

UNIVERSITY OF CALIFORNIA

Los Angeles

Data Mining of Remote Sensed Data for Stormwater Systems

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Civil Engineering

By

Hsueh-hwa Lee

2003

UMI Number: 3112756

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[®]

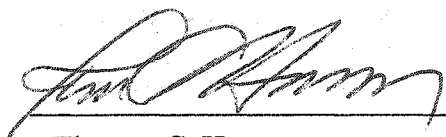
UMI Microform 3112756

Copyright 2004 by ProQuest Information and Learning Company.

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

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

The dissertation of Hsueh-hwa Lee is approved.



Thomas C. Harmon



Keith D. Stolzenbach



Irwin H. Suffet



Michael K. Stenstrom, Committee Chair

University of California, Los Angeles

2003

TABLE OF CONTENTS

LIST OF FIGURES	VII
LIST OF TABLES	XI
ACKNOWLEDGEMENTS	XIII
VITA.....	XIV
ABSTRACT.....	XVI
1. INTRODUCTION	1
1.1 Problem Definition and Significance.....	1
1.2 Objectives and Scope of Research.....	3
1.3 Organization of the Dissertation	4
1.4 References.....	5
2. DATA MINING OF THEMATIC MAPPER IMAGES FOR SANTA MONICA BAY LAND USE CHARACTERIZATION WITH GIS	7
2.1 Introduction.....	7
2.2 Research Scope and Brief Review of the Study Area.....	9
2.3 Space Image Processing.....	10
2.4 Raw Data Processing	11
2.5 Land Use Characterization.....	13
2.6 Results and Discussion	15
2.7 Conclusion and Future Works	16

TABLE OF CONTENTS (Cont'd)

2.8	References.....	16
3.	FUZZY NEURAL NETWORKS AND GIS FOR MULTI-SPECTRAL LAND USE CLASSIFICATION.....	40
3.1	Introduction.....	41
3.2	Methodology.....	42
3.2.1	Description of Raw Data.....	42
3.2.2	Description of the Artificial Neural Network Algorithm.....	42
3.2.3	Input Sensitivity Study.....	45
3.2.4	Fuzzy Logic.....	46
3.3	Results and Discussion.....	47
3.3.1	Network Construction.....	47
3.3.2	Classification Results.....	48
3.3.3	Sensitivity Analysis.....	50
3.4	Conclusions and Future Research.....	51
3.5	References.....	53
4.	NEURAL NETWORK AND GIS TO DETERMINE PIXEL-LEVEL URBAN LAND USE FOR THEMATIC MAPPER IMAGERY.....	77
4.1	Study Area and Land Use Data Processing.....	78
4.2	Methodology.....	79
4.3	Results and Discussion.....	81

TABLE OF CONTENTS (Cont'd)

4.3.1	Network Construction.....	81
4.3.2	Classification Results.....	82
4.4	Conclusions and Future Research.....	83
4.5	References.....	85
5.	STORMWATER RUNOFF SIMULATION IN MALIBU CREEK WATERSHED USING A DETERMINISTIC HYDROLOGICAL MODEL AND ARTIFICIAL NEURAL NETWORKS	103
5.1	Introduction.....	104
5.2	Watershed Description.....	106
5.3	Deterministic Hydrological Modeling.....	106
5.3.1	HEC Geo-HMS Processing.....	107
5.3.2	Meteorologic Model.....	108
5.3.3	Loss Model.....	110
5.3.4	Direct Runoff Model.....	111
5.3.5	Model Calibration	112
5.4	Artificial Neural Network Simulation	114
5.5	Results.....	115
5.5.1	HEC-HMS Simulation.....	115
5.5.2	ANN Simulation	115
5.5.2	HEC-HMS and ANN Comparisons.....	117
5.6	Conclusions.....	117

TABLE OF CONTENTS (Cont'd)

5.7 References.....119

6. CONCLUSIONS..... 154

7. FUTURE WORK..... 157

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
2.1 Santa Monica Bay (SMB) Vicinity Map.....	23
2.2 Thematic Mapper Index.....	24
2.3 Thematic Mapper Raw Image (Band1).....	25
2.4 A Clipped Subset of the Raw Image.....	26
2.5 Formation of Normal-color Multi-band Image.....	27
2.6 Normal-color Composite Image of SMB.....	28
2.7 Infrared Composite Image of SMB.....	29
2.8 Infrared Composite Image of West Los Angeles.....	30
2.9 SMB Land Use Polygon Layer.....	31
2.10 Average DN for Band 1 and Band 2.....	32
2.11 Average DN of Single and Multiple Family Land Use for Band 1 and Band 2.....	33
2.12 Average DN of Single and Multiple Family Land Use for Band 3 and Band 4.....	34
2.13 Average DN of Single and Multiple Family Land Use on Band 1, 2, and 3.....	35
2.14 Average DN of Single and Multiple Family Land Use on Band 2, 3, and 4.....	36
2.15 Average DN of Commercial and Open Land Use for Band 1 and Band 2.....	37
2.16 Average DN of Commercial and Open Land Use for Band 3 and Band 4.....	38

LIST OF FIGURES (Cont'd)

