dominated by residential land use but contains other urban features such as commercial, industrial, and transportation land uses. This area has a Mediterranean climate with an average annual rainfall of 360 mm (NOAA), which occurs mainly during winter. The area is relatively flat with the maximum elevation of approximately 60 m.

5.2 Satellite Imagery

5.2.1 Landsat TM

We used Landsat Thematic Mapper (TM) images (obtained on September 3, 1990, path 41 row 36) for classification. Figure 5.3 shows three visible color composite of the study area. A subimage of the study area, which consists of 173 pixels by 227 pixels as columns and rows respectively, was extracted from all bands 1 to 7. The spatial resolution of the image was 25 m^2 instead of 30 m^2 due to the post-processing of edge enhancement.

Figure 5.4 shows the distribution of each TM band and the distribution of each band is mostly skewed. Band 6 has narrow range with high peak since it is thermal band. Table 5.1 gives statistical information of each Landsat TM band and Table 5.2 shows correlation of each band. The table shows that bands 1, 2 and 3 are highly correlated, since they are visible bands. Bands 5 and 7 also show high correlation since both are middle infrared.

Official land use data obtained from Southern California Association of Governments (SCAG, 1993) was used as ground reference for both training data and test data for accuracy assessment. Figure 5.5 shows the SCAG land uses in the given study area. This land use data had been resampled to match the resolution of TM image using GIS (Lee, 2003). The data contained 26,614 homogeneous land use pixels.

91

Figure 5.3 Three color composite of Landsat TM visible bands for the study area
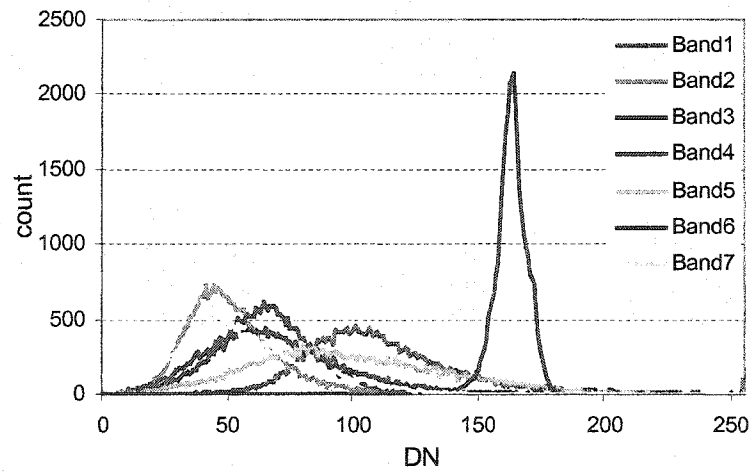
Figure 5.4 Distribution of Landsat TM bands

Table 5.1 Statistics of Landsat TM data

Input	Min	Max	Median	Mean	Standard deviation
band1	5	255	107	113	33
band2	0	255	49	52	19
band3	0	255	65	68	29
band4	0	255	65	66	23
band5	0	255	94	98	39
band6	134	193	163	163	6
band7	0	255	53	56	26

Table 5.2 Correlation of Landsat TM bands

	band1	band2	band3	band4	band5	band6	band7
band1	1						
band2	0.91	1					
band3	0.87	0.95	1				
band4	0.40	0.53	0.52	1			
band5	0.30	0.37	0.48	0.43	1		
band6	0.06	0.08	0.14	-0.02	0.29	1	
band7	0.40	0.47	0.57	0.33	0.88	0.34	1

Land use

single-family residential
multiple-family residential
commercial
public
industrial
transportation
open
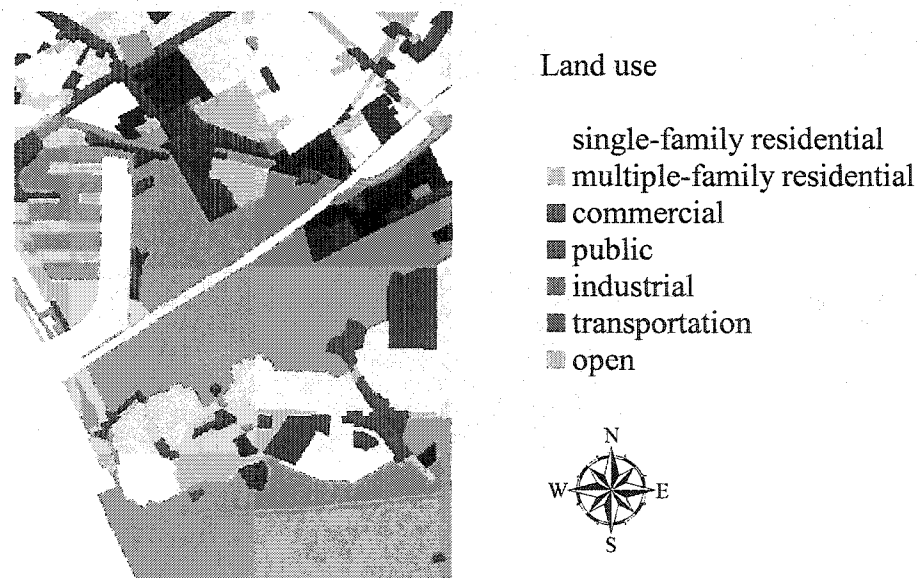
Figure 5.5 Southern California Association of Governments land use data (1993)

93

5.2.2 Landsat ETM[+]

A Landsat ETM[+] image (obtained on August 11, 2002, path 41 row 36) was also used. All bands except panchromatic band were examined for our study. A subimage of the study area is shown in Figure 5.6, which was extracted from bands 1 to 7 consisting of 153 pixels by 200 pixels as columns and rows respectively. The total number of pixel of the subimage was 30,600 but the pixels that belong to water was excluded using masking. Therefore, total number of land use pixel in the given area was 27,416. The spatial resolution of band 1 to 5 and 7 was 28.5 m^2 whereas band 6 (thermal band) has the spatial resolution of 57 m^2. The thermal band was resampled to match the resolution of 28.5 m^2 based on nearest neighbor. The training data and testing data for accuracy assessment were collected from all classes to avoid undersampling the small classes (Jensen, 1996). The total number of training data pixels was 2,067 and the total number of test data pixels was 1,033, which correspond to 8.5 % and 4.3 % of total data respectively. Figure 5.7 shows the distribution of each band and the detail information of satellite image data is given in Table 5.3. Table 5.4 provides the correlation of each band and visible bands and middle infrared bands exhibit high correlation.

As ground reference data, official land use obtained from SCAG (2003), aerial photo, and field survey were used for both training data and test data for accuracy assessment. Figure 5.8 shows the SCAG land use data. Some land use has been changed compared to 1993 data, i.e. public from open land use next to Loyola Marymount University and commercial converted from open land use on Washington Blvd.
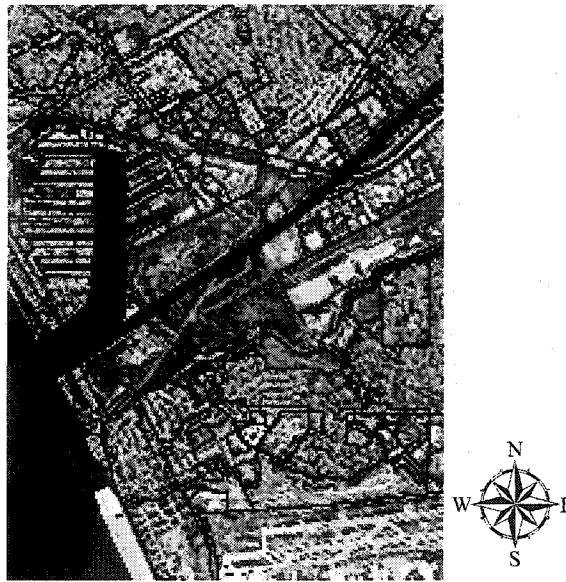
94
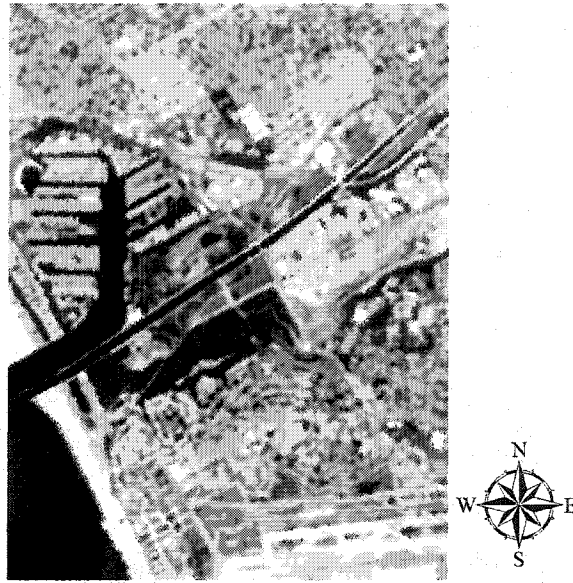
Figure 5.6 Three color composite of Landsat ETM$^+$ visible bands for the study area
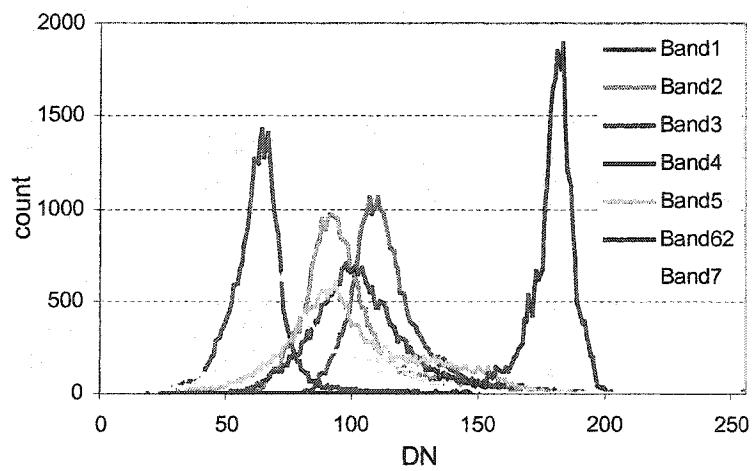
Figure 5.7 Distribution of Landsat ETM$^+$ bands

Table 5.3 Statistics of Landsat EMT$^+$ data

Input	Min	Max	Median	Mean	Standard Deviation
band1	80	255	111	114	15
band2	50	255	94	96	17
band3	43	255	102	105	22
band4	20	155	63	63	12
band5	9	255	96	101	28
band6	138	210	180	179	9
band7	8	255	75	79	25

Table 5.4 Correlation of Landsat ETM$^+$ bands

	band1	band2	Band3	band4	band5	band6	band7
band1	1						
band2	0.98	1					
band3	0.95	0.97	1				
band4	0.25	0.34	0.31	1			
band5	0.24	0.32	0.47	0.36	1		
band6	0.21	0.20	0.27	-0.18	0.25	1	
band7	0.51	0.58	0.70	0.25	0.90	0.30	1

Land use

    single-family residential
    multiple-family residential
    commercial
    public
    industrial
    transportation
    open
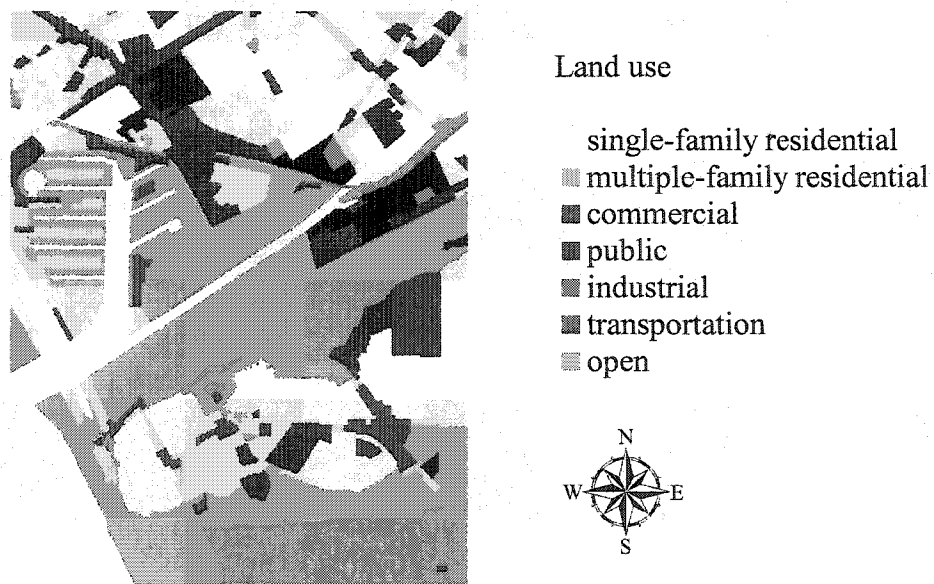
Figure 5.8 SCAG land use data of the study area (2001)

96

5.3 Ancillary Data

5.3.1 Locational Data

In order to incorporate horizontal spatial information of the study area, locational ancillary data was created, i.e. X and Y coordinate values. The coordinate values of each pixel were obtained from Landsat images using RSI ENVI 4.0 based on WGS 1984 UTM (zone 11N). The unit of the coordinate values were meters. The purpose of the ancillary data was to check whether including locational information improves the classification accuracy and to compare the result with the classification using spectral data only. Table 5.5 shows the statistics of the locational data of the study area.

5.3.2 Digital Elevation Models

To investigate the effect of the vertical elevational information on classification, we used Shuttle Radar Topography Mission (SRTM) data and National Elevation Dataset (NED) as DEM ancillary data.

Shuttle Radar Topography Mission

SRTM collected elevation radar data on a near-global scale (NASA JPL). The mission was cooperated by NASA, the National Imagery and Mapping Agency (NIMA)

97

Table 5.5 Statistics of geospatial ancillary data

Input	Min	Max	Mean	Mode	Median
X	365213.25	369545.25	367550.66	367692.75	367578.75
Y	3756969.75	3762641.25	3759933.66	3762641.25	3759962.25
SRTM	-8388607	61	-163983	11	12
NED	-0.1997	57.1982	16.70158	2.27	6.95925

Figure 5.9 SRTM image of the Marina del Rey
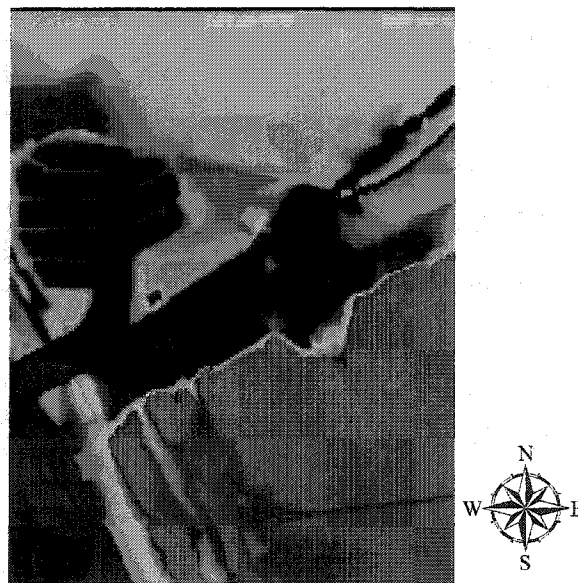
Figure 5.10 NED image of the Marina del Rey

98

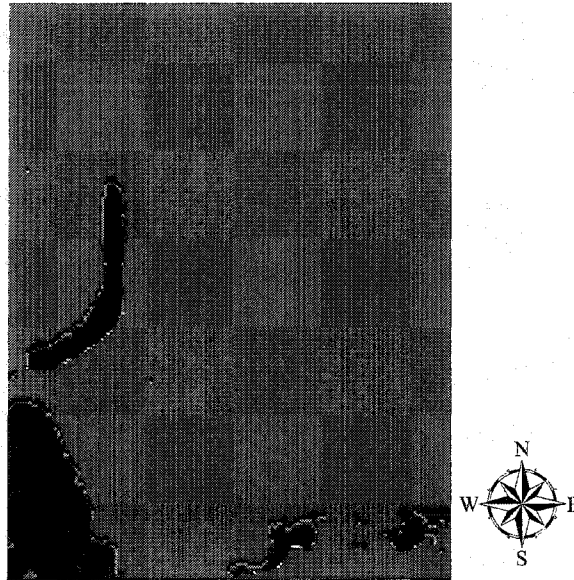of the U.S. Department of Defense and the German and Italian space agencies and is managed by NASA's Jet Propulsion Laboratory (JPL). The purpose of SRTM is to produce high-resolution digital topographic information of Earth. A radar system onboard Space Shuttle Endeavour was launched on February 11, 2000 and the mission lasted for 11 days. SRTM used radar interferometry to collect elevation data. Radar interferometry uses two radar images taken from different locations, from which surface elevation is calculated. By using interferometry, SRTM collected data over 80% of Earth's land mass.

The SRTM data were resized to match the Landsat image size of the study area. The resized data were resampled based on nearest neighbor to meet the resolution of Landsat image. Table 5.5 shows the statistics of the SRTM data. In this table, the negative elevation is below the sea level. Figure 5.9 shows the color coded SRTM image of the study area. This figure shows that the red color areas have higher elevation and the Santa Monica Bay has negative values.

National Elevation Dataset

NED is a raster elevation data assembled by the U.S. Geological Survey (USGS) (Gesch *et al.*, 2002). The purpose of NED is to provide nation wide coverage of elevation data in a seamless format with consistent bases. Tiled DEM data were assembled with edge matching to minimize artifacts of adjacent tiles. Removing artifact improves slope and shaded-relief, and accordingly information that derived from the elevation data. NED has a resolution of approximately 30 m for the continental United States, and Hawaii, and

99

60 m for Alaska. The horizontal datum is the North American Datum of 1983 (NAD 83), and the vertical datum is the North American Vertical Datum of 1988 (NAVD 88), in which the elevation values are converted to decimal meters. The projection is converted from UTM coordinate system of the source DEM to a geographic coordinate system, i.e. decimal degrees of latitude and longitude.

NED data used here was obtained on February 11, 1999 (USGS). The image was reprojected and resized to be consistent with the Landsat image of the study area and resampled using nearest neighbor since it's resolution is 30 m that is slightly different from the resolution of Landsat ETM$^+$ image (28.5m). As shown in Table 5.5, the maximum of elevation of the study area is 57m and the elevation is not much varied. The elevation below the sea level is close to 0. Figure 5.10 shows the NED image of the study area. This figure shows that the elevation of the southern part adjacent to Ballona Wetlands is higher than northern part.

5.3.3. Vegetation Index

Normalized Difference Vegetation Index

Vegetation index was calculated in order to investigate whether it assists in improving classification accuracy. NDVI was calculated using RSI ENVI 4.0 from Landsat ETM$^+$ image. The resulting NDVI image of the study area Figure 5.11, which is color coded. In this figure, green color areas are vegetated areas corresponding to open

100

areas such as Ballona Wetlands, Westchester recreation center and other parks, and single-family residential areas. Red color areas are nonvegetated or impervious surfaces such as Los Angeles International airport, commercial and industrial areas. Dockweiler state beach and the open area under construction also exhibit redish color although they belong to open land use.

Tasseled Cap Transformation

Tasseled Cap transformation was calculated using RSI ENVI 4.0 in order to assist in selecting training data and to improve classification accuracy. Figure 5.12 shows the resulting Tasseled Cap transformation of the study area where brightness, greenness and wetness were represented by red, green and blue respectively. Therefore, highly impervious surfaces such as Los Angeles International airport, commercial and industrial areas show red due to their high brightness values. Dockweiler state beach also exhibites high brightness in this case. On the other hand, vegetated areas in open land use such as Ballona Wetlands, Westchester recreational center and parks, and in other land uses such as residential areas and Loyola Marymount University are green and Santa Monica Bay and fresh water marsh of Ballona Wetlands are blue due to their water content.

101
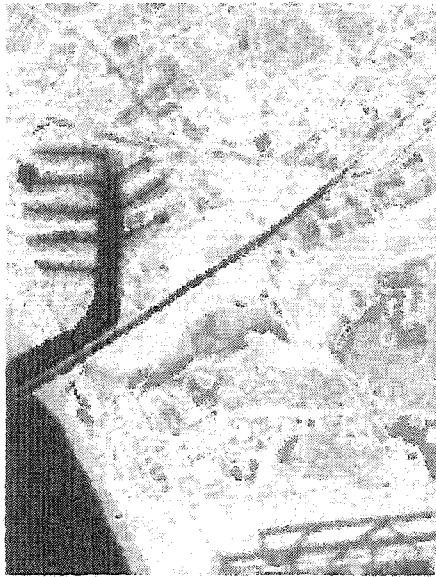
Figure 5.11 Color coded NDVI of the Marina del Rey
Green: vegetated, red: non vegetated
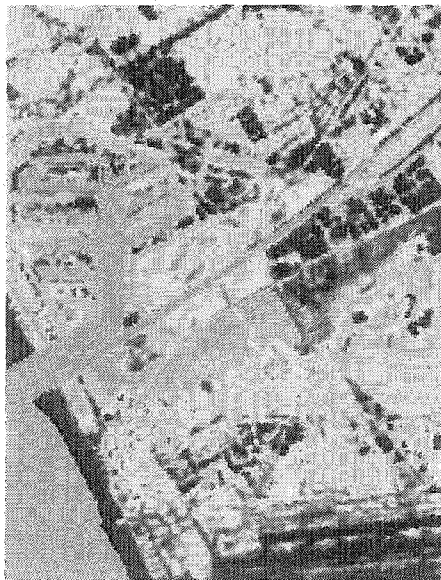
Figure 5.12 Tasseled cap image of the Marina del Rey
red: brightness, green: greenness, blue: wetness

102

Chapter VI

Urban Land Use Classification

Many approaches to land use classification with satellite imagery have been widely explored to assist urban planning and environment management. For instance, Landsat series imagery has been extensively used for land cover/use analysis (Haack, 1983; Haack et al., 1987; Pax-Lenney, 1997; Ridd and Liu, 1998; Stefanov, 2001; Clapham, 2003; Wilson et al., 2003). However, the use of Landsat imagery for urban areas poses problems since urban areas are highly heterogeneous. Most pixels in the Landsat imagery contain a mixture of different urban features with different proportions. For example, a pixel of residential land use could contain mixed signatures of buildings, pavement, driveways and vegetation such as grass yards and trees (Clapham, 2003). Another problem with urban classification arises from spectral signature similarity among different land uses. For example, pixels of commercial and industrial land uses could

103

exhibit similar spectral signatures if the pixels consist of similar roof materials (Stefanov, 2001). Therefore, a method to resolve these problems is needed.

In this chapter, we evaluate the performance of Bayesian networks to classify urban land use from Landsat images. In section 6.1, we investigate the optimal conditions of discretization and training data size for Bayesian network classification. In section 6.2, the performance of different network structures is compared for classification and the most informative input variables are identified. In section 6.3, we investigate the effect of incorporating ancillary data on accuracy. In section 6.4, we discuss the performance of Bayesian networks. We aimed to gain an understanding of the nature of urban land use classification.

6.1 Optimal Conditions for Bayesian Networks Performance

Landsat TM image data were used to investigate the effect of discretization and training data size on classification. Although the resolution allows at most level II classification (Anderson *et al.*, 1976; Ridd, 1995), it is consistent with our categories of classification. The categories we used were urban classification level II based on U.S. Geological Survey (USGS) classification system (Anderson *et al.*, 1976): residential, commercial, industrial, transportation, and open area.

104

6.1.1 Training Sample Size

Bayesian networks require prior knowledge of the study area and data set to select representative training data. In particular, training data size, which is related to the number of spectral bands and statistical properties, affects the accuracy of classification. It is important to select sufficient training data size for proper training (Hay, 1979; Foody, 2002) and homogeneous training data representing each class (Richards and Jia, 1999; Schowengerdt, 1997). The accuracy of the classification usually increases with large number of training data (Dobbertin and Biging, 1996) but large training data sets are not always affordable. There is no universal rule for selecting training data size and the required training data size from the same imagery could vary for different classifiers. For example, the maximum likelihood algorithm is known to require a smaller training data set than neural networks (Swain, 1978; Paola and Schowengerdt, 1995), although some researchers reported that neural networks needed smaller training data sets (Hepner *et al.*, 1990; Foody *et al.* 1995a; Paola and Schowengerdt, 1995). Presently there is very limited research that examines the effect of training data size on Bayesian network classification with satellite imagery. Therefore, we tried to find the sufficient training data size for Bayesian network classification. If Bayesian networks can be properly trained with small training data sets, they become more attractive as an alternative classifier of satellite imagery.

In order to test the effect of training data size, we used ten subsets of training data with sizes of 400, 800, 1600, 2000, 4000, 8000, 16000 and 20000. The proportion of each

105

land use is consistent for all different training data size. All input data were discretized into 15 values based on equal width intervals. A total of 400 pixels of test data were selected for accuracy assessment, which did not include any pixels from the training data, and were fixed for each case. These conditions are summarized in Table 6.1.

Figure 6.1 shows the relationships between training data size and accuracies. The results show that accuracy increases with the size of the training data, up to 2,000 or 4,000 pixels. The increase in accuracy with increasing numbers of pixels is monotone, except for 1,600 (MWSTs) and 1,200 (naïve Bayesian classifiers) pixels. The reason for the decline is not known, but may be an artifact of the specific data set at those pixels. For MWSTs with locational ancillary data, the accuracy and κ coefficient converged to approximately 76% and 65% respectively at the 2,000 pixels, which is 7.5% of total data. For naïve Bayesian classifiers with locational ancillary data, the accuracy and κ coefficient converged to approximately 74% and 62 % respectively at 4,000 pixels. The increase in accuracy between 2,000 and 4,000 pixels is only 1 to 2 %. These results indicate that a training data set size of 2,000 pixels, or 7.5% of the data set, will be sufficient for both networks when incorporating locational ancillary data.

The results without locational ancillary data show no significant difference in accuracies as the training data set increased from 800 to 20,000 pixels. Interestingly, both networks produced better accuracies with small training data size of 800 pixels than those with large training data size. The results exhibited a similar decrease in accuracy when using 1200 (MWSTs) and 1600 (naïve Bayesian classifiers) pixels. The accuracies

106

Table 6.1 Summary of test condition for the effect of discretization and training data set

	training data size	discretization methods	variable states
training data size	400 - 20,000	4,000	4,000
testing data size	400	400	400
discretization methods	equal width	equal width equal frequency standard deviation	equal width
input variable states	15	10 7	5 - 50
class variable states	5	5	5
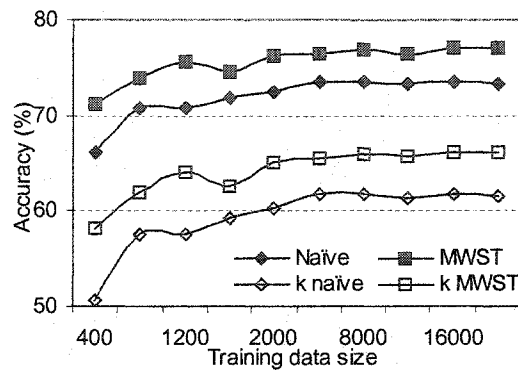
(a)
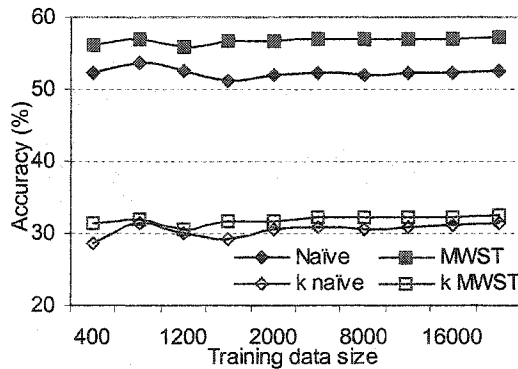
(b)

Figure 6.1 Bayesian network performance depending on training data size
(a) with spectral and geospatial data (b) with spectral data only
Note that solid lines with closed markers represent overall accuracy and those
with open markers represent kappa coefficient.

107

become almost constant when using more than 2,000 pixels. Overall, MWSTs slightly outperformed naïve Bayesian classifiers.

The number of training pixels may depend more on the reliability of training as opposed to the final accuracy. The accuracy increased by only 5 to 6% as the pixels increased from 400 to 2,000 then reaching a plateau. The size of the required training data set for these examples is comparable to Mather's (1999) suggestion. Using equation 2.2, the required size of the training dataset is 1,050 (e.g., $30 \times 7 \times 5$). The anomaly in accuracy from 1,200 to 1,600 pixels may be an artifact of the randomly selected pixels. Therefore choosing a sufficient number of pixels may be more important in avoiding anomalies than in maximizing accuracy.

6.1.2 Discretization Methods

We compared three different discretization methods widely used in classification: equal width interval, equal frequency interval and standard deviation breaks. The equal width interval method breaks the range of the observed data values with equally sized intervals whereas the equal frequency interval method divides the range with unequally sized intervals (Liu *et al.*, 2002; Yang and Webb, 2003). In case of satellite imagery, there are so many pixels with equal DN values that it is impossible to create intervals containing equal numbers of observations without splitting pixels with equal DN values into different intervals. To avoid this problem, the intervals may not have equal numbers of observations. The standard deviation method breaks the range at standard deviation

108

intervals from the mean (Minami, 2000). For example, each interval is calculated as mean $\pm \alpha$·standard deviation and α is a scale factor, e.g. 0.25, 0.5, or 1, that can be determined by the number of intervals. Those observations beyond a certain number of intervals from the mean can be aggregated.

For testing different discretization methods, training data size was set to 4,000, which corresponds to 15% of total data. Total test data was fixed to 400 for accuracy assessment. In this case, we discretized input data to 10 states for equal width and equal frequency interval and 7 states for standard deviation breaks with α equal to 1 for spectral signature data.

Figure 6.2 presents the relationships of discretization methods and the accuracies. In this figure, the equal frequency interval method gave the best accuracy among the three methods. The differences among discretization methods were 2 to 7% in overall accuracy and 3 to 11% for κ coefficient. This appears to agree with our intuition in that the distribution has a bell shape as shown in Figure 5.4. The equal frequency interval method better represented the data than the other methods, which placed a large number of pixels in just a few intervals. The equal frequency interval method also benefited more with the inclusion of locational ancillary data. The largest differences in MWSTs' overall accuracies and κ coefficients were 7% and 11%, respectively. For naïve Bayesian classifiers, the largest differences were only 3% and 4% for overall accuracy and κ coefficient, respectively.

When considering the classification accuracies with spectral data only, the difference between three discretization methods became trivial. The difference of overall

109
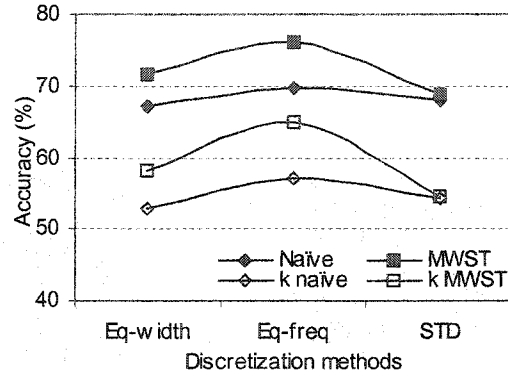
(a)
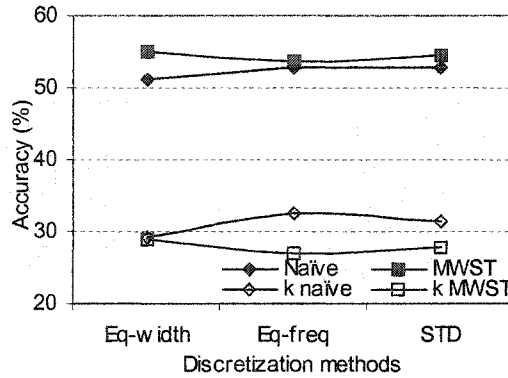
(b)

Figure 6.2 Bayesian network performance depending on discretization methods
(a) with spectral and geospatial data (b) with spectral data only
Note that solid lines with closed markers represent overall accuracy and those
with open markers represent kappa coefficient.

110

accuracy was only 1 to 2 % for both networks. Similarly, the difference in κ coefficient was only 1 to 3%. In the absence of locational ancillary data, the differences in accuracy using different discretization methods were small.

6.1.3 Number of Input Variable States

The number of input variable states is known to affect Bayesian network performance. A small number of states may improve the accuracy but may lose information from the original distribution. A large number of states can represent the original distribution more accurately but may suffer from accuracy and complexity. There is no study that systematically evaluates the optimal number of input variable states. Therefore, we try to empirically determine an optimal number of input variable states for discretization. The number of variable states may different depending on input data characteristics, but may be constant for our application since many satellite multi-spectral signature consists of 8-bit information.

The number of input variable states was varied from 5, 10, 15, 20, 25, and 50. Each variable consists of 8-bit spectral signature and we fixed the number of states for all variables in each case. The locational ancillary data was also discretized in the same way when applicable. We used a fixed size of 4,000 training data based on equal width interval to investigate the effect of the number of the input variable states. For accuracy assessment, 400 data were selected for a consistent test.

111

Figure 6.3 illustrates the relationships between the number of input states and the classification accuracy. The accuracy increased as the number of input variables states increased. This was especially true for the first half of the subsets.

For MWSTs incorporating locational data, the overall accuracy and κ coefficient increased monotonically with the number of input variable states, except for the anomalous results at 20 states. The naïve Bayesian classifiers achieved the peak accuracies with the 15 states and then decreased.

For the classification with spectral data only, larger numbers of input variable states actually lowered the accuracy. Overall accuracy of both networks was greatest using 15 states. The greatest value of κ coefficient was achieved with 15 states for MWSTs. Naïve Bayesian classifiers had the best accuracy with 20 states, although this accuracy was almost the same as the accuracy with 15 states.

These results show that fifteen states were the most effective. Slightly higher accuracy achieved with 25 and 50 states for MWSTs, but the increase was trivial. The κ coefficient for naïve Bayesian classifier was slightly greater using 20 states. For these examples, 15 states are optimal for classifying the satellite image data consisting of 8-bit DN values.

112
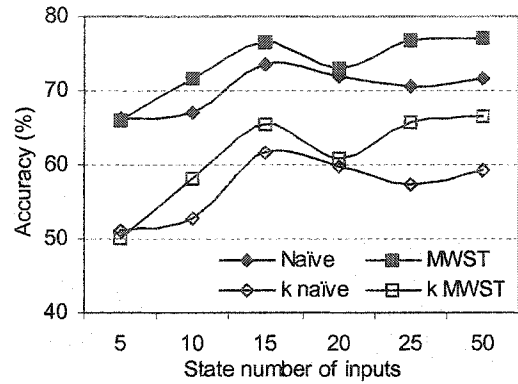
(a)
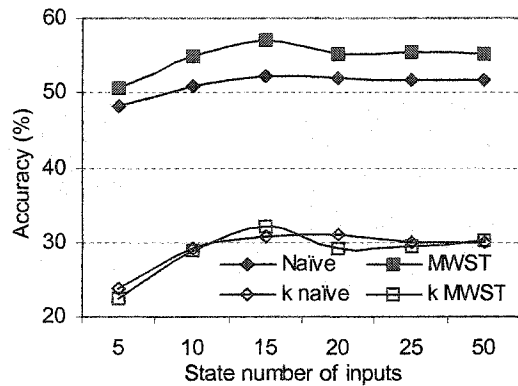
(b)

Figure 6.3 Bayesian network performance depending on the number of variable states
(a) with spectral and geospatial data (b) with spectral data only
Note that solid lines with closed markers represent overall accuracy and those
with open markers represent kappa coefficient.

113

6.2 Bayesian Network Classification

6.2.1 Characteristics of Data

For urban land use classification, we used Landsat ETM$^+$ images. We used the land use categories based on U.S. Geological Survey (USGS) level II classification system for urban (Anderson *et al.*, 1976): residential, commercial, industrial, transportation, and open area. Furthermore, we also used level III classification system: single-family residential, multi-family residential, commercial, public, light industrial, transportation, and open area.

Table 6.2 shows the percentage of each class in the training data set. All classes were collected from each class to avoid undersampling (Jensen, 2000). The statistics of each land use in training data are given in Table 6.3. The tables show that range of each land use category overlapped with other categories in each band.

Figure 6.4 shows distribution of pixels of each land use category in two dimensional spectral space. In this figure, none of the classes were separated from other classes in the space of highly correlated bands, e.g. bands 1, 2 and 3, bands 5 and 7, and bands 61 and 62.

The spectral separability of each class was calculated based on the Jeffries-Matusita equation (Richards and Jia, 1999). The potential separability range is from 0.0 to 2.0, and values greater than 1.9 indicate that the pairs have good separability. Table 6.4 shows the separability of each class in level II and III classification. Pairs of multiple-

114

Table 6.2 Ratios of land uses in the training data set

USGS level II class	USGS level III class	Ratio (%)
Residential (11)	single-family residential (111)	29
	multiple-family residential (112)	9
Commercial (12)	commercial (121)	6
	public (122)	5
Industrial (13)	light industrial (13)	5
Transportation(14)	transportation (14)	11
Open	agriculture (2) parks(122)	35
	Total	100

Note that ( ) is the USGS land use category.

Table 6.3 Statistics of training data

(a) Range of DN values of each class in level III classification system

Band	S	M	C	P	I	T	O
B1	91-124	94-192	104-211	95-214	114-228	107-181	81-183
B2	71-110	71-1982	82-192	71-210	91-216	86-168	62-160
B3	67-124	68-206	83-220	69-247	95-253	84-190	52-190
B4	53-87	46-106	43-106	35-108	44-123	41-89	40-155
B5	53-133	52-202	61-255	39-187	65-255	80-161	30-192
B61	81-131	74-131	63-159	67-170	88-152	63-131	49-145
B62	166-192	163-191	158-205	160-210	169-201	157-192	150-199
B7	38-105	38-194	46-255	37-170	55-255	68-158	23-174

(b) Mean ± std of DN values of each class in level III classification system

Band	S	M	C	P	I	T	O
B1	107±6	124±16	133±20	132±19	144±18	144±19	104±14
B2	90±7	105±18	114±21	113±21	126±20	129±20	88±18
B3	94±9	112±22	123±26	124±27	138±25	144±25	100±25
B4	68±5	62±10	61±11	64±13	65±12	66±9	67±15
B5	92±10	89±19	97±31	98±21	103±27	122±17	128±27
B61	107±7	105±10	116±15	120±17	126±12	110±9	108±20
B62	179±3	178±5	184±8	186±8	189±6	181±5	180±10
B7	69±10	72±17	83±33	81±18	88±26	113±20	94±27

115
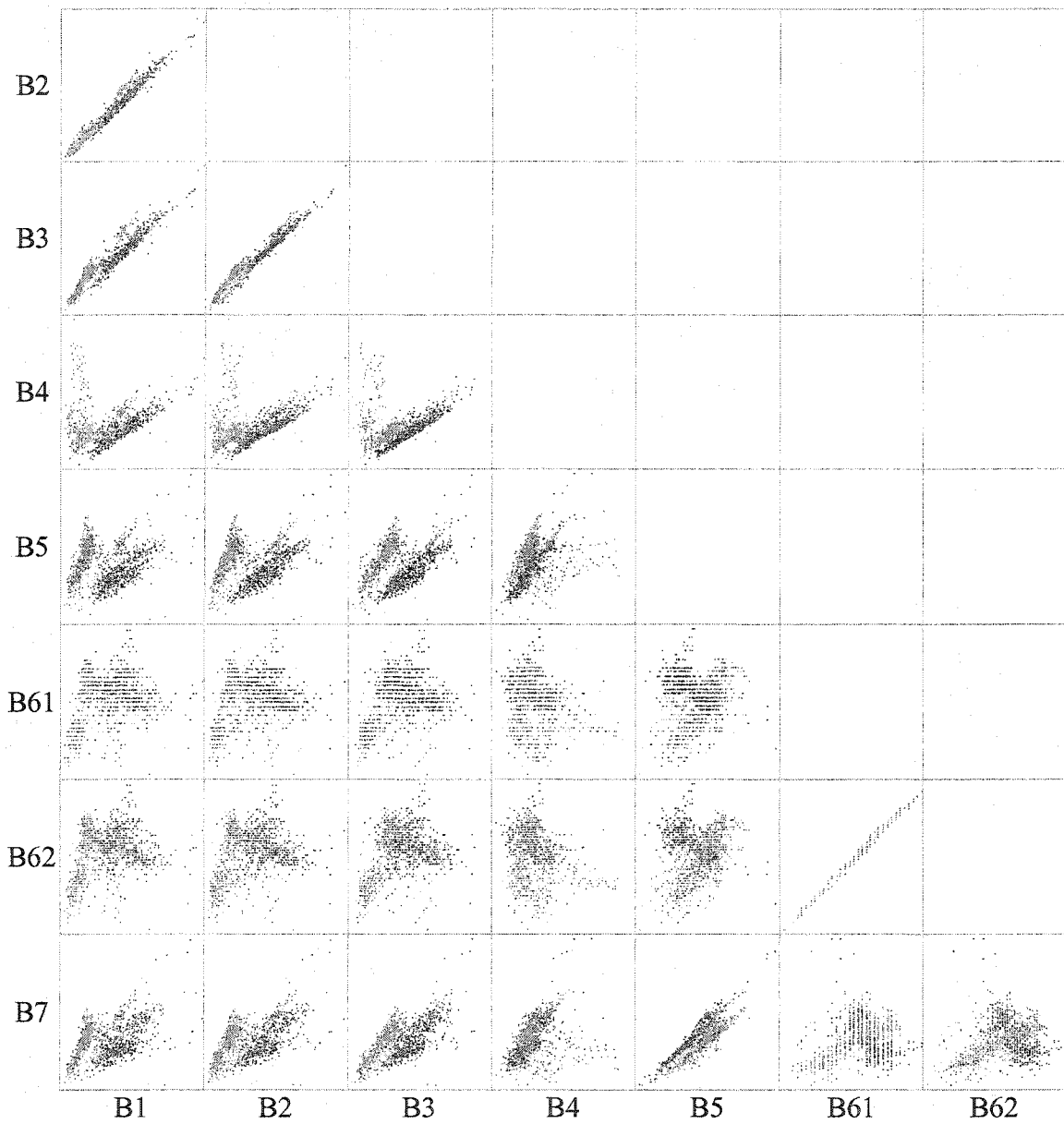
Figure 6.4 Two dimensional spectral space of distribution of ETM$^+$ training data

116

Table 6.4 Separability of each land use class in training data
(a) Level II classification system

	R	C	I	T	O
Residential					
Commercial	1.39				
Industrial	1.82	0.72			
Transportation	1.81	1.26	1.32		
Open	1.74	1.89	1.99	1.89	

(b) Level III classification system

	S	M	C	P	I	T	O
single							
multiple	1.59						
commercial	1.98	0.97					
public	1.77	0.71	0.77				
industrial	1.99	1.50	0.71	0.94			
transportation	1.99	1.54	1.21	1.49	1.32		
open	1.81	1.870	1.94	1.84	1.99	1.89	

117

family residential & public, commercial & industrial, commercial & public, public & industrial, and multiple-family residential & commercial showed low separability, less than 1.0. The separability of pairs of multiple-family residential & industrial, multiple-family residential & transportation, commercial & transportation, industrial & transportation, and public & transportation was also low ranged from 1.2 to 1.5. Conversely, pairs of single-family residential & industrial, single-family residential & transportation, industrial & open, single-family residential & commercial, and commercial & open exhibited high separability, which is more than 1.9. This result shows that single-family residential, and open land uses can be easily distinguished from other land uses. But the other land uses are not easily separable.