Figure	Page
2.17 Average DNs of Commercial and Open Space Land Use for Band 2, 3 and Band 4	39
3.1 Santa Monica Bay USGS Digital Elevation Model	62
3.2 Santa Monica Bay Slope Distribution.....	63
3.3 MLP of One Hidden Layer with D Inputs and M Outputs	64
3.4 Triangular Fuzzy Function.....	64
3.5 Fuzzy Sets for Land Use Categories Based on Average Pixel Value.....	65
3.6 Derivation of Membership from Single Input Through Fuzzifier Function	68
3.7 A Neural Network with Fuzzy Pre-processor	69
3.8 MLP with Back Propagation Learning	69
3.9 Neural Network Input Data Type.....	70
3.10 Network Outputs with Varied Bands	71
4.1 Marina Del Rey and Santa Monica Bay.....	91
4.2 Land Use Polygons	92
4.3 Land Use Pixels	93
4.4 Homogeneous Land Use Pixels	94
4.5 Study Area and Seven Thematic Mapper Spectral Bands	95
4.6 Multiple Layer Perceptron Network	96
4.7 Kohonen Self Organization Map Network	96
4.8 MLP Networks Input Data Type.....	97

LIST OF FIGURES (Cont'd)

Figure	Page
4.9 ANN Predicted Outputs vs. Desired Results for P-MLP3 Training Data.....	98
4.10 SOM-1 Network: Four Clusters.....	99
4.11 SOM-2 Network: Seven Clusters.....	100
4.12 SOM-3 Network: Nine Clusters.....	101
4.13 Aerial Photo and SOM-1 Clustering.....	102
5.1 Malibu Creek Watershed Vicinity Map.....	129
5.2 HEC-HMS Schematics	130
5.3 USGS 10-meter Digital Elevation Models.....	131
5.4 HEC-GeoHMS Processing.....	132
5.5 Sub-watersheds after Delineation	133
5.6 Precipitation Gauges and Sub-watershed Centroids	134
5.7 Inverse-distance-square Method to Calculate Sub-watershed Hytograph	135
5.8 NRCS STATSGO Soil Maps with Map Unit ID (MUID).....	136
5.9 NRCS STATSGO Soil Loss Rate Surface.....	137
5.10 Surface of LADPW Imperviousness Ratio	138
5.11 Surface of Distance to Longest Flow Path.....	139
5.12 Surface of Percentage Slope (V:H).....	140
5.13 Surface of Roughness Coefficient.....	141
5.14 LADPW Stream Gauging Station F-130R.....	142

LIST OF FIGURES (Cont'd)

<u>Figure</u>	<u>Page</u>
5.15 Daily Discharge Points Used in ANN Simulation.....	143
5.16 Hydrograph of HEC-HMS Simulated, Calibrated, and Observed Discharges.....	144
5.17 Hydrograph of HEC-HMS Simulated, Calibrated, and Observed Discharges for 1998 Wet Season	145
5.18 Residuals of Calibrated and Observed Discharges for 1998 Wet Season	146
5.19 Network Simulated and Observed Discharges	147
5.20 Problematic Discharge Gauge Readings During Nov. 15 to Dec. 9, 1997.....	150
5.21 Hydrograph of HEC-HMS Calibrated, Q-MLP4, and Observed Discharges.....	151

LIST OF TABLES

Table	Page
2.1 SCAG Land-Use Characteristics	20
2.2 Description of the Thematic Mapper Image	20
2.3 Thematic Mapper Band/Color Combinations and Potential Applications	21
2.4 Summary of the Santa Monica Bay Watershed Land Use Coverage	21
2.5 Global Average Digital Numbers (DNs) of All Seven Spectral Bands for SCAG Land Use Patterns	22
3.1 Summary of Input Data Used	57
3.2 Lists of Networks Constructed.....	57
3.3 Network Mean Square Errors	57
3.4 Testing Results.....	58
3.5 Correlation	59
3.6 Confusion Matrix	60
3.7 Sensitivity of Input Parameters.....	61
4.1 Summary of Land Use Polygons	86
4.2 Summary of Land Use Pixels	86
4.3 Summary of Input Data Used	87
4.4 Lists of MLP Networks.....	87
4.5 Lists of SOM Networks	87
4.6 MLP Network Mean Square Errors	88
4.7 Testing Results.....	88

LIST OF TABLES (Cont'd)

Table	Page
4.8 Correlation	89
4.9 Confusion Matrix	90
5.1 Major Parameters Used in HEC-HMS.....	122
5.2 Major GeoHMS Parameters of Sub-watersheds	122
5.3 Los Angeles County Department of Public Works Precipitation Gauges	123
5.4 Distance (ft)-matrix between Precipitation Gauges and Centroids of Sub-watersheds.....	123
5.5 SCS Soil Groups and Corresponding Loss Rates	124
5.6 Overland-flow Roughness Coefficients	124
5.7 Overland-flow Roughness and SCAG Land Use Correlation Table	125
5.8 MLP Networks Used in This Study	127
5.9 Residuals of the Seasonal Runoff Volume and Peak Discharge.....	127
5.10 Optimized Scale Factors and Values of Objective Function	127
5.11 ANN Simulation Results.....	128
5.12 Sensitivity Analysis of Q-MLP4 Network.....	128

ACKNOWLEDGEMENTS

I am most grateful to have had the love and support of my family, especially my parents, and my wife Lucy, during the course of my research.