6.2.2 Bayesian Network Structures

We conducted urban land use classification using naïve Bayesian classifiers and MWSTs with the given data by selecting land use category as a class node. All data values used here were discretized to 15 values based on equal frequency interval, which was found to be optimal for Bayesian network classification in the previous section. The class node had 5 or 7 values corresponding to each land use category of interest.

Figure 6.5 shows the resulting Bayesian network structures. The structures revealed the most informative inputs to the class node. As shown in the figure, all input variables, bands 1 to 7, contributed to the value of the class node in naïve Bayesian classifiers, because the number of child nodes equaled the number of input variables. In

118
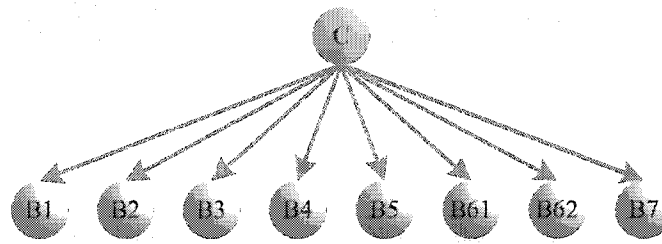
(a)
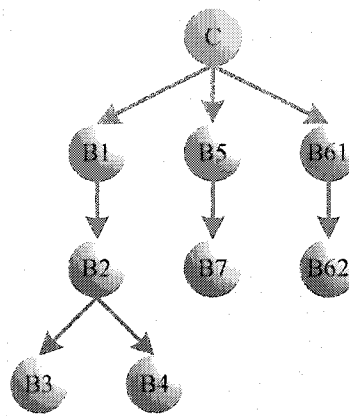
(b)

Figure 6.5 Bayesian network structure for urban land use classification
(a) naïve Bayesian classifiers (b) MWSTs

Table 6.5 Overall accuracy of Bayesian network classification

	Level II		Level III	
	accuracy	κ	accuracy	κ
Naïve	73%	63%	69%	60%
MWST	77%	67%	71%	63%

119

contrast, the structure of MWSTs showed that bands 1, 5 and 6 contributed to the value of the class node. MWSTs also provided strong dependency between visible bands (1, 2, and 3) and middle infrared bands (5 and 7), which is consistent to their high correlation.

6.2.3 Accuracy Assessment

For assessment of the performance of Bayesian network classification, the overall accuracies and κ coefficients for each case are given in Table 6.5. The MWSTs slightly outperform naïve Bayesian classifiers with overall accuracies of 77% and 71% (κ coefficients of 67% and 63%) for classification level II and III, respectively. The differences of overall accuracy between level II and III classification systems were less than 5% for both overall accuracy and κ coefficient.

Confusion matrices for level II and III classification systems are presented in Table 6.6. In level II classification, residential and open were fairly well predicted whereas commercial and industrial were not. For example, the omission and commission errors of residential and open land uses were all below 20% whereas those for commercial and industrial land uses were above 50%. In level III classification, single-family residential, transportation, and open were well assigned but multiple-family residential, commercial, public and light industrial were not. For example, commission errors of single-family residential, transportation and open land uses were all below 20% whereas those for commercial and public land uses were above 40%. The pixels

120

Table 6.6 Confusion matrix of urban land use classification
(a) Level II classification system using naïve Bayesian classifiers

	R	C	I	T	O	total	user
R	322	30	7	20	14	393	82%
C	31	61	22	17	8	139	44%
I	0	25	27	10	0	62	44%
T	8	19	4	64	3	98	65%
O	13	6	0	39	283	341	83%
total	374	141	60	150	308	1033	
prod	86%	43%	45%	43%	92%		

(b) Level II classification system using MWSTs

	R	C	I	T	O	total	user
R	348	13	5	15	12	393	89%
C	50	55	16	13	5	139	40%
I	1	29	22	8	2	62	36%
T	17	11	2	65	3	98	66%
O	14	1	0	23	303	341	89%
total	430	109	45	124	325	1033	
prod	81%	51%	49%	52%	93%		

(c) Level III classification system using naïve Bayesian classifiers

	S	M	C	P	I	T	O	total	user
S	252	6	1	4	0	0	10	273	92%
M	15	73	8	5	1	11	7	120	61%
C	1	10	41	6	10	1	3	72	57%
P	7	11	9	23	9	5	3	67	34%
I	0	0	17	2	40	3	0	62	65%
T	1	10	1	2	0	74	10	98	76%
O	2	3	5	4	0	10	317	341	93%
total	278	113	82	46	60	104	350	1033	
producer	91%	65%	50%	50%	67%	71%	91%		

(d) Level III classification system using MWSTs

	S	M	C	P	I	T	O	total	user
S	253	6	0	1	0	0	13	273	93%
M	16	79	9	2	1	4	9	120	66%
C	8	13	35	4	9	1	2	72	49%
P	13	6	8	26	9	3	2	67	39%
I	0	2	10	2	47	1	0	62	76%
T	4	5	1	3	0	78	7	98	80%
O	5	7	3	1	0	5	320	341	94%
total	299	118	66	39	66	92	353	1033	
producer	85%	67%	53%	67%	71%	85%	91%		

121

originally belonging to commercial and public land uses were often mislabeled with each other, multiple-family residential, and industrial land uses.

6.2.4 Urban Land Use Thematic Maps

The resulting urban land use thematic maps shown in Figure 6.6. For level II classification, both networks well captured residential, transportation and open land uses. The other classes visually agreed with their real land use although their prediction accuracy was not high enough, as shown in Table 6.6. For level III classification, both networks also well assigned pixels belonging to single-family residential, transportation and open land uses. However, public land uses were not well predicted especially for those pixels in Loyola Marymount University. Some pixels inside Los Angeles International Airport were misclassified as commercial land use and many pixels in Dockweiler state beach as transportation land uses. Those pixels in under construction areas in Playa del Rey exhibited transportation land use due to the construction. For both classification system, MWSTs appeared to be better than naïve Bayesian classifiers.

6.3 Incorporating Ancillary Data

In order to improve classification accuracy, several ancillary data, in addition to spectral data, were incorporated. We investigated the effect of locational data, elevational data, and image transformation data such as NDVI and Tasseled Cap transformation.
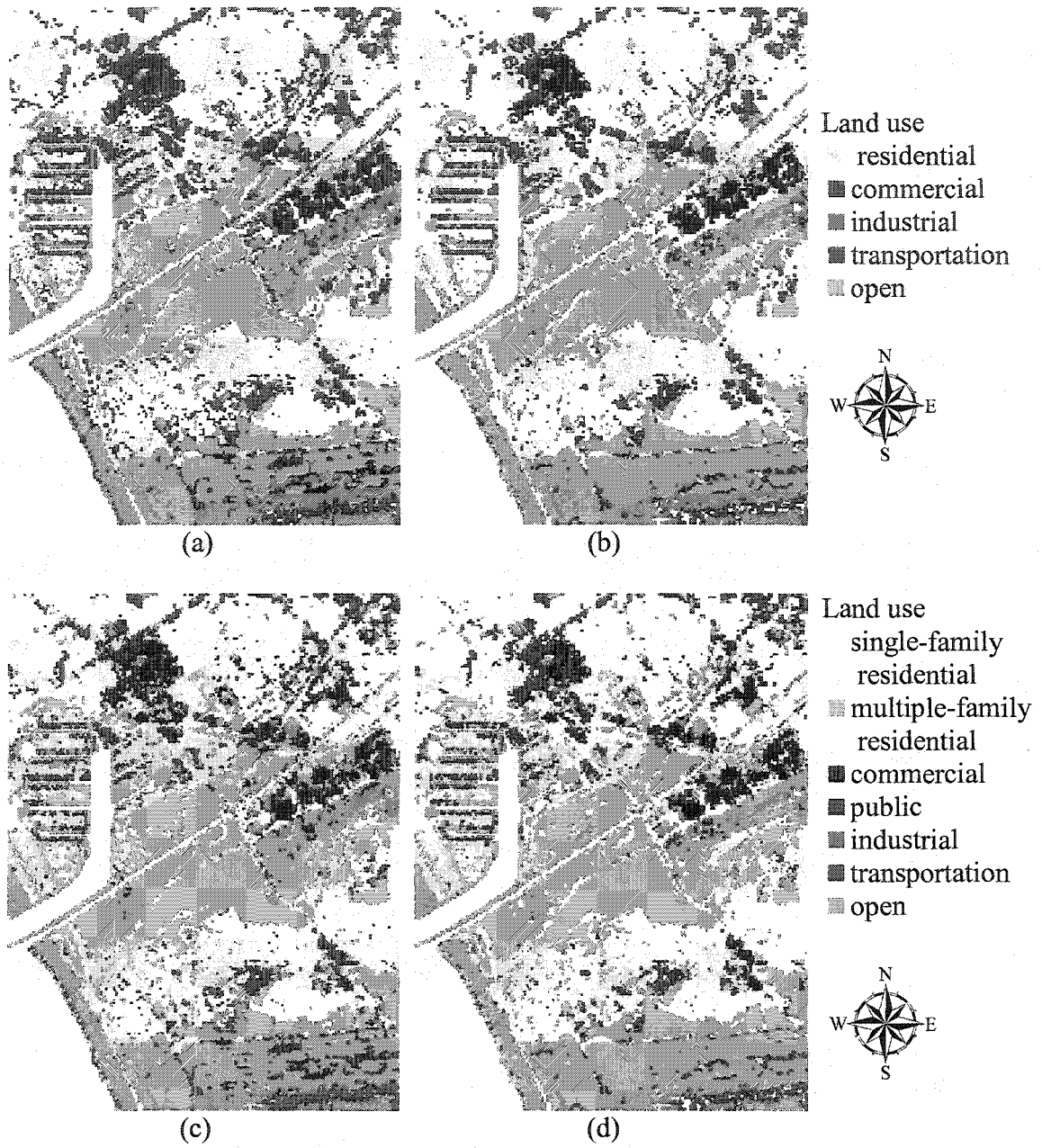
122

Figure 6.6 Urban land use maps using (a) naïve Bayesian classifiers for level II classification system (b) MWSTs for level II classification system (c) naïve Bayesian classifiers for level III classification system (d) MWSTs for level III classification system

123

6.3.1 Locational Data

Figure 6.7 shows the Bayesian network structure when incorporating ancillary data. Incorporating locational ancillary data improved urban land use classification accuracy. Figure 6.8 shows that the average improvements of overall accuracies were up to 7% and 10% for classification level II and III, respectively. Even the differences for the same network between classification level II and III became smaller when incorporating locational ancillary data. For example, the difference of overall accuracies when using spectral data only was 5% on the average whereas the difference when including locational ancillary data was 2%. The κ coefficients were also improved by 9% and 14% for classification level II and III, respectively for both network structures. In the same way, the differences between level II and III classification for the same network become smaller. For example, the range of different κ coefficients when using only spectral data was up to 4% whereas the range when including locational ancillary data was up to 2%.

Incorporating locational ancillary data also reduced the omission and commission errors for each land use. For example, omission errors of multiple-family residential, commercial, public, industrial and transportation land uses were reduced by 12% to 35% and commission errors of the land uses were reduced by 9% to 33%. In this case, MWSTs slightly outperformed naïve Bayesian classifiers when incorporating locational ancillary data.
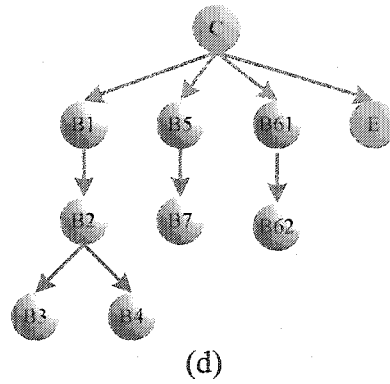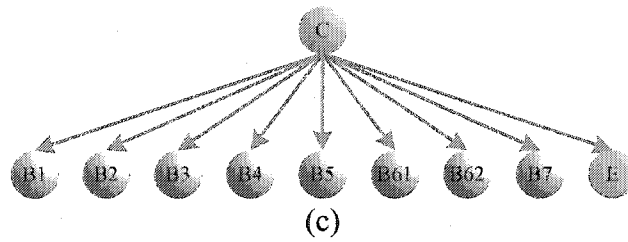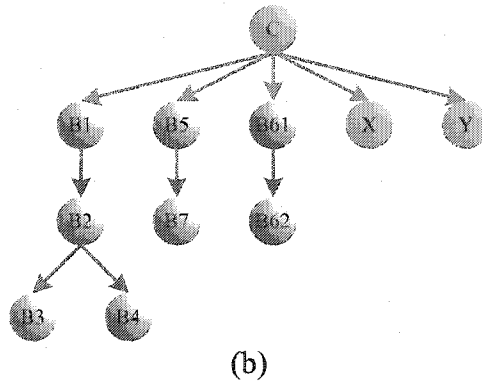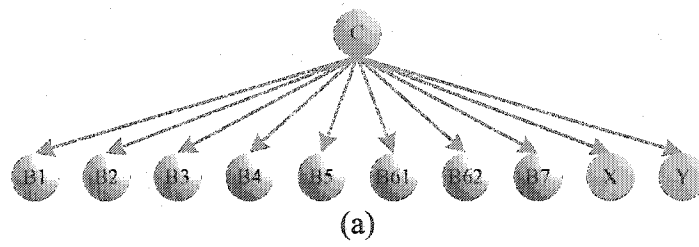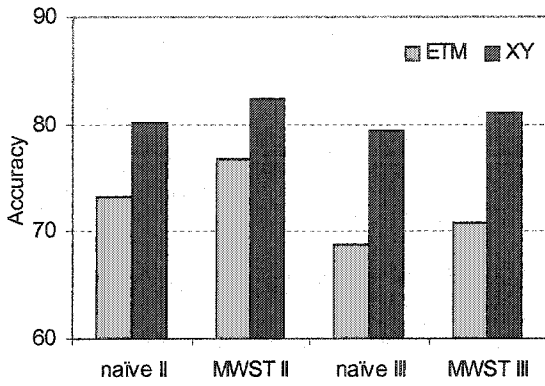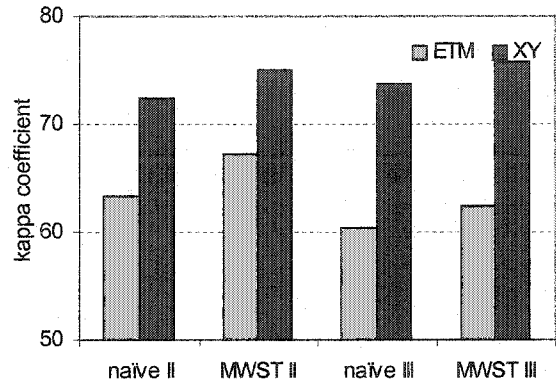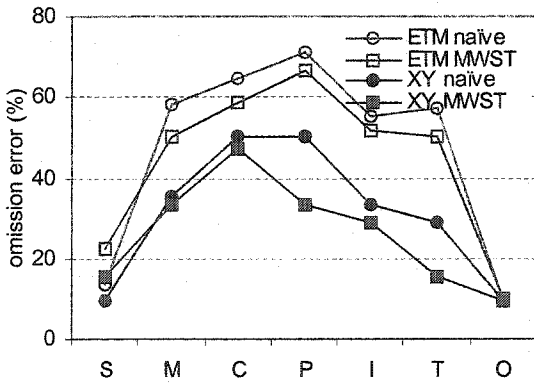
124

Figure 6.7 Bayesian network structure with ancillary data (a) naïve Bayesian classifiers with locational data (b) MWSTs with locational data (c) naïve Bayesian classifiers with elevation data (d) MWSTs with elevation data
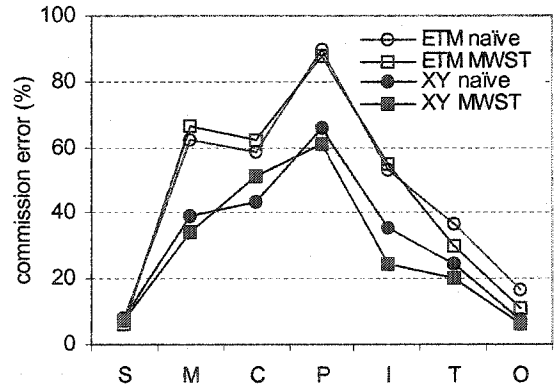
125

Figure 6.8 Level III classification accuracy (a) overall accuracy (b) kappa coefficient
(c) omission error (d) commission error

126

The land use map in Figure 6.9 shows improved urban land use maps when incorporating locational ancillary data. Compared to the land use map with spectral data only, those pixels in Los Angeles International Airport were correctly assigned as transportation land use and those in Dockweiler state beach were as open land use. Moreover, the pixels in Loyola Marymount University were predicted as public land use. On the whole, incorporating locational ancillary data achieved better classification.

6.3.2 DEM

Incorporating elevational ancillary data also improved the classification accuracy. For example, incorporating SRTM and NED data improved overall accuracies by up to 3% and 6% and κ coefficients by up to 5% and 7%, respectively for level III classification system. Incorporating elevational ancillary data also improved individual class prediction. For level III system, omission errors were reduced by up to 17% and 26%, and commission errors were reduced by up to 19% and 22% by incorporating SRTM and NED, respectively. Especially public, industrial, and transportation land uses were better assigned. Interestingly, the commission errors of single-family residential and commercial land uses incorporating DEM were not lower than those using spectral data only.