I would like to express my deep appreciation to my advisor, Professor Michael K. Stenstrom, for his inspiration and encouragement throughout my graduate study at UCLA. Special thanks go to Dr. Walter Karplus, my former committee member and Ph.D. minor field advisor in Scientific Computing. Dr. Karplus passed away before my term of completion, but his inspiration, specifically in the field of mathematical modeling, will always be remembered. I also wish to acknowledge the other members of my committee, Dr. Thomas Harmon, Dr. Keith Stolzenbach, and Dr. Irwin Suffet, for their valuable assistance and suggestions.

Others who contributed directly to my research include Professor Jeff Kuo, who led me all the way to the Environmental Engineering discipline, and Megan Miller for her editing direction.

VITA

September 29, 1970	Born, Taipei, Taiwan
1993	B.S., Civil Engineering National Taiwan University, Taipei
1996	M.S., Civil Engineering University of California, Los Angeles
1993 – 1995	Platoon Leader, Second Lieutenant Kimen Combat Engineer Battalion Taiwanese Army Corps of Engineers
1995 – 1999	Research and Teaching Assistant Department of Civil Engineering University of California, Los Angeles
1999 – 2001	Water Resources Engineer Fusco Engineering, Inc., Irvine, CA
2001 – 2003	Environmental Planner Planning and Property Management Division County Sanitation Districts of Los Angeles County Whittier, CA

PUBLICATIONS AND PRESENTATIONS

- Lee, H. H. and Stenstrom, Michael. K. (2003), Data Mining of Remote Sensed Data for Stormwater Modeling with GIS, Submitted to Urban Water Journal
- Lee, H. H. and Stenstrom, Michael. K. (2003), Stormwater Runoff Simulation in Malibu Creek Watershed Using a Deterministic Hydrological Model and Artificial Neural Networks, Submitted to Journal of Water Resources Research.
- Lee, H. H., Guttman G., and S. Highter (2003), Using GIS for Wastewater Reclamation Site Suitability Study, Submitted to 2003 ESRI International User Conference.

Stenstrom, M. K., H. H. Lee, and J. S. Ma (1998), Mathematical Modeling and Computer Simulation, Dynamic and Control of Waste Water Systems, vol.6, Technomic Publishing Co. Inc.

Stenstrom, M. K., K. M. Wong, J. S. Ma and H. H. Lee "The Impact of Land Use on Stormwater Quality in a Large Urban Watershed," US EPA EMAP Meeting, April 8, 1999, San Francisco, CA

Lee, H. H., J. S. Ma and M, K, Stenstrom "Calibration of a GIS-Based Empirical Urban Runoff Model on Ballona Creek Watershed", Proc. Of AGU Meeting, May 26, 1998, Boston, MA.

ABSTRACT OF THE DISSERTATION

Data Mining of Remote Sensed Data for Stormwater Systems

By

Hsueh-hwa Lee

Doctor of Philosophy in Civil Engineering

University of California, Los Angeles, 2003

Professor Michael K. Stenstrom, Chair

The main objective of this research was to enhance understanding of the Santa Monica Bay area stormwater systems. The runoff system is proven too intricate for conventional approaches, as most of its parameters are complex and spatially distributed. A novel three-part approach was developed to redress problems that had long prevented stormwater research from breaking new ground.

Although remote sensed data have been available for many years, no stormwater model has used them as direct model input. This dissertation proposes establishing a geographic information system (GIS) with remote sensed data, using data mining techniques to explore land use information for stormwater modeling, and performing hydrological analysis on one of Santa Monica Bay's discrete watersheds. The unique technological combination of the Geographic Information System (GIS) and artificial

neural network (ANN) algorithms is the approach that stormwater research has been waiting for.

The study focused on the geographical characteristics of the Santa Monica Bay stormwater systems, but the method can be extrapolated to other watersheds.

1. INTRODUCTION

1.1 Problem Definition and Significance

Stormwater runoff has historically been one of the most complex and consequently one of the least understood environmental systems under scrutiny by environmental researchers. Numerous factors such as geology (surface elevation, slope, etc.), precipitation characteristics, and human activities (land use pattern, land surface improvement, and drainage networks) affect runoff quality and quantity, and standardized measurements are difficult both to achieve and to evaluate [Corbit, 1989]. Preparing effective yet efficient runoff models requires technologies that have only become widely available in the last decade; innovations like the data mining technology, geographic information system (GIS), and others have streamlined previously laborious processes, making them more accessible and feasible to researchers. This is critical to the continuing efforts and success of controlling non-point source pollution.

Stormwater runoff models vary in their complexity and data, personnel and computational requirements [Charbeneau and Barrett, 1998]. Obviously, technology that can facilitate any or all of these is invaluable to the field as a whole. Among all the required parameters, land use information is the most essential to monitor runoff quality and quantity. Based on land use patterns, stormwater management models define imperviousness and event mean concentrations of pollutants. In recent decades, land use mapping has been enhanced by remote sensed satellite images, which are now a standard source of planning, tracing, and classifying land use patterns. However, again, the

computing resources required for processing and analyzing these images have long posed a challenge.

Advances in computing hardware and software and the innovation of the data mining technology of the last two decades have provided a solution to this hardship. In particular, data mining using artificial neural networks with the addition of fuzzy logic has been very effective. Data mining uses an efficient way to discover new, valuable, and non-obvious information from a huge collection of data [Bigus, 1996; Mesrobian et al., 1996]. With digital databases of remote sensed data now more affordable and accessible, data mining techniques have become a popular cost-effective way to extract precious land information.

An *Artificial Neural Network* (ANN) is a computational structure inspired by human biological neural processing [Rao, 1995]. ANN simulation has gained enormous attention as a technique for classifying remote sensed data and has proved to be one of the key algorithms used in data mining. Whereas ANNs deal with learning and probabilistic reasoning, fuzzy logic is concerned with imprecision and uncertainty [Zadeh, 1994]. Neural Fuzzy (NeuroFuzzy) networks are networks that use fuzzy logic in the processing units or in the connection weight representations [Okada et al., 1992]. The integration of ANN and fuzzy logic produces a system that is greater than the sum of its parts—a synergy that overcomes the weaknesses of each technology and thus results in a holistic approach to data mining.

Most established database management systems were originally developed to manipulate tabular-format numeric data sets. However, nearly all the information

required for stormwater modeling is distributed spatially. GIS technology now provides for the conversion of complex geographic reality into a finite number of database records or objects, which includes lines, points, or areas, and also possesses descriptive attributes [Goodchild, 1993]. GIS therefore complements remote sensing techniques as a framework for integrated spatial analysis and helps develop and tailor remote sensed data sets. The complex spatial data requirement forges vital links among GIS, remote sensed data, and stormwater management models. More specifically, GIS could be defined as a computer hardware and software system designed for the storage and processing of geographic data in both geographic and analytical forms [Sanchez and Canton, 1999]. A GIS application typically requires extensive geographic, cartographic, engineering, and statistical knowledge and experience as well as considerable analytical skills.

1.2 Objectives and Scope of Research

California's entire Santa Monica Bay (SMB) area, and in particular two of its discrete watersheds, the Ballona Creek Watershed and the Malibu Creek Watershed, were selected as the study area for this project. Extensive development and population growth in this area have led to severe degradation of the surface water quality. It is important to explore the causes of the damage in order to restore the Bay to its pristine state.

The goal of this study is to develop a GIS integrated with data mining system that is powered by artificial neural network algorithms, modern database servers, and conventional surface hydrological models. The GIS based data mining system contains more than 80 geographic layers or images and occupies 40 Gigabytes (GB) of disk

storage. It will glean data necessary for understanding and depicting the stormwater runoff parameters and mechanisms of SMB. The methodology used in and the data derived from this research project will contribute to a better understanding of the stormwater system in the SMB watershed, and thus will be useful in developing a better monitoring program for runoff pollution control.

The primary objectives of this research are:

1. To establish a centralized GIS containing remote sensed images and other GIS layers for the entire SMB watershed.
2. To preprocess NeuroFuzzy model input parameters with GIS and present statistical observations between the SMB land use patterns and satellite imagery.
3. To perform polygon level land use classifications using NeuroFuzzy models for SMB.
4. To perform pixel level land use classification using supervised and unsupervised learning algorithms near the Ballona Wetland and the surrounding areas.
5. To build a GIS to support U.S. Army Corps of Engineer HEC-HMS hydrological modeling program parameters, to perform HEC-HMS stormwater quantity modeling, and to use the ANN simulation as an alternative for imitating the model output in the Malibu Creek Watershed.

1.3 Organization of the Dissertation

This dissertation is composed of four research manuscripts ready to be submitted for publication. Each paper addresses one or more objectives listed in Section 1.2:

Chapter 2 describes the GIS techniques used for satellite image processing and presents graphs of statistical observations. Chapter 3 introduces a NeuroFuzzy network to perform polygon-level land use classification. Pixel-level land use classification using both supervised and unsupervised learning algorithms is proposed in Chapter 4. In Chapter 5, a GIS integrated with both a traditional hydrological model and ANN is built to perform watershed analysis. Chapter 6 makes conclusions about this study and Chapter 7 describes the future works.

1.4 References

- Bigus, Joseph. (1996). *Data Mining with Neural Networks*. McGraw Hill, Inc., New York, NY.
- Charbeneau, R. J. and Barrett M. E. (1998). "Evaluation of Methods for Estimating Stormwater Pollutant Loads", *Water Environment Research*. Vol. 70, No. 7, pp. 1295-1302.
- Corbit, R.A. (1989). *Standard Handbook of Environmental Engineering*. McGraw Hill, Inc., New York, NY.
- Goodchild, M. F. (1993). "The State of GIS for Environmental Problem-Solving". At *Environmental Modeling with GIS*. Oxford University Press, New York, NY.
- Mesrobian, E, Muntz, R, Shek, E, Nittel, S and et al (1996). "Mining Geophysical Data for Knowledge". *IEEE Expert* 0885-9000/96. Vol. 11, No. 5, October 1996. pp.34-44.

- Okada, H., Watanabe, N., Kawamura, A., Asakawa, K., Taira, T., Isida, K., Kaji, T., and Narita, M. (1992). "Initializing multilayer neural networks with fuzzy logic", *Proceedings of the IEEE International Conference on Neural Networks*, Vol. 1, pp. 239 – 250.
- Rao, V. B. and Rao, H. V. (1995). *C++ Neural Networks & Fuzzy Logic*. MIS Press. New York, NY.
- Sanchez, Julio, and Canton, M. P. (1999). *Space Image Processing*. CRC Press LLC. Boca Raton, FL.
- Zadeh, L. A. (1994). "Fuzzy logic, neural networks, and soft computing" in *Communications of the ACM*, 3, pp. 77 – 84.