Figure 6.10 shows the comparison of accuracies using different geospatial ancillary data. This figure shows that all accuracies were improved by using ancillary data. However, the improvement by including elevational data was not as significant as

127
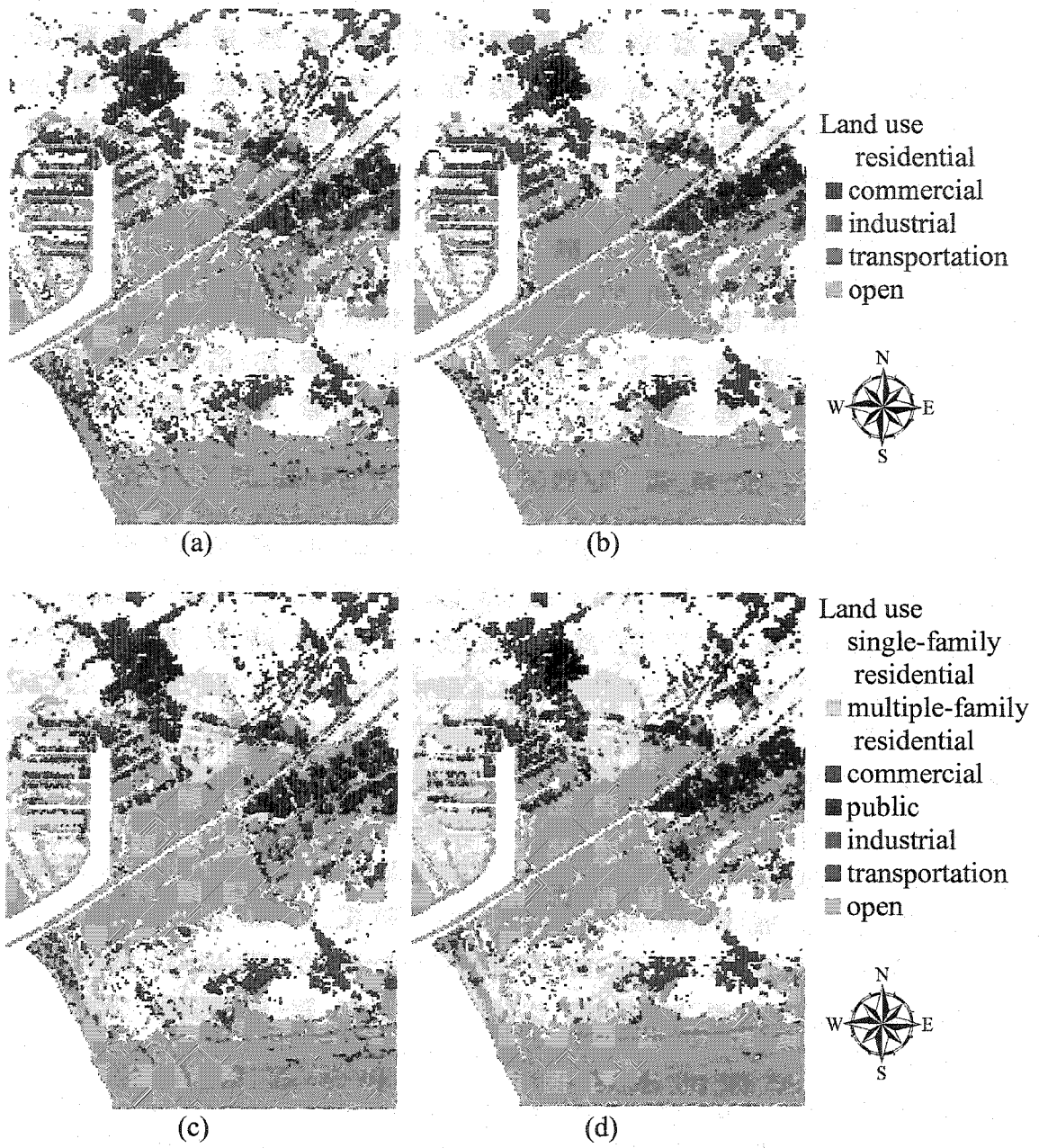
Figure 6.9 Urban land use maps incorporating locational data using (a) naïve Bayesian classifiers for level II classification system (b) MWSTs for level II classification system (c) naïve Bayesian classifiers for level III classification system (d) MWSTs for level III classification system

128
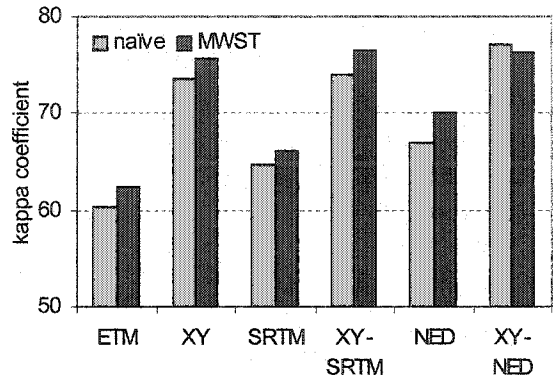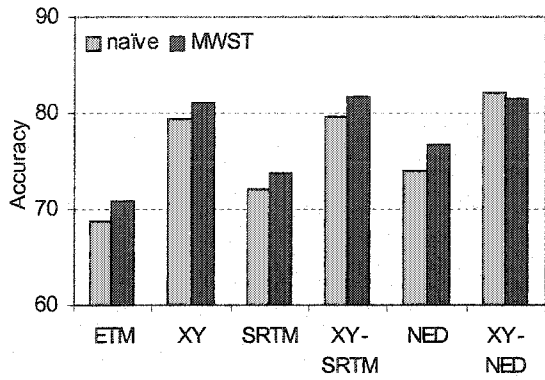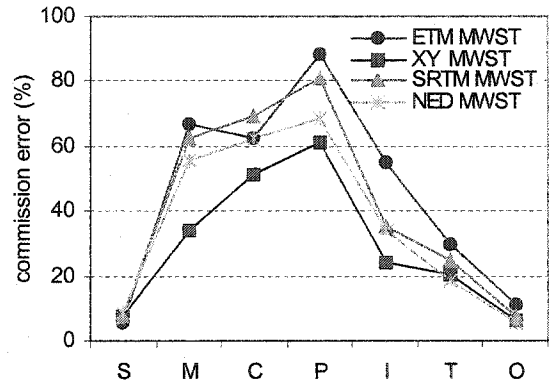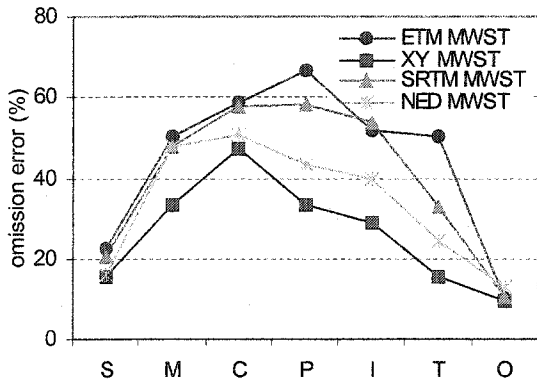
Figure 6.10 Level III classification accuracy incorporating geospatial ancillary data
(a) overall accuracy (b) kappa coefficient (c) omission error (d) commission
error

129

incorporating locational data. Moreover, omission and commission errors were more reduced by incorporating locational data. Using locational ancillary data was even better than or using both locational and elevation data for many cases.

6.3.3 Vegetation Index

Apart from spatial and elevation ancillary data, we also investigated the effect of incorporating other ancillary data obtained from image transformation such as NDVI and brightness, greenness and wetness from Tasseled Cap. The result shows that these ancillary data did not significantly improve accuracies compared to using spectral data only or incorporating locational ancillary data. For example, incorporating NDVI improved the overall accuracy of 1% and κ coefficient of 2% for level III classification. It reduced omission errors of single-family residential, public and transportation land uses by 8% to 17% and commission errors of multiple-family residential, commercial, public, and industrial land uses by 2% to 13%.

Incorporating Tasseled Cap transformation did not significantly improves classification accuracy. The overall accuracy was improved only up to 2% and κ coefficient was reduced by up to 6% to 14%. Therefore, additional information from the image transformation was not useful as other ancillary data. Figure 6.11 shows the comparison of accuracies incorporating different vegetation indices.
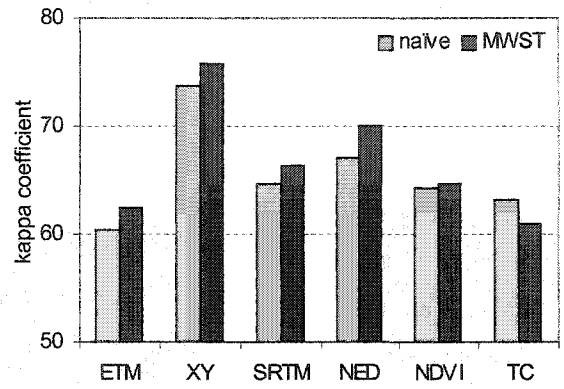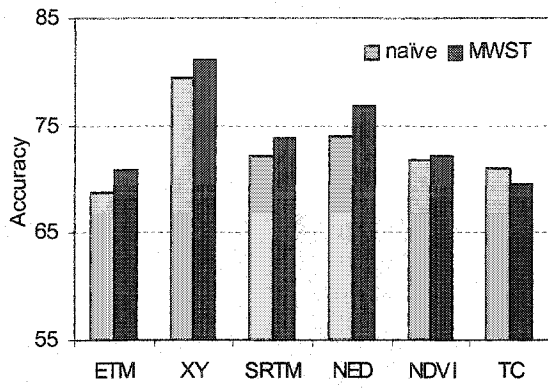
130
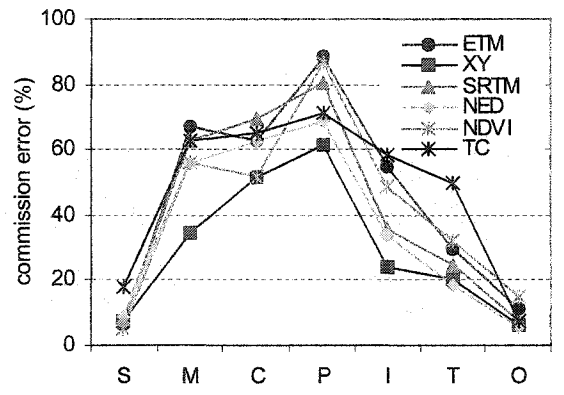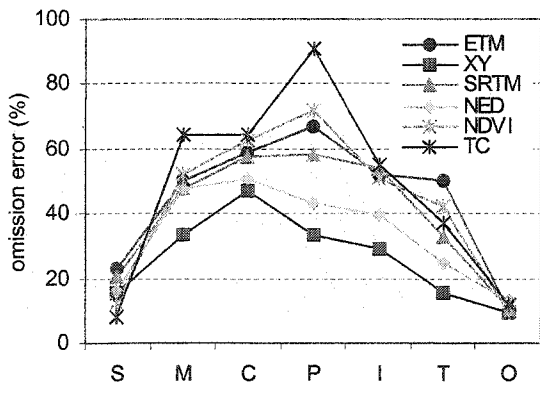
Figure 6.11 Level III classification accuracy incorporating vegetation indices (a) overall accuracy (b) kappa coefficient (b) omission error (c) commission error

131

6.4 Discussion

Using satellite imagery for land use classification is an effective alternative to the conventional ground surveyed land use. Estimating land use from spectral signatures of satellite imagery reduces time to collect information. Moreover potential errors in ground surveyed land use can be avoided since ground surveyed land use that are not developed for environmental purposes often group environmentally different land uses into the same category. For example, the vegetation areas inside the Los Angeles International Airport were assigned to open land use instead of transportation as in SCAG land use. Vegetated areas and parking lots in Loyola Marymount University were assigned to open and transportation land uses respectively, which both belonged to public in the SCAG land use data. The areas under construction in Playa Vista showed transportation and industrial land uses depending on the degree of construction.

Bayesian networks are useful in urban land use classification from satellite imagery. The networks can be trained with small size of training data. Bayesian networks can be successfully applied to a broad range of datasets, which may have been collected without considering data distribution.

Both naïve Bayesian classifiers and the networks based on MWST algorithm were useful, but MWSTs slightly outperformed naïve Bayesian classifiers. Naïve Bayesian classifiers have all the inputs connected to the class node under conditional independence assumption among the input nodes. In other words, the input nodes are not dependent on one another given the value of the class node and the value of the class node is affected

132

by all the input nodes. This assumption could be violated since visible bands 1, 2 and 3 and middle infrared bands 5 and 7 are highly correlated. Therefore, this might inhibit the performance of naïve Bayesian classifiers. In addition, naïve Bayesian classifiers do not present information of which input nodes more contribute to prediction of the class node value. Conversely, the structure of MWSTs explicitly showed the dependence relationships among variables. The structure of MWSTs reveal that bands 1, 5 and 6 are the most contributing input to the class node value. The mutual information of band 1 and class node was the largest which, shows that band 1 is critical in the urban land use classification using Landsat TM data. This is consistent with the existing finding that band 1 is important in urban features (Jensen, 2000). MWSTs also exhibit strong dependency between visible bands (1, 2, and 3) and middle infrared bands (5 and 7), which is consistent with their high correlation. Compared with naïve Bayesian classifiers and other existing AI techniques, MWSTs are an effective structure to understand the relationships among input variables for classification and identify the most important input variables for determining target node value.

In order to validate the accuracy improvement, the overall accuracies of Bayesian networks were compared with the accuracy by random classification. For example, classification with two states can provide 50 % of accuracy even with random prediction. The comparison between these two accuracies is given in Table 6.7 Bayesian networks provide approximately 80% of accuracy whereas the random classification accuracies are 20% and 14% for level II and III classification system. Therefore, the classification accuracy provided by Bayesian networks was at least four times better than random

133

Table 6.7 Bayesian network accuracy compared with random classification accuracy

	Level II		Level III	
	accuracy	ratio	accuracy	ratio
Naïve	80	4.0	79	5.6
MWST	82	4.1	81	5.8
Random	20	1.0	14	1.0

classification. This demonstrates that Bayesian network classification of urban land use classification from Landsat ETM$^+$ image was successful and effective even for level III classification system.

Incorporating ancillary data, especially locational information, is important in predicting the class node value. Incorporating locational ancillary data achieved better classification for both network structures. This is a logical conclusion since the same types of land uses tend to be located together and incorporating locational ancillary data help lumping the surrounding pixels. This is confirmed from the thematic maps. Those poorly assigned pixels using spectral data only were corrected when incorporating locational ancillary data. In fact, incorporating locational data was better than including elevational data or combination of both. Incorporating locational ancillary data was favorable and practical since they can be simply obtained from the given Landsat ETM$^+$ Image. However, this might be case specific since our study area is rather flat so the effect of the elevation might be insignificant.

Incorporating other ancillary data from image transformation such as vegetation indices derived from the image did not significantly improve classification accuracies. This is logical conclusion since these data were calculated from image DN values, which were correlated to the spectral data. Correlated data are not helpful in learning of Bayesian networks since the ancillary data are more related to the spectral data not to the class node.

135

The inaccuracies might result from the fact that urban features are difficult to distinguish spectrally using Landsat ETM$^+$ image. The results of training data distribution and separability showed the spectral signatures from some land uses were similar to one another. Most land uses were hard to separate except single-family residential and open land uses. Some pairs exhibited low separability less than 1, which might result from the similar spectral signature from the similar roof materials. This is often reported as a problem of urban land use classification due to the heterogeneity and small size of the target materials (Ridd, 1995; Foody, 2000; Stefanov, 2001) and different urban land uses often have similar characteristics. This problem becomes worse when more land use categories are classified (Stefanov, 2001). These facts might lower the overall accuracy from the given resolution of the data.

The land use classification of TM image referenced only to SCAG data is particularly problematic. The potential sources of error in the classification might stem from SCAG land use data. The SCAG data set is typical of available land use data, but is collected for tax purposes such that some of their land uses are not necessarily representative environmental land uses. Furthermore the land use data defined in SCAG data have been lumped into neighborhoods so they can contain varying degrees of mixed land use information, which cannot be distinguished by spectral information. Another explanation is the difference in timing in that SCAG data and Landsat TM data were collected three years apart. The development and changes that occurred during the three years are undefined.

136

As a result, the Bayesian network classification gives reasonable accuracy. For some cases, the accuracy of classification measured from spectral signatures can be more accurate than shown in many cases.

Chapter VII

Stormwater Pollution Estimates

Monitoring stormwater runoff and predicting pollutant loads to receiving waters are not easy tasks since these demand expertise and accumulated data (Duncan, 1995; Vaze and Chiew, 2003), and the event specific data, i.e. EMCs of stormwater pollutants, and site specific data measurements, i.e. runoff coefficient and rainfall, are not always available (Vaze and Chiew, 2003). Alternative approach has been developed using land use and rainfall data (Stenstrom *et al.*, 1984; Chiew and McMahon, 1999). However, few attempts have been made to spatially estimate the pollutant loads in an inexpensive and consistent way. This chapter proposes the use of satellite imagery to estimate stormwater pollutant loads in a typical urban area.

In this chapter, we generate spatial estimate of pollutant loads from stormwater runoff in our study area. In section 7.1, we use land use that was classified from satellite

138

images using Bayesian networks and convert the land use data to related runoff coefficients associated with imperviousness, EMCs of a particular pollutant, and corresponding stormwater pollutant loads. In section 7.2, we predict stormwater pollutant loads directly from satellite imagery. For both approaches, we examined ten water quality parameters: chemical oxygen demand (COD), biochemical oxygen demand (BOD_5), total suspended solids (TSS), nutrients such as total Kjeldahl nitrogen (TKN), nitrite and nitrate ($NO_{2\&3}$) and total phosphorus (TP), heavy metals such as total copper (Cu), total lead (Pb), total zinc (Zn), and oil and grease (O&G). We identify the areas that generate high pollutant loadings into receiving waters. In section 7.3, we propose new classification system optimized for stormwater management purpose replacing conventional USGS land cover/use classification system. In section 7.4, we discuss different approaches to estimate stormwater pollution load and compare the pollutant loading areas with SCAG data. We also discuss the benefits of our approach in stormwater monitoring.

7.1 Pollutant Loading Estimate from Land use Classification

Land use classified from a Landsat ETM^+ image based on USGS Land Cover/Use level III classification system in the previous chapter was converted to runoff coefficients, EMCs and pollutant loading maps. These land uses were selected since the limited number of land use types can be used for more reliable stormwater quality information (Burian *et al.*, 2002).

139

Figure 7.1 shows of impervious surface maps associated with runoff coefficients in the study area. Impervious surfaces were classified into four states. The areas of the lowest runoff coefficient with the value of 0.1 corresponded to open land use. The areas of the next lowest runoff coefficient with the value of 0.4 corresponded to single-family residential land use. The areas with runoff coefficient of 0.6 corresponded to multiple-family residential land use. The areas of the highest runoff coefficient with the value of approximately 0.7 corresponded to all the other land uses such as commercial, public, industrial and transportation.

Figure 7.2 shows of EMCs maps for each pollutant in the study area. We should stress that the EMCs in the Santa Monica Bay watershed tend to be higher than those reported in US EPA's Nationwide Urban Runoff Program (NURP) database (Driscoll *et al.*, 1990; Wong *et al.*, 1997). EMCs were generally classified into three states but COD, Zn, and O&G maps displayed only two states. Residential and transportation land uses tended to increase the COD EMC, while the rest of land uses tended to lower the EMC. All land uses but single-family residential and open land uses increased BOD_5 EMC. Open land use tended to increase the TSS EMC while the other land uses except single-family residential lowered the TSS EMC. Single-family residential land use was the most significant contributor of nutrients. The EMC maps of Cu and Pb were identical in that residential and transportation land uses were responsible for high EMC and open land use corresponded to low EMC areas. The EMC map of Zn exhibited different behavior than other water quality parameters; commercial, public and industrial land uses showed high EMC and residential, transportation and open land uses showed low EMC. The high

140

Figure 7.1 Impervious surface map associated with runoff coefficients converted from land use classification
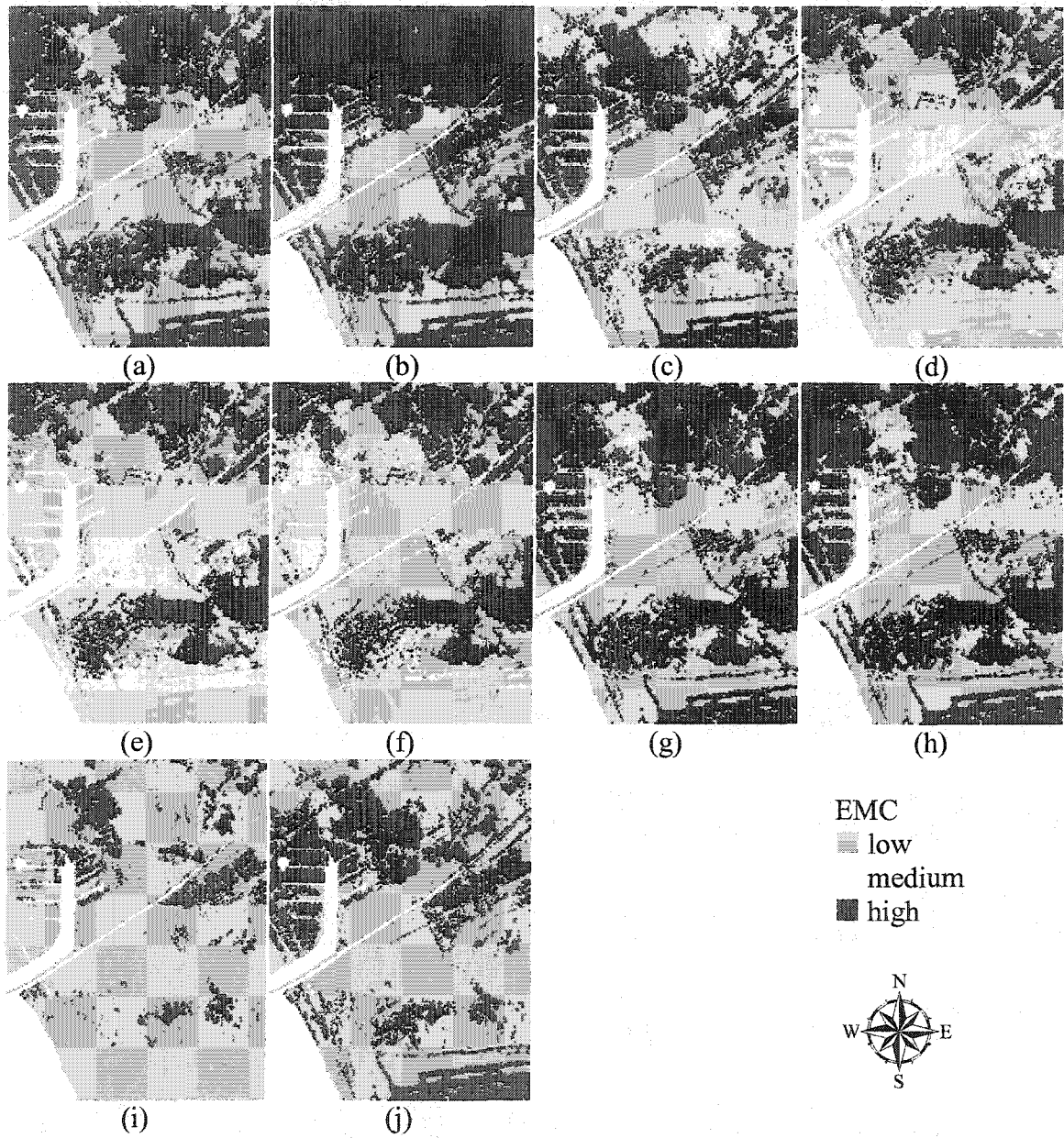
Figure 7.2 EMC maps converted from land use classification (a) COD (b) BOD$_5$ (c) TSS (d) TKN (e) NO$_{2\&3}$ (f) TP (g) Cu (h) Pb (i) Zn (j) oil and grease

142

EMC of O&G is mainly caused by multiple-residential, commercial, public, industrial, and transportation land uses and contribution of the other two land uses was trivial.

Figure 7.3 shows stormwater pollutant loading maps for each pollutant type, in which the pollutant load is proportional to the product of runoff coefficient and EMC. Pollutant loads were also classified into three states but TSS, TKN, TP and O&G maps had only two states such that the medium state was not represented in the final result. High COD emissions were generated from multiple-family residential, commercial, industrial, and transportation land uses. Single-family residential and public land uses corresponded to medium load and open land use to low load. High BOD_5 emission areas were similar to high COD emission areas in addition to public land use. The pollutant loading maps of TSS, TKN and TP were identical because all land uses except open area significantly contributed to high load. High $NO_{2\&3}$ loads came from single-family residential, commercial, public, industrial, and transportation land uses. The Cu map shows that single-family residential land use corresponded to medium load, open land use to low load, and all the other land uses to high load. For the Pb, the high load mainly came from multiple-family residential and transportation land uses. Open land use also generated low load and the other four land uses created medium load. Zn map shows low loads from open land use, medium loads from residential and transportation land uses, and high loads from commercial, public and industrial land uses. Multiple-family residential, commercial, public, industrial, and transportation land uses created high O&G loads whereas single-family residential and open land uses produced low loads. O&G loading map was identical to its EMC map, which was not true for the other pollutants.
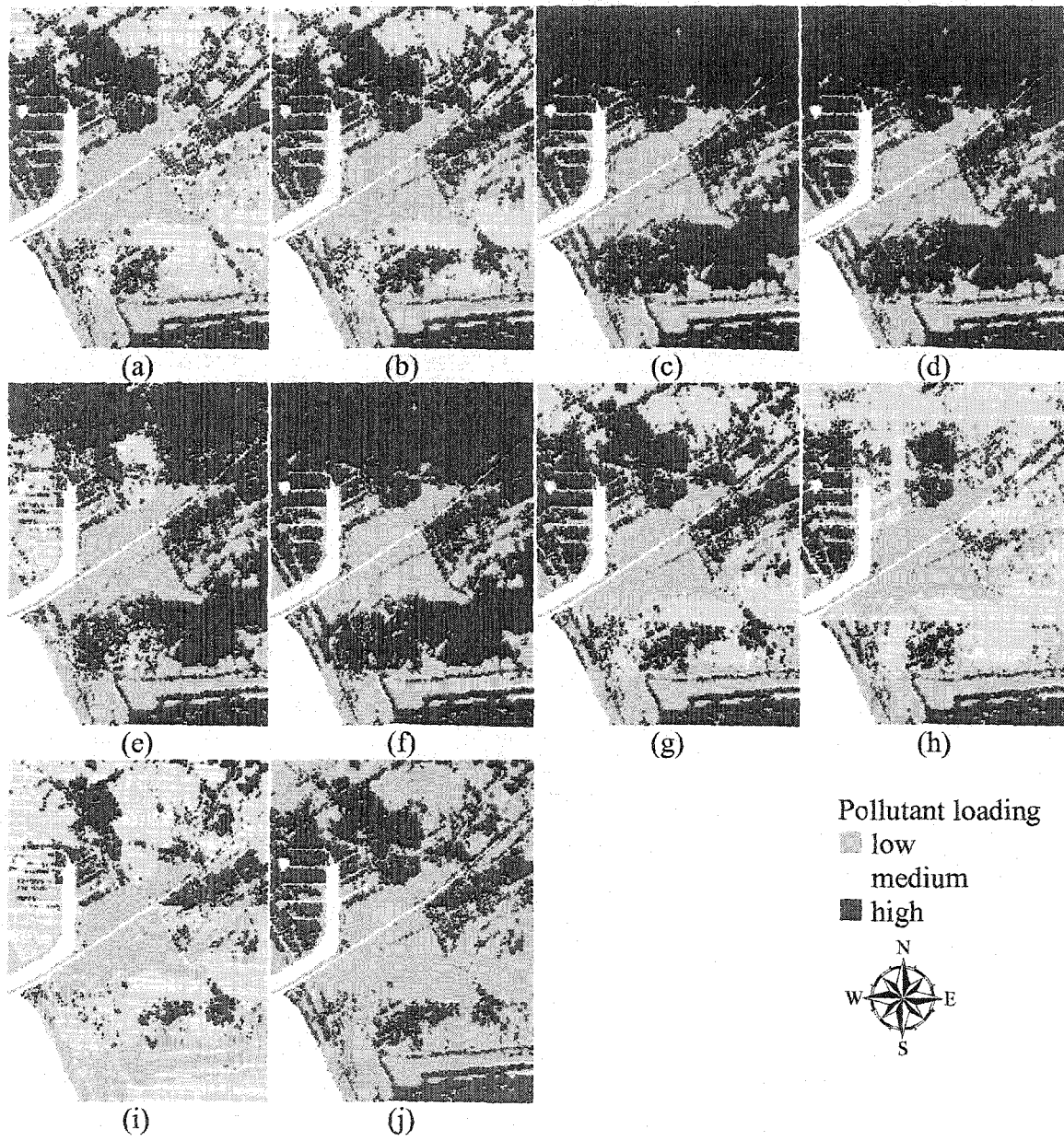
143

Figure 7.3 Pollutant loading maps converted from land use classification (a) COD (b) BOD$_5$ (c) TSS (d) TKN (e) NO$_{2\&3}$ (f) TP (g) Cu (h) Pb (i) Zn (j) oil and grease

144

Figure 7.4 compares the pollutant loads and EMCs as a percentage of total area. The percentage of areas corresponding to the high, medium and low categories are compared. The result shows that the pollutant loads did not necessarily correspond to the EMCs. The differences were due to the varying imperviousness or runoff coefficient. Over 65% of the area generated high pollutant loads for TSS, TKN, and TP. Only 15% of the area generated high Zn loads, and only 22% of the area generated high Pb load. Open areas tended to reduce pollutant loads, and 33% of the areas were classified as low due to the contribution of the open area, for all pollutants except O&G.

7.2 Pollutant Loading Estimate from Satellite Imagery

7.2.1 Pollutant Loading with Low, Medium and High Scheme

As an alternative approach, pollutant loads were estimated from satellite imagery using Bayesian networks. MWSTs were constructed from data using mutual information with the pollutant loads as the class node. For each pollutant, unit pollutant loads were calculated from runoff coefficients and EMCs based on equation (4.8) developed for the Santa Monica Bay watershed. The class node had 3 states, which corresponded to low, medium, and high loads. The network structures in Figure 7.5 shows that bands 1, 5, and 6 mainly contributed to the class node values for TSS, TKN, TP, $NO_{2\&3}$, Pb, and Zn whereas Bands 4 also contributed to class node for COD, BOD_5, Cu and O&G.
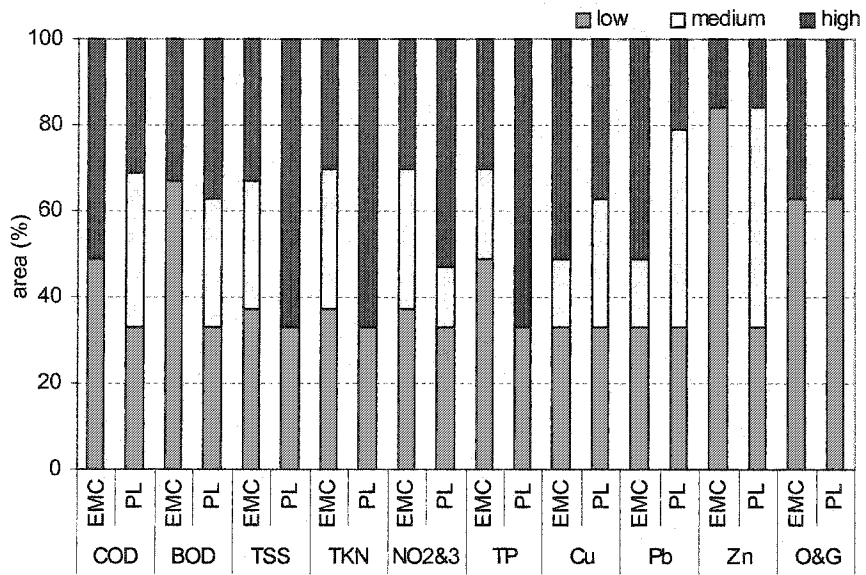
145

Figure 7.4 Proportion of each water quality parameter state (a) EMC (b) pollutant loading
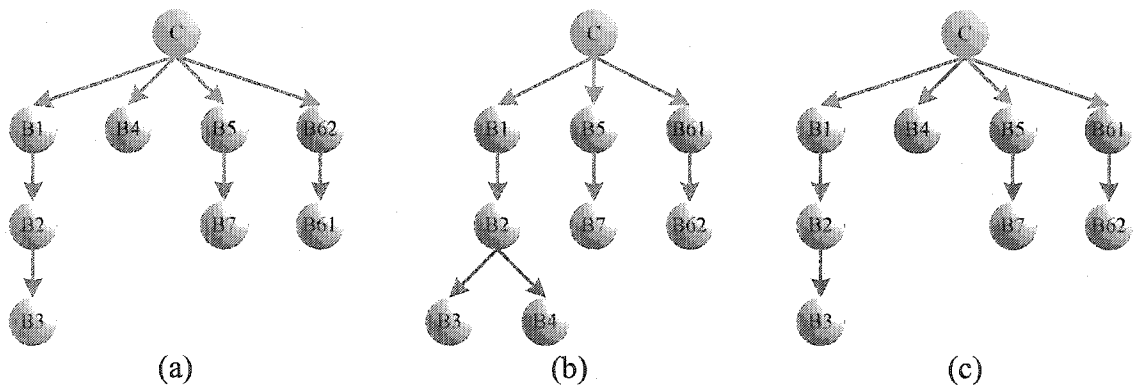

(a)　　　　　　　　　(b)　　　　　　　　　(c)

Figure 7.5 Bayesian network structures for loading of (a) COD (b) TSS, TKN, $NO_{2\&3}$, TP, Pb, and Zn (c) BOD_5, Cu, and oil and grease

146

Figure 7.6 shows each stormwater pollutant loading maps. The figure also shows that each pollutant shows different pattern. For example, TSS, TKN, TP, and O&G were classified into two classes whereas the other maps were classified into three classes. Each map also shows different high pollutant loading areas although transportation areas corresponded to high pollutant loading areas for most water quality parameters except Zn. Multi-family residential, commercial, public, and industrial land uses often contributed to high pollutant loads for most pollutants although it could be case specific. For all cases, open land use corresponded to low pollutant loading area.

The overall accuracies and κ coefficients of classification were varied as shown in Figure 7.7. Overall accuracies ranged from 80% to 94% and κ coefficients from 66% to 86%. Classification of BOD_5, TSS, TKN, TP, Cu and O&G loads was highly accurate i.e. more than 90% of overall accuracy and 80% of κ. The omission errors of high pollutant loads ranged from 6% to 36%. Especially omission errors of BOD5, TSS, TKN, TP, Cu and O&G loads were much lower than that of other pollutants, i.e. less than 11% errors. The commission errors of low pollutant loads were not so varied for each pollutant, ranged from 7% to 13%.

The accuracies were improved by incorporating ancillary data for $NO_{2\&3}$, Pb and Zn by up to 6% and 11% for overall accuracy and κ coefficient respectively. For other pollutants, the improvement was trivial. Accuracy improvement by incorporating locational ancillary data was better than incorporating elevational data or including both in most cases. Moreover, incorporating locational data reduces the omission errors of high pollutant loading areas especially for $NO_{2\&3}$, Pb, and Zn by up to 13% and the
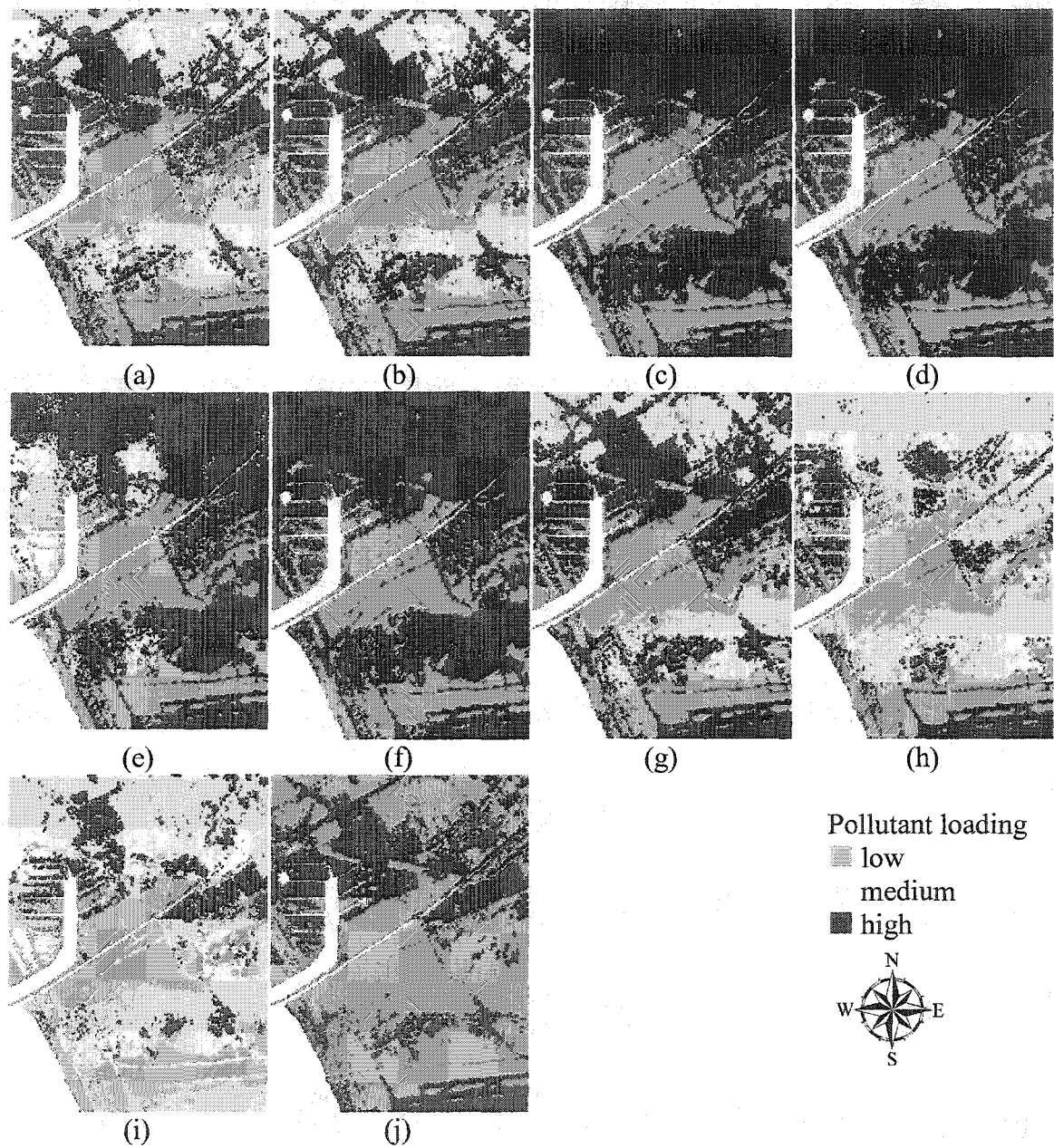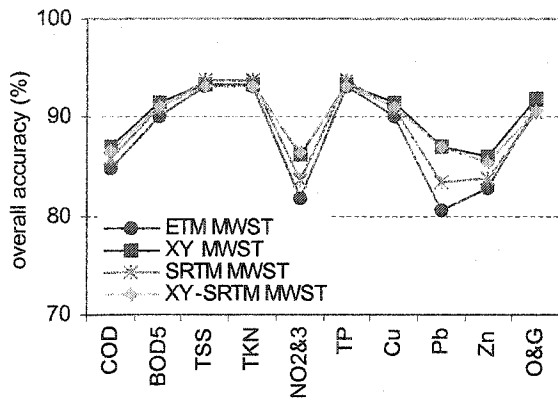
147

Figure 7.6 Pollutant loading maps using MWSTs with spectral data and geospatial data (a) COD (b) BOD$_5$ (c) TSS (d) TKN (e) NO$_{2\&3}$ (f) TP (g) Cu (h) Pb (i) Zn (j) oil and grease

148
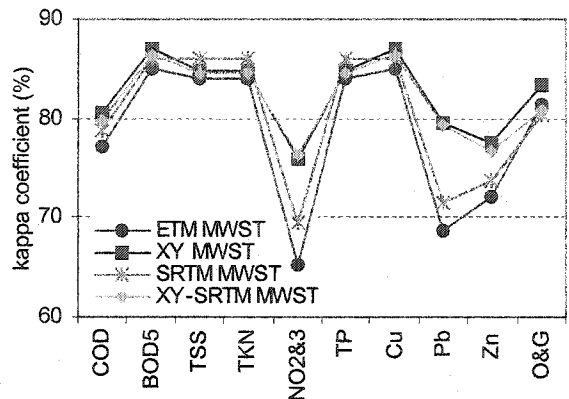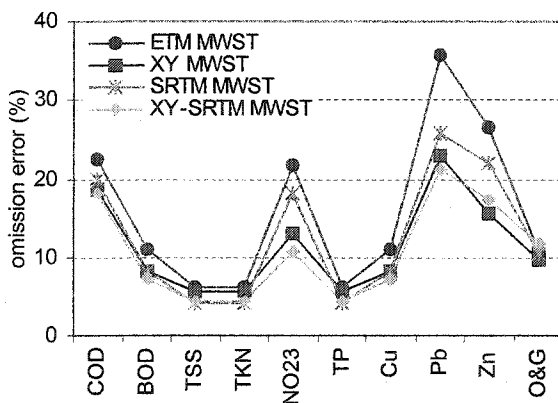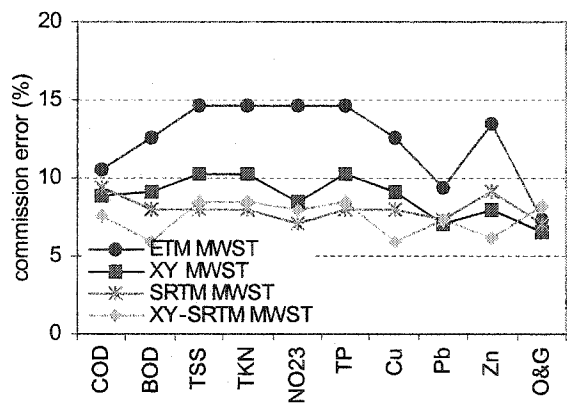
Figure 7.7 Accuracies of pollutant loading estimates from satellite imagery (a) overall accuracy (b) kappa coefficient (c) omission errors (d) commission errors

149

commission errors of low pollutant loading areas especially for BOD_5, $NO_{2\&3}$, Cu and Zn by 5%. Therefore, incorporating locational ancillary data was also more useful and practical for pollutant loading classification in our study area.

7.2.2 Normalized Pollutant Loading Scheme

We classified normalized pollutant loads in order to examine which pollutants have more severe impact on receiving water. The class node had seven states corresponding to the degree of pollutant loads: very low, low, medium-low, medium, medium-high, high, and very high. In this case, 'very low' state indicates that no pollutant load is discharged. The state of 'low' is the minimum pollutant load except zero loading. The rest of the states were normalized based on 'low' state of unit pollutant loading values as shown in Table 7.1. The resulting Bayesian network structures are shown in Figure 7.8. For all cases, bands 1, 5 and 6 mainly contributed to the class node values. In addition, band 4 was also connected to the class node in the structure for Cu, Zn, and O&G.