2. Data Mining of Thematic Mapper Images for Santa Monica Bay

Land Use Characterization with GIS

H H. Lee, Michael K. Stenstrom, Jiun-shiu Ma, and Kenneth M. Wong

Abstract

This paper was developed mainly for quantity and quality research of stormwater runoff in the Santa Monica Bay Watershed. It suggests the integration of remotely sensed data and the land use layer with a geographic information system to redress the potentials of using these data to classify complex urban land use patterns (specifically, the challenges of the numerical characterization of spectral signatures and the additional problem of selecting an appropriate methodology). Using theories of spectral signatures and GIS spatial analyses, Digital Numbers of each spectral band were extracted from the Landsat Thematic Mapper images. The relationships of land use patterns and spectral signatures suggested the potential for multi-band spectral signatures to yield significant information of land use characterization. This paper focuses on the methodologies used to collect and extract information, and presents the first-stage observations of a multi-phase research project.

2.1 Introduction

The terms *land use* and *land covers* are used to describe different surface features of landscape. Land use refers to the economic and functional roles of land surface in human activities [Sanchez and Canton, 1999]. Land cover is an integrated expression of

the physical, climatic, and biotic environment as well as of the history of human land use [Gopal et al., 1999]. Land use focuses on the socioeconomic issues while land cover refers to the physical components. The data from both are not only useful for stormwater modeling, but are also influential over the local, state, and federal policy-making.

Land use data is essential for monitoring the quality and quantity of stormwater runoff. As such, it is required for almost all stormwater runoff models. Traditionally, to determine the land use patterns, various techniques of mapping are applied to a designated area. In the last two decades, the field of land use mapping was greatly enhanced by the innovation of satellite imagery. Since then, these images and the extraordinary detail that they depict have been integral to the planning, tracing, and classifying of land use patterns. The introduction of data mining techniques in the 1990's likewise has allowed for efficient discoveries of valuable and non-obvious information from a large body of data [Bigus, 1996; Mesrobian et al., 1996].

The consensus among environmental researchers is that a combination of techniques is the best way to extract information from the satellite imagery just mentioned, as well as other spatial data: data mining integrated with a geographic information system (GIS) covers all areas fundamental to information discovery. Typical of economic constraints, a data mining operation is only successful when the value of the extracted information exceeds the cost of processing the raw data. The methodology and example presented in this paper follow this logic, and therefore provide a unique cost-effective approach to the process of eliciting information from remotely sensed data.

2.2 Research Scope and Brief Review of the Study Area

The Santa Monica Bay (SMB) Watershed in the State of California was selected as the study area because of its huge socioeconomic impact and significant geographical location. The SMB (Figure 1) Watershed serves approximately nine million people, the largest population in the West Coast of the United States—as well as its commercial center. It extends from the shoreline of Los Angeles and the adjacent cities of Santa Monica. Factors affecting the quality and quantity of stormwater runoff in the SMB area are complicated and not fully understood. The goal of this study is to:

1. Establish a GIS for the SMB Watershed with satellite images and other geographic data.
2. Perform GIS spatial analyses and observe the relationships between land use patterns and satellite images.
3. Provide a better characterization of land use in the SMB Watershed and possibilities for future studies of more complicated classification methods.

The land use data used here came from Southern California Association of Governments [AIS, 1996]. A slightly modified definition of the Anderson Land Use Classification [Anderson et al., 1976] was used. Anderson's theory of Level II classification represents the land use categories by two digital codes, and is utilized to describe the prominent geographic attributes of the study area.

The land use patterns play a key role in stormwater runoff quantity and quality by affecting the imperviousness and the specific pollutant's Event Mean Concentration

(EMC). Previous studies have used land use polygons to estimate nonpoint source pollutant loadings of urban watershed [Wong et al., 1997] and sewershed [Ventura et al., 1993]. The runoff coefficient, RV , is defined as the ratio of overall average runoff to rainfall, and is greatly correlated to the imperviousness (IMP) of surface area [Discoll et al., 1990]. The relationship between RV and IMP is presented in the following equation:

$$RV = 0.007 IMP + 0.1 \quad (\text{Eq. 1})$$

Based on the SMB drainage area characteristics from the Los Angeles County Department of Public Works, CA, the values of IMP and calculated RV for each of the land use pattern are described in Table 1.

2.3 Space Image Processing

Remote sensing is based on radiation emitted and reflected from the Earth's surface. Instruments of remote sensing measure the relative brightness of objects over a range of wavelengths. The *spectral signature* of an object can be defined as the result of a set of observations accounting the variations in the response of the same object and the difference with other objects [Sanchez and Canton, 1999]. In practice, spectral signatures have helped identify various features of the Earth's surface, such as crops, forests, minerals, and land use patterns. Due to the temporal and spatial variations of the spectral signatures of various objects, the identification process can be quite complicated. Objects may also present different signatures under different temperatures. Even with great

variations, investigation of spectral properties is an indispensable component of remote sensing. While unique signatures of specific objects may not be easily defined, the associated spectral patterns of the objects can always be discovered.

Since the late 1960's, the National Aeronautics and Space Administration (NASA) has developed and launched its Earth-monitoring satellites (Landsat). Landsat images with Thematic Mapper (TM) as sensors are used in this study. Landsat TM supports an on-board analog-to-digital conversion with a value range from 0 to 255 (8 bits). This value is referred to as *digital number* (DN) in this study. A TM *scene* is composed of seven spectral bands; the usage of the DNs of seven bands to define spectral signatures can greatly enhance the accuracy, since "significant differences or similarities may well remain hidden if the variables are considered one at a time and not simultaneously" [Mather, 1976].

Landsat data used for land characterization include images generated by the Landsat Multi-spectral Scanner Sensors (MSS) [Benediktsson et al., 1990; Lee et al., 1990], Landsat TM [Hepner et al., 1990; Civco, 1993; Yoshida and Omatu, 1994; Moody et al., 1996], Advanced Very High Resolution Radiometer (AVHRR) [Gopal et al., 1994; 1999], and other data sets. Both statistical and soft-computing approaches have been applied to land use classification.

2.4 Raw Data Processing

Only one Landsat TM scene (Figure 2, USGS scene no. 5237717480), composed of seven spectral bands, was used in this study (Features described in Table 2). Each of

the seven bands of the raw TM data has a file size of approximately 75–80 MB (megabytes) in BIP (Band Interleaved by Pixel) format.

The TM sensor has a spatial resolution of 30 meters for bands 1 through 5, and band 7, and a spatial resolution of 120 meters for band 6. However, it should be noted that the TM scene used in this study has been post-processed with “edge-enhancement” technique and has a spatial resolution of 25 meters for all spectral bands. The map projection of the raw data is in Universal Transverse Mercator (UTM) system with map units in meters. In UTM, the globe is divided into six zones, each spanning six degrees of longitude with its own central median [ESRI, 1994]. Figure 3 presents the spectral Band 1 in gray scale. The red line overlaying the raw image is the boundary of the SMB Watershed. A subset of image covering study area (Figure 4) was clipped from the raw data in order to save the storage space and processing time.

Because a TM scene is composed of seven spectral bands and most video display systems only have three-color schemes (blue, green, and red), any combination of three bands can form a composite image and be displayed on a video system. Table 3 presents band combinations and their corresponding potential applications [Sanchez and Canton, 1999].

Forming multi-band composite images demands the complex task of image processing. First, the raw TM image for each band is clipped and converted to GIS grid format. Then any three grids of bands can be converted to an image file. Figure 5 schematically illustrates the process of creating a normal-color image by superimposing three spectral bands.

The normal multi-band composite image is created by coloring Band 1 blue, Band 2 green, and Band 3 red (Figure 6). The infrared composite image can be generated by coloring Band 2 blue, Band 3 green, and Band 4 red (Figure 7). Figure 8 shows a closer look at the west portion of Los Angeles on the infrared composite image. Each of these images contains 4,158,720 pixels (1,520 rows \times 2,736 columns). In reality, each pixel of a TM scene is 25 meters wide and 25 meters long and contains a unique combination of 7 digital numbers from 7 various spectral bands. It is assumed that the information of spectral signatures for land use pattern characterization is hidden within these digital numbers.

2.5 Land Use Characterization

There are 5,241 land use polygons defined in the study area; Figure 9 shows the GIS vector-based polygon coverage. A summary of land use polygons in Santa Monica Bay is listed in Table 4. The global average DN of each land use pattern in the entire study area can be calculated by the following equation.

$$DN_i = \frac{\sum_{j=1}^{N_i} dn_j}{N_i} \quad (\text{Eq. 2})$$

where DN_i = global average DN for land use pattern i

dn_j = DN of a pixel j in the domain of land use pattern i

N_i = total number of pixels of land use pattern i

To calculate the average DN, it is necessary to first overlay two GIS sources: the SMB land use coverage (5,244 vector polygons) in vector format and the TM scene (4,158,720 pixels) in raster grid format, and then perform spatial analyses. Several GIS zonal functions were used to calculate the global average DN values of each spectral band for each land use pattern in the whole study area.

The next task is to characterize each land use polygon with DN. The average DN of a specific spectral band for a single land use polygon can be calculated by the following equation.

$$DN_{kp} = \frac{\sum_{j=1}^{N_p} dn_{kj}}{N_p} \quad (\text{Eq. 3})$$

where DN_{kp} = average DN of spectral band k in polygon p

dn_{kj} = DN of spectral band k for pixel j (j is inside polygon p)

N_p = total number of pixels inside polygon p

Again, the average DN of a specific spectral band for a specific polygon can be calculated by overlaying each land use polygon spatially with the corresponding TM bands by using GIS zonal functions.

2.6 Results and Discussion

GIS zonal functions (Eq. 3) were used to calculate the global average DN values of each spectral band for each land use pattern in the whole study area. The results are presented in a 7 x 7 matrix (Table 5).

After performing the GIS spatial analysis, the average DNs of each TM spectral band (seven total) of each land use polygon (5,244 total) were calculated and stored in a 7 x 5,244 matrix. Figure 10 plots the average DNs (Band 1 and Band 2) of random-sampled land use polygons among all polygons; the distributions of points for the different land uses at this point are almost indistinguishable. However, when the number of land uses is reduced from seven to two (Single and Multi-family), two distinct land use patterns start to emerge in Figure 11. They become even more separable when the average DNs of Band 3 and Band 4 are used in Figure 12.

When the average DNs of Bands 1, 2, and 3 for Single and Multi-family land uses are three-dimensionally plotted on Figure 13, the patterns themselves become more distinguishable. Similar to the increased detail between Figures 11 and 12, Figure 14 shows what happens when the average DNs of Band 2, 3, and 4 are plotted.