Figure 7.9 shows the normalized pollutant loading maps. Each pollutant shows a different level of classification: BOD_5, TSS, TKN, and TP were classified into two classes; COD, $NO_{2\&3}$, Cu, and Pb were classified into three classes; and Zn and O&G were classified into four classes. Only O&G loading map displayed the level of 'very low' and only BOD_5 and Pb loading maps displayed 'very high' level. The normalized pollutant loading maps show the lowest pollutant loading areas mostly correspond to

150

Table 7.1 Classification states for water quality parameter

State	Normalized loading values
very low	0
low	1
medium low	≤ 4
medium	≤ 8
medium high	≤ 12
high	≤ 16
very high	> 16

(a)   (b)
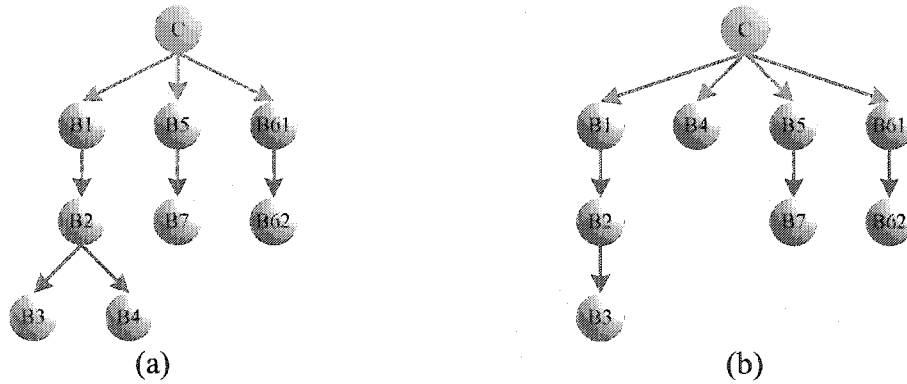
Figure 7.8 Bayesian network structures for normalized pollutant loading of (a) COD, BOD_5, TSS, TKN, $NO_{2\&3}$, TP, and Pb (b) Cu, Zn, and oil and grease

151
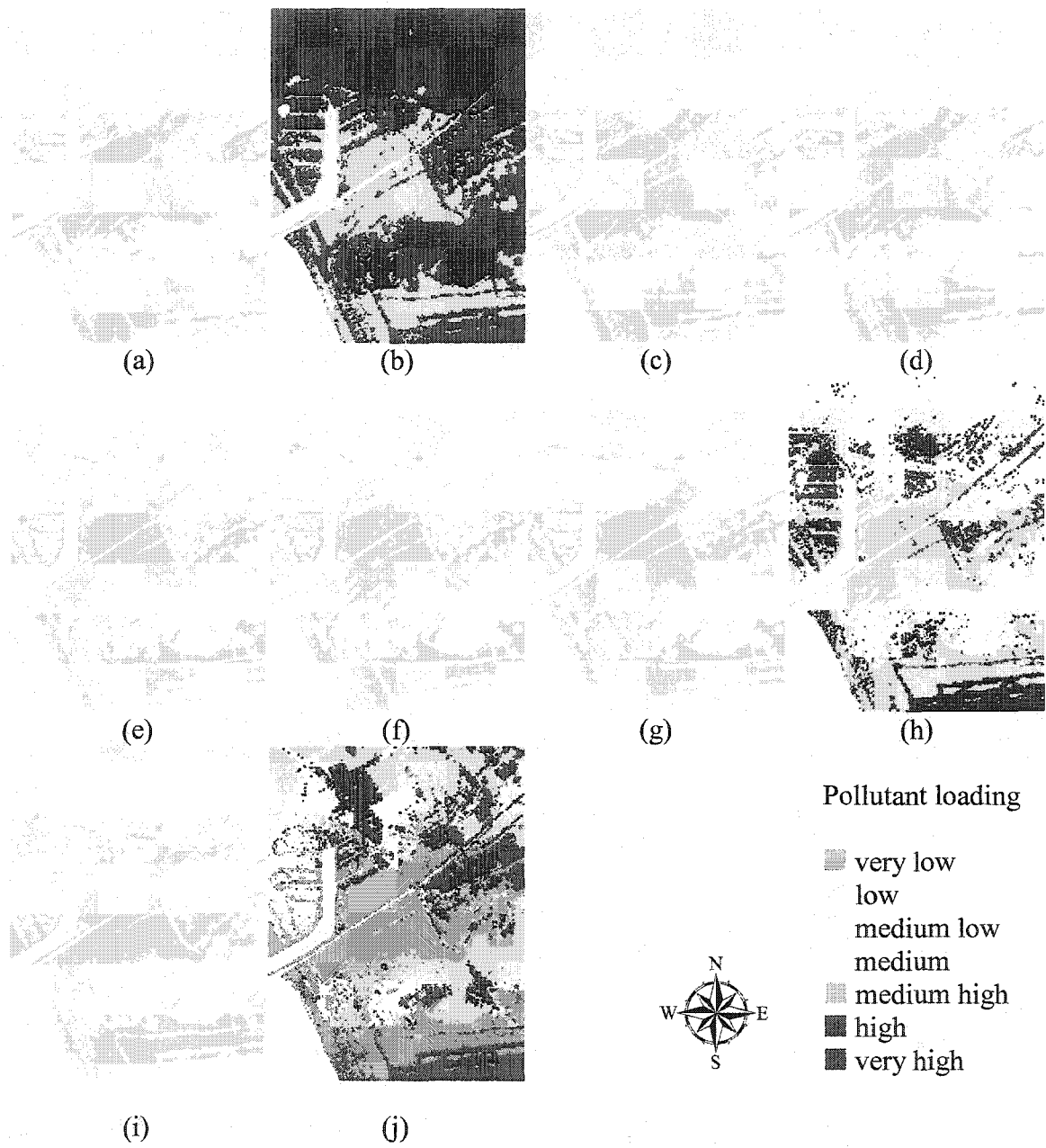
Figure 7.9 Normalized pollutant loading maps from Landsat ETM^{+} image (a) COD (b) BOD$_5$ (c) TSS (d) TKN (e) NO$_{2\&3}$ (f) TP (g) Cu (h) Pb (i) Zn (j) oil and grease

152

open land uses. Transportation areas correspond to the highest pollutant loading areas for all pollutants except Zn, which was highest in commercial and industrial land uses. This shows the significance of transportation land uses for stormwater management. Multiple-family residential areas are also important for Pb, which may result from greater numbers of vehicles associated with the land use. Table 7.2 shows percentage of normalized pollutant loading areas. Most pollutants had only approximately 1/3 of the areas as low loading areas and the others generated higher loads except O&G.

The accuracies for each case are given in Figure 7.10. Accuracies for BOD_5, TSS, TKN, TP, and Cu were higher than other water quality parameters. The overall accuracies were all above 90 % and κ coefficients were above 85% using spectral data only. For other pollutants, overall accuracies ranged from 80% to 88% and κ coefficients from 66% to 79% when using spectral data only. Omission errors of the highest pollutant load varied depending on the pollutant types. Omission errors for BOD_5, TSS, TKN, TP and Cu were the lowest, less than 10 %. The errors for other pollutants were above 20% and especially the error of COD was above 40%. Commission errors of the lowest pollutant load from other areas showed little variation for different pollutants compared with the omission errors. The commission errors of all pollutants were varied only between 9% and 14%. The magnitudes of commission errors of COD, $NO_{2\&3}$, Cu, Pb, Zn and O&G became smaller compared with the omission errors whereas those of BOD_5, TSS, TKN, and TP became larger.

Including locational ancillary data improved overall accuracies of COD, $NO_{2\&3}$, heavy metals and O&G by up to 7%. The improvement of κ coefficient was more

153

Table 7.2 Percentage of normalized pollutant loading area of each pollutant

	COD	BOD$_5$	TSS	TKN	NO$_{2\&3}$	TP	Cu	Pb	Zn	O&G
very low	0	0	0	0	0	0	0	0	0	31
low	30	29	29	29	30	29	31	31	33	26
med-low	0	0	71	0	13	0	0	0	27	0
medium	63	71	0	71	57	71	27	0	22	0
med-high	7	0	0	0	0	0	41	51	18	17
high	0	0	0	0	0	0	0	0	0	26
very high	0	0	0	0	0	0	0	18	0	0
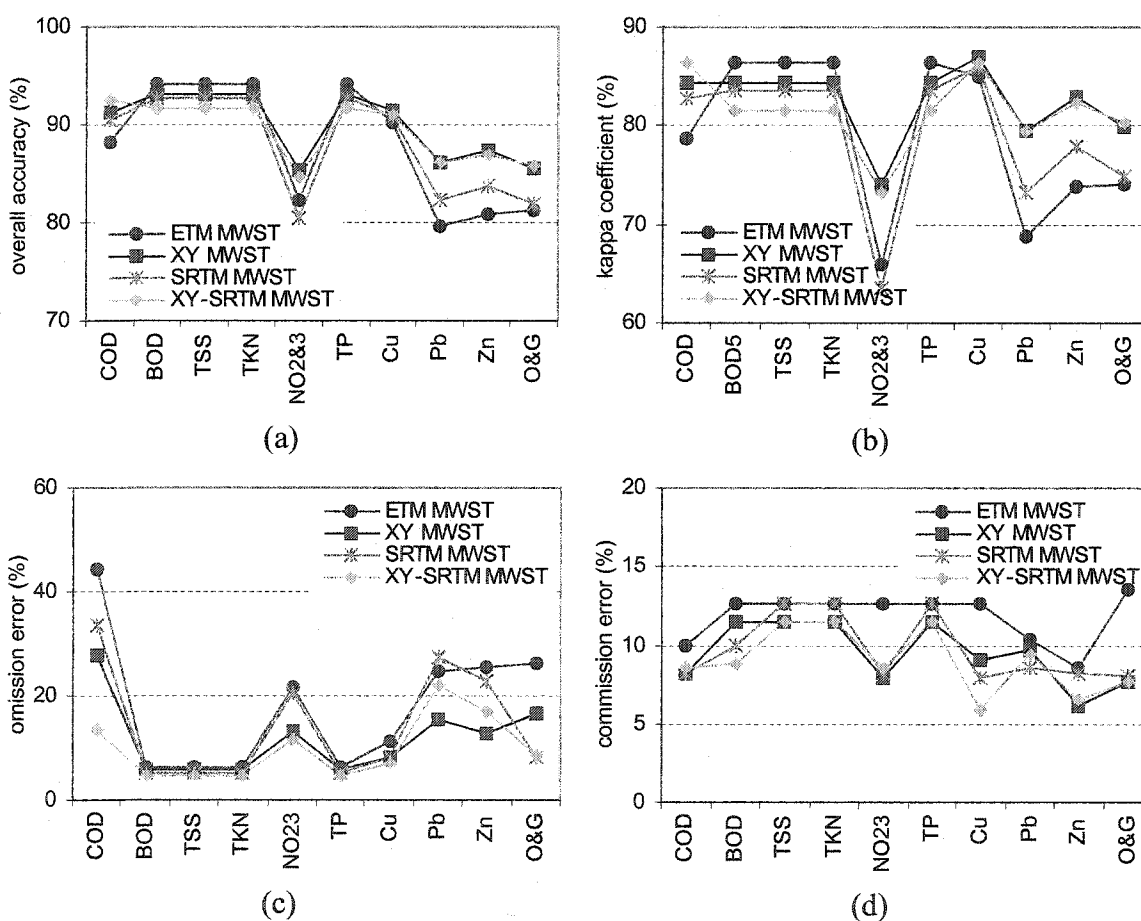
Figure 7.10 Accuracies of normalized pollutant loading estimates from satellite imagery and ancillary data (a) overall accuracy (b) kappa coefficient (c) omission errors (d) commission errors

154

significant, which is up to 11%. Including locational data reduced omission errors for COD, $NO_{2\&3}$, heavy metals, and O&G by up to 16%. Commission errors were also reduced especially for $NO_{2\&3}$, Cu, Zn, and O&G by up to 6% but the decrease was not as significant as omission errors. However, including elevational data was not so useful as incorporating locational data for improving accuracies and reducing omission and commission errors.

7.3 New Classification System for Stormwater Modeling

The results given in the previous chapters imply that classification system for stormwater pollutant loads should be developed differently for individual pollutant. The only common class for all pollutants is low loading area. For the other classes, each pollutant exhibited different classification scheme. For example of normalized scheme, COD had only transportation land use as higher pollutant loading areas than other land uses. But other pollutants often included commercial, public, and industrial land uses as higher pollutant loading areas. Conversely Zn does not have transportation land use as higher pollutant loading area.

The new classification system for managing stormwater pollutant loads was proposed as shown in Table 7.3. In this system, residential, commercial, public, industrial and transportation land uses are lumped together as the very high pollutant loading areas if BOD_5 is of concern. Similarly, multiple-family residential and transportation land uses belong to the very high Pb loading areas. For O&G, commercial, public, industrial, and

155

Table 7.3 New classification system for normalized stormwater pollutant loadings

	COD	BOD$_5$	TSS	TKN&TP	NO$_{2\&3}$	Cu	Pb	Zn	O&G
very low									o
low	o	o	o	o	o	o	o	o	s
med-low			s,m,c,p,i,t		m			s	
medium	s,m,c,p,i			s,m,c,p,i,t	s,c,p,i,t	s		m, t	
med-high	t					m,c,p,i,t	s,c,p,i,	c,p,i,	m
high									c,p,i,t
very high		s,m,c,p,i,t					m,t		

Note that 's' is single-family residential, 'm' is multiple-family residential, 'c' is commercial, 'p' is public, 'i' is industrial, 't' is transportation and 'o' is open land use.

transportation land uses become high loading areas. These areas should be addressed first to reduce the stormwater pollution impact on receiving water quality.

7.4 Discussion

The methodologies we used have shown that satellite image classification is useful in estimating stormwater pollutant loads. Our methodology has advantages over conventional approaches based on ground surveys due to its applicability to areas where land use data are not available, such as in developing countries.

Estimating stormwater pollutant loads converted from land use classification is useful although some urban land uses in the satellite image exhibit similar spectral signatures and are not easily distinguished. Commercial and public land uses were often misclassified. Omission and commission errors for commercial and public land uses were more than 40%. Fortunately, misclassification of commercial and public land uses does not significantly affect the spatial estimates of pollutant loads since they often generate similar pollutant loads.

The stormwater pollutant loading maps converted from land use classification show that low pollutant loading areas correspond to open land uses, which exhibit low runoff coefficient, whereas high pollutant loading areas mostly correspond to the areas with high runoff coefficient, i.e. commercial, public, industrial and transportation land uses. Those land uses with high runoff coefficient corresponded to at least medium or high pollutant loading areas. Therefore, it is important to identify areas with high runoff

157

coefficient to properly manage stormwater pollutant loads. This is logical conclusion because stormwater may be retained by vegetation or infiltrated through the pervious ground whereas stormwater overland runoff is conveyed into receiving water through impervious surfaces. With the exception of Zn, multiple-family residential land use also generates high pollutant loads, despite its medium runoff coefficient.

The comparison of between EMC maps and pollutant loading maps shows that high EMC generating areas do not always match high pollutant loading areas. For example, open areas generated high TSS concentrations, but generated only low loads due to the very low runoff coefficient. Furthermore, areas with low EMCs are often high pollutant loading areas; for example, COD, TSS, TKN, and TP maps show that commercial, public, and industrial land uses had low EMCs, but high pollutant loads. These results emphasize that low EMC generating areas should not be underestimated or ignored for their potential impact on the receiving water quality. This results from the ratio of high to low EMCs ranges from 1.6 to 3.1 (except for O&G), while the ratio of runoff coefficients is 7.4. These differences translate into higher ranges of pollutant loads. For example, the ratio between low and high EMCs for Cu, Pb, and Zn is only 2 to 3 fold whereas the ratio between low and high pollutant loads is 12 to 21 fold.

The new methodology predicted pollutants from satellite imagery with accuracies ranging from 80% to 94%. The land use classification approach using a Bayesian network was less accurate, yielding 81% accuracy with locational data (71% without locational data). Therefore, the new methodology improved accuracy up to 13% when using locational data and up to 23% without locational data. The new methodology

158

provides better prediction and is a promising method for future environmental planning and management. Moreover, the new method does not require potentially expensive land use data based on ground surveys.

Figure 7.11 compares each pollutant loading areas from SCAG land use data, land use classification, and satellite imagery. The result shows that percentage of each pollutant loading area for different approaches are similar. However, estimates directly from satellite imagery gives more high pollutant loading areas for TSS, TKN, TP and $NO_{2\&3}$ and less low loading areas for all pollutants except O&G. This result also demonstrates that the new approach is valuable alternative and useful for managing stormwater pollution.

The normalized pollutant loading maps show important differences among pollutant types. Heavy metals and O&G emission rates varied by 12 to 21 fold between the lowest and the highest emission rates, whereas COD, TSS, TKN, TP, and $NO_{2\&3}$ varied by only 3 to 9 fold. BOD_5 was the extreme case since the difference was over 50 fold. This suggests that there is greater opportunity to impact BOD_5, heavy metals, and O&G emission rates by identifying high emitters. Environmental planners and regulators need to be aware of these important differences. The new strategy provides this information along with spatial information to locate environmental opportunities.

The example in this study is small and is intended to be illustrative of larger applications. The simple ratios of pollutant loads can be inferred from the runoff and EMC data, but cannot be spatially identified. The value of the thematic maps is not only to determine the high loading areas but also to locate and visualize them so that planners

159
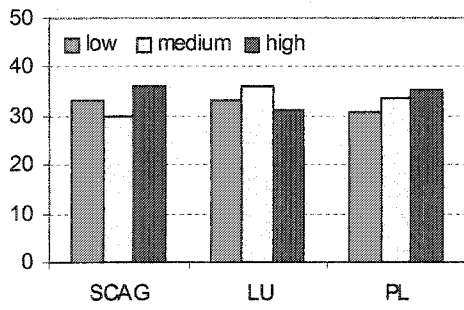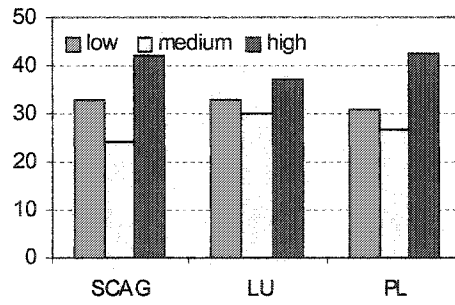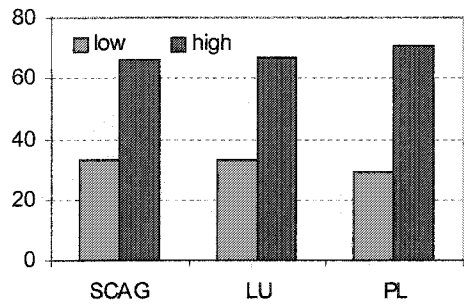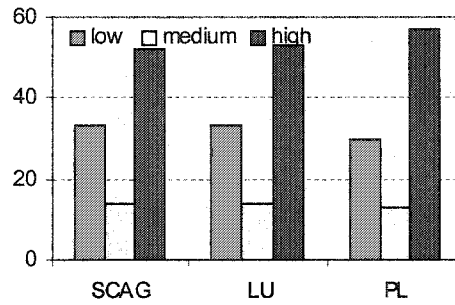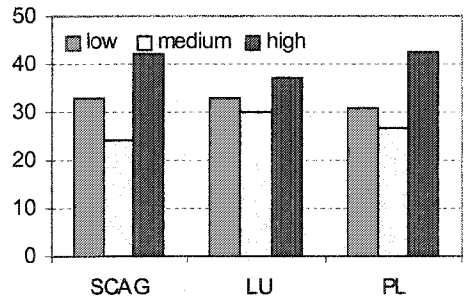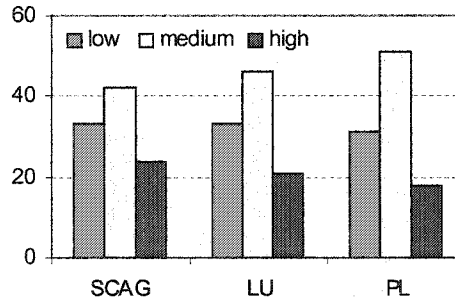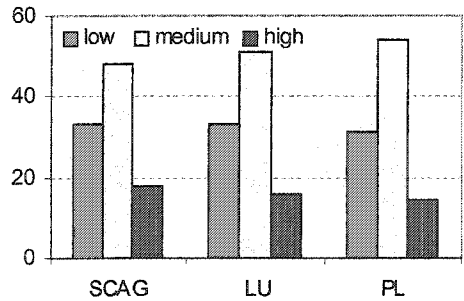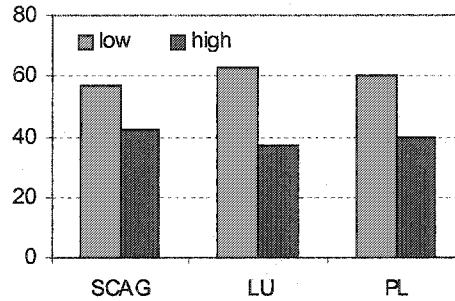
Figure 7.11 Comparison of pollutant loading areas (a) COD (b) BOD_5 (c) TSS, TKN and TP (d) $NO_{2\&3}$ (e) Cu (f) Pb (g) Zn (h) oil and grease

160

can better develop BMPs and regulation agencies can establish total mass daily limits (TMDL). The Bayesian technique also reclassified areas that are chronically misclassified, providing more accurate estimates.