Again, Figure 15 plots the average DNs of Commercial and Open land uses on Band 1 and Band 2, and Figure 16 plots the average DNs on Band 3 and Band 4. Based on this study, Commercial and Open land uses are made more discernible by using the average DNs of Band 3 and Band 4. And, Commercial and Open land uses are even most differentiable when plotted three-dimensionally with the average DNs of Band 2, 3, and 4 as in Figure 17.

2.7 Conclusion and Future Works

The results achieved by this study provide a valuable database of information for future endeavors in the area. The integrated GIS is able to depict the land features of a watershed numerically, and the geographically related spectral signatures provide a great resource for other environmental researches in the Santa Monica Bay Watershed. Specifically, the usage of TM spectrum signatures facilitates the differentiation of land use patterns, and a combination of two or more spectral bands offers the potential to define spectrally distinct land use patterns.

The methods used to elicit these findings could become a new model for the field of stormwater data mining in general. The numerical matrix of spectral signatures corresponding to each land use polygon is established in a format readily accessible to other statistical and soft computing based applications. Perhaps most significantly, the methodology utilized in this research, such as numerical spectral signature extraction from TM images and spatial statistical analysis performed by GIS, can be applied as a blueprint to the study of other watersheds.

2.8 References

Aerial Information Systems (1996). *Southern California 1990 Aerial Land Use Study: Land Use Level III/IV Classification*. Aerial Information Systems, Redlands, California.

- Anderson, J. R., Hardy, E. T., and Witmer, R. E. (1976). "A land use and land cover classification system for use with remote sensor data", U.S. Geological Survey Professional Paper 964.
- Benediktsson, J. A., Swain, P. H., and Ersoy, O. K. (1990). "Neural Network Approaches versus Statistical Methods in Classification of Multisource Remote Sensing Data." *IEEE Trans. Geosci. Remote Sens.* Vol. 28, pp. 540-552.
- Bigus, Joseph. (1996). *Data Mining with Neural Networks*. McGraw Hill, Inc., New York, NY.
- Civco, D. L. (1993). "Artificial Neural Networks for Land-cover and Mapping." *International Journal of Geographical Information Systems*. Vol. 7, pp. 173 – 186.
- Driscoll, E.D., Shelley, P.E., and E.W. Strecker (1990). *Pollutant Loadings and Impacts from Stormwater Runoff, Volume III: Analytical Investigation and Research Report*. FHWA-RD-88-008, Federal Highway Administration, Washington, DC.
- ESRI (1994). *GIS by ESRI: Map Projections*. Environmental Systems Research Institute, Inc. Redlands, CA.
- Gopal, Sucharita, Woodcock, C. E., and Strahler, A. H. (1999). "Fuzzy Neural Network Classification of Global Land Cover from a 10 AVHRR Data Set", *Remote Sens. Environ.* Vol. 67, pp. 230-243.
- Gopal, S., Sklarew, D. M. and Lambin, E. (1994). "Fuzz-neural Networks in Multi-temporal Classification of Land-cover Change in the Sahel." In *Proceedings of the DOSES Workshop on New Tools for Spatial Analysis*, Lisbon, Portugal, DOSES, EUROSTAT, ECSC-EC-EAEC, Brussels, Luxembourg, pp. 55 - 68.

- 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." *Photogramm. Eng. Remote Sens.* Vol. 56, pp. 469 – 473.
- Lee, J., Weger, W. C., Sengupta, S. K., and Welch, R. M. (1990). "A Neural Network Approach to Cloud Classification." *IEEE Trans. Geosci. Remote Sens.* Vol. 28, pp. 846 – 855.
- Mather, P. M (1976). *Computational Methods of Multivariate Analysis in Physical Geography*. Wiley, Chichester.
- Mesrobian, E, Muntz, R, Shek, E, Nittel, S and et al (1996). "Mining Geophysical Data for Knowledge". *IEEE Expert* 0885-9000/96. Vol. 11, No. 5, October 1996. pp.34-44.
- Moody, A., Gopal, S. and Strahler, A. H. (1996). "Artificial Neural Network Response to Mixed Pixels in Coarse-Resolution Satellite Data." *IEEE Trans. Geosci. Remote Sens.* Vol. 58, pp. 329 – 343.
- Sanchez, Julio, and Canton, M. P. (1999). *Space Image Processing*. CRC Press LLC. Boca Raton, FL.
- Ventura, S. J., and Kim, K. (1993). "Modeling urban nonpoint source pollution with a geographic information system." *Water Resource Bull.*, Vol. 28, No. 5, pp. 189 – 198.

Wong, K. M., Strecker, E. W., and Stenstrom, M. K. (1997). "GIS to estimate storm-water pollution mass loadings". *J. of Environ. Eng.*, Vol. 123, No. 8, pp. 737 – 745.

Yoshida, T., and Omatu, S. (1994). " Neural Network Applications to Land-cover Mapping." *IEEE Trans. Geosci. Remote Sens.* Vol. 32, pp. 1103 – 1109.