161

Chapter VIII

Conclusion

This study has shown that satellite image classification is useful in estimating urban land use and stormwater pollutant loads. Compared with conventional approaches based on ground surveys, satellite imagery offers several benefits. Satellite imagery can be applicable to areas where land use data are not available. Information from satellite images can be optimized for environmental purposes since potential errors in conventional land use data are avoided.

The study has shown that Bayesian networks are useful for satellite image classification of urban land use and pollutant loads in an example watershed. Both naïve Bayesian classifiers and Bayesian networks based on MWST algorithm were useful, but the latter structure was better in many ways. First, they reveal the relationships among variables such that we can predict the class node value even though particular band

162

information is missing. The missing band information can be inferred from the relationship from other bands. This is a valuable property since it does not require correction for sensor system error when a detector fails to function during a scan. Secondly, the networks identify the most informative input variables to classification. For urban land use and pollutant loading classification, bands 1(blue), 5(middle IR), and 6(thermal IR) are the most important in most cases. Thirdly, the problem domain can be minimized with a reduced number of input variables required for classification once the most contributing inputs were found. Finally, they reduce the risk of overestimating or underestimating stormwater pollutant loads as well as urban land uses in the given area.

For both urban land use and stormwater pollutant loading classification, incorporating ancillary data improved the classification performance. Locational information improved the classification accuracy and reduced omission and commission errors more significantly than other ancillary data such as digital elevation model (DEM) and image transform data. Therefore, incorporating locational ancillary data is useful and practical since they can be simply obtained from the given satellite image.

The stormwater pollutant loading maps show that pollutant loads were more affected by runoff coefficients than EMCs for most of the land uses. This results from the fact that there is more variation in runoff coefficients than EMCs among the land uses. Therefore, elimination or reduction of impervious surfaces may be more important than BMPs that reduce the EMCs by partial treatment. In typical urban areas, multiple-family residential, commercial, industrial, and transportation land uses contribute proportionally

163

more pollutant loads. These land uses should be targeted for the first application of BMPs.

The new approach to estimate stormwater pollutant loads directly from satellite imagery has advantages over approaches based upon land use information. Compared to ground surveyed land use, the new approach provides information optimized for stormwater management purpose. The new approach also provides better accuracy compared to the pollutant loading estimates converted from land use classification with satellite images. Therefore, training Bayesian network classification based upon pollutant loadings is a promising alternative to conventional land use models.

The normalized pollutant loading maps show different impacts among pollutant types on receiving waters. The maps demonstrate that there is greater opportunity to impact BOD_5, heavy metals, and oil and grease loads by identifying high emitters. Urban planners and regulation agencies need to recognize these important differences. The new strategy provides this information along with spatial information to locate environmental opportunities.

This approach can be extended to regional and global scales. The spatial information from stormwater pollutant load thematic maps is important to detect the high loading areas and to locate and visualize them. This will be useful to develop BMPs and TMDLs for urban planners and local regulation agencies. The Bayesian networks were an effective classification method in that they provide more accurate estimates by reclassifying areas that are chronically misclassified.

164

Chapter IX

Future Study

For urban land use classification, Landsat images have limited ability to separate urban features based upon the spatial resolution. Therefore, higher resolution image such as IKONOS and aerial photos will be useful to delineate more detailed urban features. Moreover, higher resolution images can be used to resolve the mixed pixel problem using subpixel analysis. Unfortunately, using high resolution images will increase computational cost. In this case, high resolution images might not be applicable to a large coverage for watershed management. For stormwater management, we can use both low and high resolution images. For example, low resolution can be used for watershed scale and high resolution image can be used for local scale.

In order to enhance image classification, other ancillary data, in addition to locational data, DEM and image transformation, can be incorporated. For example, radar

165

images can identify man-made structures on the landscape since man-made structures have strong radar returns. Texture and contexture information can be included for further classification. In this case, information fusion of multisource data can be useful (Slonecker, 2001). Alternatively, Bayesian networks can be particularly useful since they provide the most contributing information to classification among multisource data. In addition, other study areas can be explored to investigate the effect of incorporating DEM since our study did not show much elevational variation.

An alternative way of estimating pollutant loads might be to use impervious surfaces instead of land use, since impervious surface estimates may be more easily predicted from satellite imagery. Urban land uses are highly heterogeneous such that they can contain a mixture of different urban features with different portions. For example, residential and commercial land uses could contain buildings, pavement, driveways and vegetation (Clapham, 2003), which have different imperviousness and impact on stormwater quality. Moreover, different land uses might have similar imperviousness and impact on stormwater quality if the land uses consist of similar urban features, i.e. roof materials (Stefanov, 2001). Therefore, impervious surface measurements could be more effective than land use classification for stormwater management. In this case, the relationship between EMCs and/or pollutant loads of each water pollutant type, and imperviousness should be investigated.

A future research topic might be to estimate land use change in Santa Monica Bay watershed and predict urban sprawl and stormwater pollutant loading changes in a given watershed. Landsat data are useful in this sense because they provide temporally

166

accumulated data over last three decades. For example, the temporal changes of Ballona

Wetlands can also be monitored from satellite image.

167

References

Ackerman, D, and Schiff, K. (2003) Modeling storm water mass emissions to the Southern California Bight, *Journal of Environmental Engineering, ASCE*, 129 (4), 308-317

Anderson, J.R., Hardy, E., Roach, J., and Witmer, R. (1976) *A land-use and land-cover classification system for use with remote sensor data*, U. S. Geological Survey Profession Paper 964

Arnold, C.A., Jr., and Gibbons, C.J. (1996) Impervious surface coverage: the emergence of a key urban environmental indicator, *Journal of the American Planning Association*, 62(2), 243–258

Asaf, L., Native, R., Shain, D., Hassan M. and Geyer, S. (2004) Controls on the chemical and isotopic compositions of urban stormwater in a semiarid zone, *Journal of Hydrology*, in press

Baeza, J.A., Gabriel, D. and Lafuente, J. (2002) Improving the nitrogen removal efficiency of an A^2/O based WWTP by using an on-line Knowledge Based Expert System, *Water Research*, 36(8), 2109-2123

Ballona Wetlands Foundation, http://www.ballona-wetlands.org

Bang, J. (2002) *Hidden Nodes in Bayesian Networks and Their Application to Prognostic Analysis of Hepatitis C*, Ph.D. Dissertation, Imperial College

Bang, J.W. and Gillies, D.F. (2002) Using Bayesian networks to model the prognosis of hepatitis C, in *Proceedings of the 7th Intelligent Data Analysis and Pharmacology (IDAMAP) workshop, 15th European Conference on Artificial Intelligence*, Lyon, France, 7-12

Bannerman, R.T., Owens, D.W., Dodds, R.B. and Hornewer, N.J. (1993) Sources of pollutants in Wisconsin stormwater, *Water Science and Technology*, 28 (3-5), 241-259

Bay, S., Jones, B.H., and Schiff, K (1999) *Study of the Impact of Stormwater Discharge on the Beneficial Uses of Santa Monica Bay*, Executive Summary Prepared for Los Angeles County Department of Public Works, Alhambra, California

Bay, S., Jones, B.H., Schiff, K., and Washburn, L. (2003) Water quality impacts of stormwater discharges to Santa Monica Bay, *Marine Environmental Research*, 56, 205–223

Benediktsson, J. A., Swain, P. H., Ersoy, O. K., and Hong, D. (1990) Classification of very high dimensional data using neural networks, in *Proceedings of IGARSS*, Washington, D.C., 1269-1272

Birth, G.S. and McVey, G. (1968) Measuring the color of growing turf with a reflectance spectrophotometer, *Agronomy Journal*, 60, 640-643

Bischof, H., Schneider, W., and Pinz, J. (1992) Multispectral classification of landsat-images using neural networks, *IEEE Transactions on Geoscience and Remote Sensing*, 30, 482-490

Bolstad, P.V., and Lillesand T. M. (1992) Rule-base classification models: Flexible integration of satellite imagery and thematic spatial data, *Photogrammetric Engineering and Remote Sensing*, 58(7), 965-971, 1992

Borsuk, M.E. and C.A. Stow (2000) Bayesian parameter estimation in a mixed-order model of BOD decay, *Water Research*, 34, 1830-1836.

Borsuk, M.E., Stow, C.A. and Reckhow K.H. (2004) A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis, *Ecological Modeling*, 173, 219–239

Brabec, E., Schulte, S., and Richards, P.L (2002) Impervious surfaces and water quality: A review of current literature and its implications for watershed planning, *Journal of Planning Literature*, 16(4), 499-514

Brezonic, L.P. and Stadelmann, H.T. (2002) Analysis and predictive models of stormwater runoff volume, loads, and pollutant concentrations from watersheds in the Twin cities metropolitan area, Minnesota, USA, *Water Research*, 36, 1743-1757

Brion, G.M. and Lingireddy, S. (1999) A neural network approach to identifying non-point sources of microbial contamination, *Water Research*, 33(14), 3099-3106

Burian, S.J., Brown, M.J., and McPherson, T.N. (2002) Evaluation of land use/land cover datasets for urban watershed modeling, *Water Science and Technology*, 45 (9), 269-276

Carpenter, G.A., Gopal, S., Macomber, S., Martens, S., Woodcock, C.E., and Franklin J. (1999) A Neural Network Method for Efficient Vegetation Mapping, *Remote Sensing of Environment*, 70(3), 326-338

Catlett, J. (1991) On changing continuous attributes into ordered discrete attributes in *Proceedings of the European Working Session on Learning*, Berlin, Germany, 164-178

CCRS (Canadian Centre for Remote Sensing, 2002), *Fundamentals of Remote Sensing*, http://www.ccrs.nrcan.gc.ca/ccrs/learn/tutorials/fundam/fundam_e.html

Cermak, R.J., Feldman, A. and Webb, R.P. (1979) Hydrologic land use classification using Landsat, in *Satellite Hydrology*, Deutsch, M., Wiesnet, D.R., and Rango, A. (edi), American Water Resources Associations, Minneapolis, MN, 262-269

Cerquides, J. and Mantaras, R.L. (1997) Proposal and empirical comparison of a parallelized distance-based discretization method, in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA, 139-142

Chan, C.C., Batur, C., and Srinivasan, A. (1991) Determination of quantization intervals in rule based model for dynamic systems, in *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, Charlottesville, VA, 1719-1723

Charniak, E. (1991) Bayesian network without tears, *AI Magazine*, 12(4), 50-63

Cheng, J., Greiner, R., Kelly, J., Bell, D., and Liu, W. (2002) Learning Bayesian networks from data: An information-theory based approach, *Artificial Intelligence*, 137(1-2), 43-90

Chiew, F.H.S. and McMahon, T.A. (1999) Modelling runoff and diffuse pollution loads in urban areas, *Water Science and Technology*, 39(12), 241-248

Chiu, D.K.Y., Cheung, B., and Wong, A.K.C. (1990) Information s synthesis based on hierarchical entropy discretization, *Journal of Experimental and Theoretical Artificial Intelligence*, 2, 117-129

Chmielewski, M.R. and Grzymala-Busse, J.W. (1994) Global discretization of continuous attributes as preprocessing for Lachine learning, in *Third International Workshop on Rough Sets and Soft Computing*, 294-301

Chong, H.G. and Wally, W.J. (1996) Rule-based versus probabilistic approached to the diagnosis of faults in wastewater treatment processes, *Artificial Intelligence in Engineering*, 1, 265-273

Chow, C.K., and Liu, C. N. (1968) Approximating discrete probability distributions with dependence trees, *IEEE Transactions on Information theory*, IT-14, (3), 462-467

Civco, D. L. (1991) Landsat TM image classification with an artificial neural network, in *Proceedings of ASPRS-ACSM Annual Meeting*, Baltimore, MD, 67-77

Civco, D.L. and Hurd, J.D. (1997) Impervious surface mapping for the state of Connecticut, in *Proceedings of ASPRS/ACSM Annual Convention*, Seattle, WA, 3, 124-135

Clapham, W.B. (2003) Continuum-based classification of remotely sensed imagery to describe urban sprawl on a watershed scale, *Remote Sensing of Environment*, 86, 322-340

Clark, P. and Niblett, T. (1987) Induction in noisy domains, in *Proceedings of 2nd European Machine Learning Conference (EWSL)*, Bled, Yugoslavia, 11-30

Comas, J., Rodríguez-Roda, I., Sànchez-Marrè, M., Cortés, U., Freixó, A., Arráez, J. and Poch, M. (2003) A knowledge-based approach to the deflocculation problem: integrating on-line, off-line, and heuristic information, *Water Research*, 37(10), 2377-2387

Congalton, R. (1991) A Review of assessing the accuracy of classifications of remotely sensed data, *Remote Sensing of Environment*, 37, 35-46

Corbett, C.W., Wahl, M., Porter, D.W., Edwards, D., and Moise, C. (1997) Nonpoint source runoff modeling: A comparison of a forested watershed and a n urban watershed on the South Carolina coast, *Journal of Experimental Marine Biology and Ecology*, 213, 133-149

Crist, E.P. (1985) A Thematic Mapper Tasseled Cap equivalent for reflectance factor data, *Remote Sensing of Environment*, 17, 301-.06

Davis. A. P., Shokouhian, M., and Ni, S (2001) Loading estimates of lead, copper, cadmium and zinc in urban runoff from specific sources, *Chemosphere*, 44, 997-1009

Deguchi, C. and Sugio, S. (1994) Estimation for percentage of impervious area by the use of satellite remote sensing imagery, *Water and Science Technology*, 29, 135-144

Dobbertin, M. and Biging, G.S. (1996) A simulation study of the effect of scene autocorrelation, training size and sampling method on classification accuracy, *Canadian Journal of Remote Sensing*, 22, 360-367

Dojiri, M., Yamaguchi, M., Weisberg, S.B., and Lee H.J. (2003) Changing anthropogenic influence on the Santa Monica Bay watershed, *Marine Environmental Research*, 56, 1–14

171

Dougherty, J., Kohavi, R., and Sahami, M (1995) Supervised and unsupervised discretization of continuous features, in *Proceedings of the 12th International Conference on Machine Learning*, Los Altos, 194-202

Driscoll, E.D., Shelly, P.E., and Strecker, E.W. (1990) *Pollutant Loadings and Impacts from Stormwater Runoff, Volume III:* Analytical investigation and research report, FHWA-RD-88-008, Federal Highway Administration

Duda, R. O., Hart, P. E., and Stork, D. G. (2001) *Pattern Classification*, Wiley, USA

Duncan, H.P. (1995) *A Review of Urban Stormwater Quality Processes*, Report No. 95/9, Cooperative Research Centre for Catchment Hydrology, Melbourne, Austrailia

El-Din, A.G. and Smith, D.W. (2002) A neural network model to predict the wastewater inflow incorporating rainfall events, *Water Research*, 36(5), 1115-1126

Environment Protection Agency (EPA, 1994) *The Quality of Our Nation's Water: 1992*, #EPA-841-S-94-002, Washington, D.C., USEPA office of water

EPA http://www.epa.gov/owow/estuaries/programs/ smb.htm

EPA, http://www.epa.gov/owow/nps/nps_edu/urbanx3.htm

Fahsi, A., Tsegaye, T., Tadesse, W., and Coleman, T. (2000) Incorporation of digital elevation models with Landsat-TM data to improve land cover classification accuracy, *Forest Ecology and Management*, 128, 57-64

Fayyad, U.M. and Irani, K.B. (1996) Discretizing continuous attributes while learning Bayesian networks, in *Proceedings of 13th International Conference on Machine Learning*, 157-165

Fitzpatrick-Lins, K. (1981) Comparison of sampling procedures and data analysis for a land-use and land-cover map, *Photogrammetric Engineering and Remote Sensing*, 47(3), 343-351

Flanagan, M. and Civco, D.L. (2001) Subpixel impervious surface mapping, in *Proceedings of 2001 ASPRS Annual Convention*, St. Louis, MO

Flygare, A. M. (1997) A comparison of contextual classification methods using Landsat TM, *International Journal of Remote Sensing*, 18, 3835-3843

Foody, G. M. (2000) Estimation of sub-pixel land cover composition in the presence of untrained classes, *Computers and Geosciences*, 26, 469-478

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

Foody, G.M., Campbell, N.A., Trood, N.M., and Wood, T.F. (1992) Derivation and applications of probabilistic measures of class membership from the maximum-likelihood classification, *Photogrammetric Engineering and Remote Sensing*, 58(9), 1335-1341.

Foody, G.M., McCulloch, M.B., and Yates, W.B. (1995a) Classification of remotely sensed data by an artificial neural network: issues related to training data characteristics, *Photogrammetric Engineering and Remote Sensing*, 61, 391-401

Foody, G. M., McCulloch, M. B., and Yates, W. B. (1995b) The effects of training set size and composition on artificial neural network, *Photogrammetric Engineering and Remote Sensing*, 58, 1459 - 1460

Förster, J. (1996) Patterns of roof runoff contamination and their potential implications on practice and regulation of treatment and local infiltration, *Water Science and Technology*, 33 (6), 39-48

Friedman, N. and Goldszmidt, M. (1996) Discretizing continuous attributes while learning Bayesian networks, in *Proceedings of 13th International Conference on Machine Learning*, 157-165

Friedman, N., and Koller, D. (2003) Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks, *Machine Learning, 50 (1-2)*, 95-126

Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian network classifiers, *Machine Learning*, 29, 131-163

Gannon, J.J., and Busse, M.K. (1989) *E.Coli* and enterococci levels in urban stormwater, river water and chlorinated treatment plant effluent, *Water Research*, 23 (9), 1167-1176

Gesch, D., Oimoen,M., Greenlee, S., Nelson, C., Steuck, M., and Tyler, D (2002) The national elevation dataset, *Photogrammetric Engineering and Remote Sensing*, 68 (1), 5-12

Goldman, R. (1990) *A Probabilistic Approach to Language Understanding*, Technical Report, CS-90-34, Dept. of Computer Science, Brown University

Gong P., Marceau, D. J., and Howarth, P. J. (1992) A Comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data, *Remote Sensing of Environment*, 40, 137-151

Gopal, S., Woodcock, C. E., and Strahler, A. H. (1999) Fuzzy neural network classification of global land cover from a 1° AVHRR data set, *Remote Sensing of Environment*, 67, 230–243

Goward S.N., Masek, J.G., Williams, D.L., Irons, J.R., and Thompson, R.J. (2001) The Landsat 7 mission terrestrial research and applications for the 21st century, *Remote Sensing of Environment*, 78, 3-12

Gromaire, M.C., Garnaud, S., Saad, M. and Chebbo, G (2001) Contribution of different sources to the pollution of wet weather lows in combined sewers, *Water Research*, 35 (2), 521-533

Haack, B. (1983) An analysis of Thematic Mapper Simulator data for urban environments, *Remote Sensing of Environment*, 13, 265–275

Haack, B., Bryant, N., and Adams, S. (1987) An assessment of Landsat MSS and TM data for urban and near-urban land-cover digital classification, *Remote Sensing of Environment*, 21, 201–213

Haralick, R. M., Shanmugam, K. S., and Dinstein, I. (1973) Textural features for image classification, *IEEE Transactions on Systems, Man and Cybernetics*, 3, 610-622

Hay, A.M. (1979) Sampling designs to test land-use map accuracy, *Photogrammetric Engineering and Remote Sensing*, 45, 529-533

He, W., Wallinder, I.O. and Leygraf, C. (2001) A laboratory study of copper and zinc runoff during first flush and steady-state conditions, *Corrosion Science*, 43, 127-146

Heckerman, D. (1990) *Probabilistic Similarity Networks*, Technical Report, STAN-CS-1316, Dept. of Computer Science and Medicine, Stanford Univ.

Heckerman, D. (1995) *A Tutorial on Learning With Bayesian Networks*, Technical Report MSR-TR-95-06, Microsoft Research

Heermann, P. D., and Khazenie, N. (1992) Classification of multispectral remote sensing data using a back-propagation neural network, *IEEE Transactions on Geoscience and Remote Sensing*, 30, 81-88

Hepner, G. F., Logan T., Ritter, N. and Bryant, N. (1990) Artificial neural network classification using a minimal training set: comparison to conventional supervised classification, *Photogrammetric Engineering and Remote Sensing*, 56, 469-476

Hiirsalmi, M. (2000) *Method Feasibility Study: Bayesian Networks*, MODUS-Project Waste Water Case Study, Research Report TTE1-2000-29, VTT Information Technology, Espoo, Finland

Ho, K.M. and Scott, P.D. (1997) Zeta: A global method for discretization of continuous variables, in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA, 191-194

Holte, R.C. (1993) Very simple classification rules perform well on most commonly used datasets, *Machine Learning*, 11, 63-90

Hsu, C.N., Huang, H.J., and Wong, T.T (2000) Why discretization works for navie Bayesian classifiers, in *Proceedings of the 17th International Conference on Machine Learning*, 309-406

Huber W.C. (1993) Contaminant transport in surface water, in *Handbook of Hydrology*, Maidment, D.R. (edi) McGraw Hill Inc. New York, NY

Hubert-Moy, L., Cotonnec, A., Le Du, L., Chardin A. and Perez, P. (2001) A Comparison of Parametric Classification Procedures of Remotely Sensed Data Applied on Different Landscape Units, *Remote Sensing of Environment*, 75(2), 174-187

Hung, M-C, and Ridd, M.K. (2002) A subpixel classifier for urban land-cover mapping based on a maximum-likelihood approach and expert system rules, *Photogrammetric Engineering and Remote Sensing*, 68 (11), 1173-1180

Hutchinson, C, F. (1982) Techniques for combining Landsat and ancillary data for digital classification improvement, *Photogrammetric Engineering and Remote Sensing*, 48(1), 123-130

Jackson, P. (1990) *Introduction to Expert Systems*, Addison-Wesley

Jackson, T.J. (1975) Computer aided techniques for estimating the percent of impervious area from Landsat data, in *Proceedings of Workshop on the Environmental Applications of Multispectral Imagery, American Society of Photogrammetry*, Fort Belvoir, VA, 140-155

Jackson, T.J. and McCuen, R.H. (1979) Accuracy of impervious area values estimated using remotely sensed data, *Water Resources Bulletin*, 15 (2), 436-446

Jain, A. K, Mao, J. and Mohiuddin, K. M. (1996) Artificial neural networks: A tutorial, *Computer*, 29 (3), 31-44

Janssen, L. L. F. and van der Wel, F. J. M. (1994) Accuracy assessment of satellite derived land-cover data: a review, *Photogrammetric Engineering and Remote Sensing*, 60, 419-426

Jensen, F.V. (2001) *Bayesian Networks and Decision Graphs,* New York: Springer

Jensen, J.R., (1996) *Introductory Digital Image Processing: A Remote Sensing Perspective*, Prentice Hall, Upper Saddle River, NJ

Jensen, J.R. (2000) *Remote Sensing of the Environment: An Earth Resource Perspective*, Prentice Hall, Upper Saddle River, NJ

Ji, M. and Jensen, J.R. (1999) Effectiveness of subpixel analysis in detecting and quantifying urban imperviousness from Landsat Thematic Mapper imagery, *Geocarto International*, 14 (4), 31-39

Justice, C. O., Wharton, S. W. and Holben, B. N. (1981) Application of digital terrain data to quantify and reduce the topographic effect on Landsat data, *International Journal of Remote Sensing*, 2(3), 213-230

Kanellopoulos, I, Wilkinson, G.G., and Mégier, J. (1993) Integration of neural network and statistical image classification for land cover mapping, in *Proceedings of IGARSS*, Tokyo, Japan, 511-513

Kartikeyan, B., Majumder, K. L. and Dasgupta, A. R. (1995) An expert system for land cover classification, *IEEE Transactions on Geoscience and Remote Sensing*, 33(1), 58-66

Kelsey, H., Porter, D.E., Scott, G., Neet, M., and White, D. (2004) Using geographic information systems and regression analysis to evaluate relationships between land use and fecal coliform bacterial pollution, *Journal of Experimental Marine Biology and Ecology*, 298, 197-209

Kerber, R. (1992) Chimerge: Discretization of numeric attributes, in *Proceedings of the 10th National Conference on Artificial Intelligence*, San Jose, CA, 123-128

Kim, J., and Pearl, J. (1983) A Computational model for combined causal and diagnostic reasoning in inference systems, in *Proceedings of 8th International Joint Conference on Artificial Intelligence*, 190-193

Kittler, J., and Föglein, J. (1984) Contextual classification of multispectral pixel data, *Image and Vision Computing*, 2, 13-29

Kononenko, I. (1991) Semi-naïve Bayesian classifier, in *Proceedings of the 6th European Working Session on Learning*, 206-219, 1991

Kouth, R.J. and Thomas, G.S. (1976) The tasseled cap – A graphic description of the spectral-temporal development of agricultural crops as seen by Landsat, in *Proceedings of Machine Processing of Remotely Sensed Data*, West Lafayette, IN, 41-51

LADPW (Los Angeles County Department of Public Works, 2000) *Los Angeles County 1994-2000 Integrated Receiving Water Impacts Report*

Langley, P., Iba, W. and Thompson, K. (1992) An analysis of Bayesian classifiers, in *Proceedings of the 10th National Conference on Artificial Intelligence*, San Jose, CA, 223-228

Langley, P. and Sage, S. (1994) Induction of selective Bayesian classifiers, in *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, Seattle, WA

Lauritzen, S.L. and Spiegelhalter, D.J. (1986) Local computations with probabilities on graphical structures and their application to expert systems, *Journal of the Royal Statistical Society*, Vol. 50, No. 2, pp. 157-224

Lee, H-H. (2003) *Data Mining for Stormwater Systems with GIS and NeuroFuzzy Models*, Ph.D. Dissertation, University of California, Los Angeles

Li, H., Liu, Z., and Sun, W. (1993) A new approach to pattern recognition of remote sensing image using artificial neural network, in *Proceedings of IGARSS*, Tokyo, Japan, 713-715

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

Liu, H. and Setiono, R (1995) Chi2: Feature selection and discretization of numeric attributes, in *Proceedings of the 7th IEEE International Conference on Tolls with Artificial Intelligence*, Herndon, VA, 388-391

Liu, H., Hussain, F., Tan, C.L., and Dash, M. (2002) Discretization: An enabling technique, *Data Mining and Knowledge Discovery*, 6, 393-423

Lucas, P. and Gaag, L.V.D. (1991) *Principles of Expert Systems*, Addison-Wesley

Lucas, P., and Abu-Hanna, A. (1999) Prognostic methods in medicine, *Artificial Intelligence in Medicine*, 15, 105-119

Macdonald R., Hall, K., and Schreier, H. (1997) *Water Quality and Stormwater Contaminants in the Brunette River Watershed, British Columbia*, 1994/1995, Research report, Westwater Research unit, IRE, University of British Columbia

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

Mason, D. C., Corr, D. G., and Cross, A., Hoggs, D. C., Lawrence, D., Petrou, M., and Tailor, A. M. (1988) The use of digital map data in the segmentation and classification of remotely sensed data, *International Journal of Geopraphicl Information Systems*, 2(3), 195-215

Mather P. M. (1999) *Computer Processing of Remotely-Sensed Images: An Introduction*, Wiley, Chichester, UK

McPherson, T.N, Burian, S.J., Turin, H.J., Stenstrom, M.K., and Suffet I.H. (2002) Comparison of the pollutant loads in dry and wet weather runoff in a southern California urban watershed, *Water Science and Technology*, 45 (9), 255-261

Michie, D. and Al-Attar, A. (1991) Use of sequential Bayes with class probability trees, *Machine Intelligence*, 12, 187-202

Michie, D., Spiegelhalter, D.J., and Taylor, C.C. (1994) *Machine Learning, Neural and Statistical Classification*, Ellis Horwood Limited, Great Britain

Mikkelsen, P.S., Weyer G., Berry C., Walden, Y., Colandini, V., Poulsen, S., Grotehusmann, D., and Rohlfing, R. (1994) Pollution from urban stormwater infiltration, *Water Science and Technology*, 29, 293-302

Minami, M. (2000), *Using ArcMap*, ESRI

Mitchell, T.M. (1997) *Machine Learning*, McGraw Hill, Singapore

Morgan, K.M. Newland, L.W., Weber, E., and Busbey, A.B. (1993) Using Spot satellite data to map impervious cover for urban runoff predictions, *Toxicological and Environmental Chemistry*, 40, 11-16

Morrison, G.M., Revitt, D.M, Svensson, G., and Balmer, P. (1984) Variations of dissolved and suspended solid heavy metals through an urban hydrographic, *Environmental Technology Letters*, 7, 313-318

NASA JPL (National Aeronautics and Space Administration, Jet Propulsion Laboratory), http://www2.jpl.nasa.gov/srtm/

Neapolitan, R.E. (1990) *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*, New York: Wiley

Niblack, W. (1986) *An introduction to digital image processing*, Prentice-Hall

NOAA (National Oceanic and Atmospheric Administration), http://www.nwsla.noaa.gov/climate/cvc.html

Ozgur, N.H. and Stenstrom, M.K. (1994) Development of a knowledge-based expert system for process control of nitrification in the activated sludge process, *Journal of the Environmental Engineering Division*, ASCE, 120, 87-107

Pal, C., Swayne, D., and Frey, B. (2001) The automated extraction of environmentally relevant features from digital imagery using Bayesian multi-resolution analysis, *Advanced in Environmental Research*, 5(4) 435-444

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

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

Paola, J.D., and Schowengerdt, R.A. (1995b) A detailed comparison of backpropagation neural network and maximum-likelihood classification for urban land use classification, *IEEE transactions of Geoscience and Remote Sensing*, 33, 981-996

Park, M., and Stenstrom, M.K (2003) Landuse classification for stormwater modeling using Bayesian networks, in *Proceedings of the 7th International Specialised IWA Conference, Diffuse Pollution and Basin Management*, Dublin, Ireland

Pax-Lenney, M., and Woodcock, C. E. (1997) The effect of spatial resolution on the ability to monitor the status of agricultural lands, *Remote Sensing of Environment*, 61, 210-220

Pazzani, M.J. (1995) An iterative improvement approach for the discretization of numeric attributes in Bayesian classifiers, in *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, Montreal, Canada

Pearl, J. (1982) Fusion, Reverend Byes on inference engines: A distributed hierarchical approach, in *Proceedings of National Conference on Artificial Intelligence*, 133-136

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA

Pearl, J. (1999) *Bayesian Networks*, UCLA Cognitive Systems Laboratory, Technical Report (R-246), Revision I, July 1997 in MIT Encyclopedia of the Cognitive Sciences, Cambridge, MA

Peddle, D. R. and Franklin, S. E. (1991) Image texture processing and data integration for surface pattern discrimination, *Photogrammetric Engineering and Remote Sensing*, 57 (4), 413-420

Pitt, R. (1999) Source characterization, in *Innovative Urban Wet Weather Flow Management Systems*, Heaney, P.J., Pitt, R., Field, R. (edi), EPA/600/R-99/029, 1-64

Plunk, D.E., Morgan, J. and Newland, L. (1990) Mapping impervious cover using Landsat TM data, *Journal of Soil and Water Conservation*, 45 (5), 589-591

Quinlan, J.R. (1986) Introductions of decision trees, *Machine Learning*, 1, 81-106

Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA

Ready, P.J. and Wintz, P.A. (1973) Information extraction, SNR improvement and data compression in multispectral imagery, *IEEE Transactions on Communications*, COM-21 (10), 1123-1131

Rees, W.G. (2001) *Physical Principles of Remote Sensing*, Cambridge University Press, Cambridge, UK

Richards, J. A., and Jia X., (1999) *Remote sensing digital image analysis: an introduction*, Springer, Heidelberg, Germany

Ridd, M.K. (1995) Exploring a V-I-S (vegetation-impervious surface-soil) model for urban ecosystem analysis through remote sensing: comparative anatomy for cities, *International Journal of Remote Sensing*, 16, 2165–2185

Ridd, M.K., and Liu, J. (1998) A comparison of four algorithms for change detection in an urban environment, *Remote Sensing of Environment*, 63, 95–100

Riordan, E.J., Grigg, N.S, and Hiller, R.L (1978) Measuring the effects of urbanization on the hydrologic regimen, in *Proceedings of International Symposium on Urban Strom Drainage Conference*, Southampton, UK, 496-511

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*, New York: Cambridge University Press

Rouse, J.W., Haas, R.H., Schell, J.A. and Deering, D.W. (1974) Monitoring vegetation systems in the Great Pains with ERTS, in *Proceedings of 3rd Earth Resources Technology Satellite-1 Symposium*, Washington D.C., NASA SP-351, 309-317

Rumelhart, D.E., Hinton, G.E., and Williams, R. J. (1986) Learning internal representations by error propagation, *Parallel Distributed Processing: Explorations in the Microstruction of Cognition*, edited by Rumelhart, D. E. and McClelland, J. L., The MIT Press

Russell, S., and Norvig, P. (1995) *Artificial Intelligence: A Modern Approach*, Upper saddle river, NJ: Prentice Hall

Sabins, F.F. (1997) *Remote Sensing: Principles and Interpretation*, Freeman and company, New York, NY

Sahely, B.S.G.E. and Bagley, D.M. (2001) Diagnosing upsets in anaerobic wastewater treatment using Bayesian belief networks, *Journal of Environmental Engineering*, ASCE, 127(4), 302-310.

Salmerón, A., Cano, A., and Moral, S. (2000) Importance sampling in Bayesian networks using probability trees, *Computational Statistics and Data Analysis*, 34(4), 387-413

Sangüesa. R. and Burrell, P. (2000) Application of Bayesian network learning methods to waste water treatment plants, *Applied Intelligence*, 13, 19-40

SCAG (Southern California Association of Governments) http://wagsdata.scag. ca.gov

Schowengerdt, R. A. (1997), *Remote Sensing, Models, and Methods for Image Processing*, Academic Press, San Diego, CA

Schueler, T.R. (1994) The importance of imperviousness, *Watershed Protection Techniques*, 1(3), 100-111

Shannon, C. and Weaver, W. (1949) *The Mathematical Theory of Information*, Urbana: University of Illinois Press

Shinya, M., Tsuruho, J., Konoshi, T., and Ishikawa, M. (2003) Evaluation of factors influencing diffusion of pollutant loads in urban highway runoff, *Water Science and Technology*, 47 (7-8), 227-232

Short, N. (1991) A real-time expert system and neural network for the classification of remotely sensed data, in *Proceedings of ASPRS-ACSM Annual Meeting*, Baltimore, MD, 406-418

Sleavin, W., Civco, D.L., Prisloe, S., and Giannotti, L. (2000) Measuring impervious surfaces for non-point source pollution modeling, in *Proceedings of 2000 ASPRS Annual Convention*, Washington, D.C.

Sliva, L. and Williams, D.D. (2001) Buffer zone versus whole catchment approaches to studying land use impact on river water quality, *Water research*, 35 (14) 3462-3472.

Slonecker E.T., Jennings D. B., Garofalo, D. (2001) Remote sensing of impervious surfaces, *Remote Sensing Reviews*, 20, 227-255

Smith, A.J. (2000) *Subpixel Estimates of Impervious Surface Cover from Landsat TM Imagery*, M.A. Scholarly paper, Geography Department, University of Maryland, College Park

Smullen, J.T., Shallcross, A.L., and Cave, K.A. (1999) Updating the U.S. nationwide urban runoff quality data base, *Water Science and Technology*, 39 (12) 9-16

Spiegelhalter, D.J. (1986) Probabilistic Reasoning in predictive expert systems, in Kanal, L.N. and Lemmer, J.F. *Uncertainty in Artificial Intelligence*, Elsevier Science Publishers, North-Holland, Amsterdam, 47-67

Stefanov, W.L., Ramsey, M.S., and Christensen, P.R. (2001) Monitoring urban land cover change; An expert system approach to land cover classification of semiarid to arid urban centers, *Remote Sensing of Environment*, 77 (2), 173-185

Stehman, S.V., and Czaplewski, R.L. (1998) Design and analysis for thematic map accuracy assessment: Fundamental principles, *Remote Sensing of Environment*, 64(3), 331-344

Stenstrom, M. K., Silverman, G.S. and Bursztynsky, T.A. (1984) Oil and grease in urban stormwaters, *Journal of the Environmental Engineering, ASCE*, 110(1), 58-72

182

Stenstrom, M.K. and Strecker, E. (1993) *Assessment of Storm Drain Sources of Contaminants to Santa Monica Bay*, Vol. I, Annual Pollutants Loadings to Santa Monica Bay from Stormwater Runoff, UCLA-ENG-93-62, I, 1-248

Stocker, J. (1998) *Methods for Measuring and Estimating Impervious Surface Coverage*, Nonpoint education for Municipal Officials Technical Paper, No.3

Stuckens, J., Coppin, P. R., and Bauer, M. E. (2000) Integrating contextual information with per-pixel classification for improved land cover classification, *Remote Sensing of Environment*, 71, 282–296

Sucar, L.E. and Gillies, D.F. (1994), Probabilistic reasoning in high-level vision, *Image and Vision Computing*, 12 (1), 42-60

Swain, P.H. (1978) Fundamentals of pattern recognition, in *Remote Sensing: the Quantitative Approach*, edited by Swain, P.H. and Davis, S. M, New York, McGraw-Hill, 136-187

Swamikannu, X., Radulescu, D., Young, R., and Allison, R. (2003) A comparative analysis: storm water pollution policy in California, USA and Victoria, Australia, *Water Science and Technology*, 47 (7-8), 311-317

Thornton, C.J. (1992) *Techniques of Computational Learning: An Introduction*, Chapman and Hill

Tong, S.T.Y., and Chen, W. (2002) Modeling the relationship between land use and surface water quality, *Journal of Environmental Management*, 66, 377-393

US Geological Survey (USGS), http://gisdata.usgs.net/ned

USGS, http://landsat7.usgs.gov/history.html

Van de Merckt, T. (1993) Decision trees in numerical attribute spaces, in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambery, France, 1016-1021

Varis, O. (1995) Belief Networks for Modeling and Assessment of Environmental Change, *Environmetrics*, 6, 439-444

Vaze, J. and Chiew, F.H.S. (2003) Comparative evaluation of urban storm water quality models, *Water Resources Research*, 39(10), 1280-1289.

Wang, F. (1990a) Improving remote sensing image analysis through fuzzy information representation, *Photogrammetric Engineering and Remote Sensing*, 56(8), 1163-1169

Wang, F. (1990b) Fuzzy supervised classification of remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, 28(2), 194-201

Webb, G. and Pazzani, M. (1998) Adjusted probability naïve Bayesian induction, in *Proceedings of 10th Australian Joint Conference on Artificial Intelligence*, Perth, Australia

Widrow, B. and Lehr, M. A. (1990) 30 years of adaptive neural network: perceptron, madaline, and backpropagation, in *Proceedings of IEEE*, 78(9), 1415-1442

Wilson, E.H., Hurd, J.D., Civco, D.L., Prisloe, M.P., and Arnold, C (2003) Development of a geospatial model to quantify, describe and map urban growth. *Remote Sensing of Environment*, 86, 275–285

Wong, K., Strecker, E.W., and Stenstrom, M.K. (1997) A geographic information system to estimate stormwater pollutant mass loadings, *Journal of Environmental Engineering, ASCE*, 123, 737-745

Wu, C., and Murray, A.T. (2003) Estimating impervious surface distribution by spectral mixture analysis, *Remote Sensing of Environment*, 84, 493-505

Yang, L., Huang, C., Homer, C.G., Wylie, B.K., and Coan, M.J. (2003) An approach for mapping large-area impervious surfaces: synergistic use of Landsat-7 ETM+ and high spatial resolution imagery, *Canadian. Journal of Remote Sensing*, 29(2) 230–240

Yang, Y. and Webb, G.I. (2003) *Discretization for Naïve-Bayes Learning: Managing Discretization Bias and Variance*, Technical Report 2003/131, School of Computer Science and Software Engineering, Monash University

Yoshida, T. and Omatu, S. (1994) Neural network approach to land cover mapping, *IEEE Transactions on Geoscience and Remote Sensing*, 32 (5), 1103 –1109

Yuan, Y., Hall, K., and Oldham, C. (2001) A preliminary model for predicting heavy metal contaminant loading from an urban catchment, *The Science of the Total Environment*, 266, 299-307

Zhang, Q. and Stanley, S.J. (1997) Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modeling, *Water Research*, 31(9), 2340-2350

Zhang, Y. (2001) Texture-integrated classification of urban treed areas in high- resolution color-infrared imagery, *Photogrammetric Engineering and Remote Sensing*, 67(12), 1359-1366

Zhu, Z., Yang, L., Stehman, S. V. and Czaplewski, R. L. (2000) Accuracy assessment for the U. S. Geological Survey regional land-cover mapping programme: New York and New Jersey region, *Photogrammetric Engineering and Remote Sensing*, 66, 1425-1435

Zweig, G. (2003) Bayesian network structures and inference techniques for automatic speech recognition, *Computer Speech and Language*, 17(2-3), 173-193