Table 1 SCAG Land-Use Characteristics [Wong et al., 1997]

Land Use Pattern	Land Use Code	Impervious Surface Area [%]	Runoff Coefficient
Single-family	11	42	0.39
Multi-family	12	68	0.58
Commercial	20	92	0.74
Public	30	80	0.66
Light Industrial	40	91	0.74
Other Urban	50	80	0.66
Open	60	0	0.10

Table 2 Description of the Thematic Mapper Image

Projection Layer:	UTM 10/11
Source:	US Geological Survey/UC Santa Barbara
Capture Method:	LANDSAT on-board multi-spectral sensor
Data Format:	Geo-referenced bip.(raw binary image)
Resolution:	25 meters
File Size:	Approximately 75 – 80 MB per band per scene
Image Size	8540 (rows) × 9110 (columns) 213.5 km (width) × 227.75 km (length)
Data Updated:	Data acquired in 1990

Table 3 Thematic Mapper Band/Color Combinations and Potential Applications

Blue	Green	Red	Composite	Possible Applications
1	2	3	Normal color	Water sediment pattern
2	3	4	Infrared	Urban features recognition
3	4	5	False color	N/A*
3	4	7	False color	N/A*
3	5	7	False color	Vegetation enhancement
4	5	7	False color	N/A*
1	4	7	False color	N/A*

* Not Available

Table 4 Summary of the Santa Monica Bay Watershed Land Use Coverage

Land Use Pattern	Land Use Code	Total Number of Polygons
Single-family	11	1,297
Multi-family	12	1,062
Commercial	20	981
Public	30	602
Light Industrial	40	151
Other Urban	50	271
Open	60	877

Table 5 Global Average Digital Numbers (DNs) of All Seven Spectral Bands for SCAG Land Use Patterns

Land Use Pattern	Code	Global Average DN						
		Band1	Band2	Band3	Band4	Band5	Band6	Band7
Single-family	11	100	45	57	68	84	162	162
Multi-family	12	112	50	65	61	79	167	167
Commercial	20	116	52	67	54	75	169	169
Public	30	109	49	64	67	87	164	164
Light Industrial	40	120	55	72	59	90	170	170
Other Urban	50	116	55	75	65	106	168	168
Open	60	83	35	45	58	94	164	164

Figure 1
Santa Monica Bay (SMB) Vicinity Map

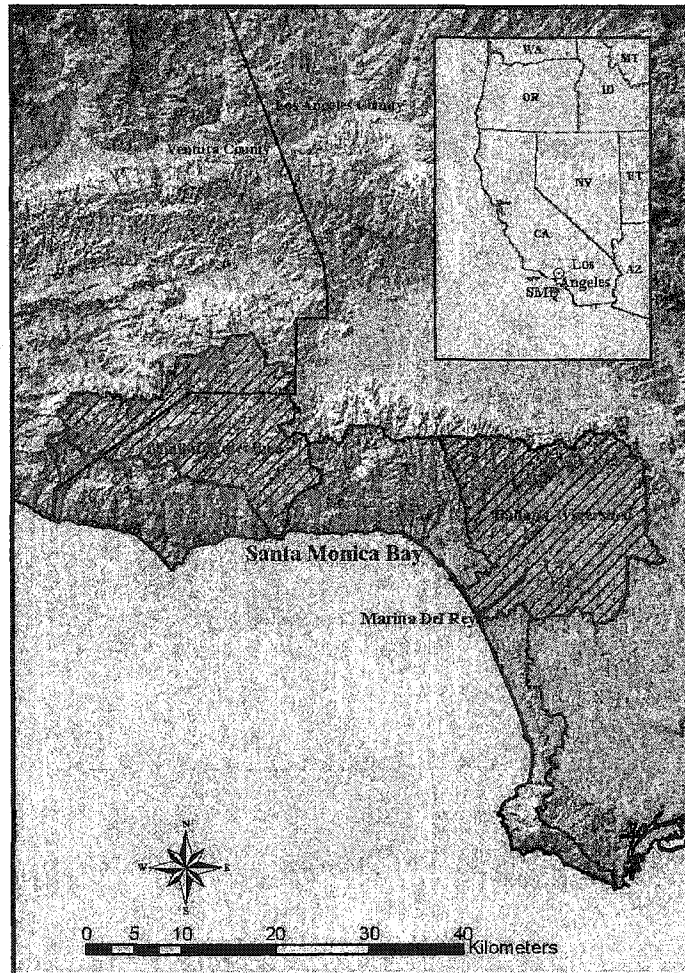


Figure 2
Thematic Mapper Index

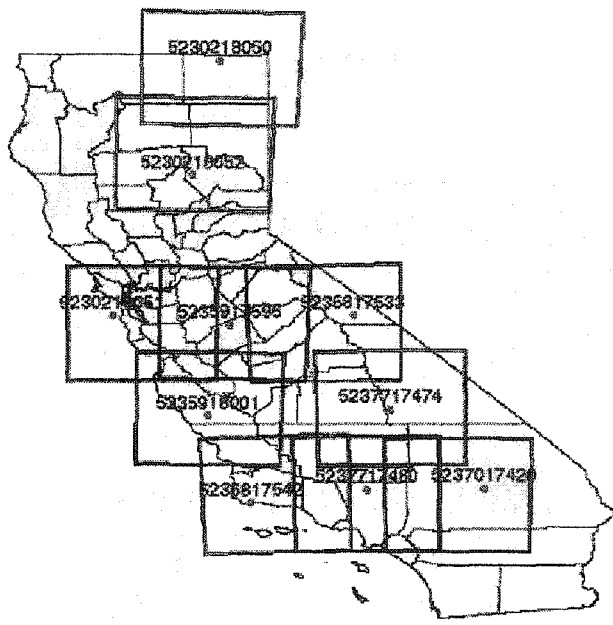
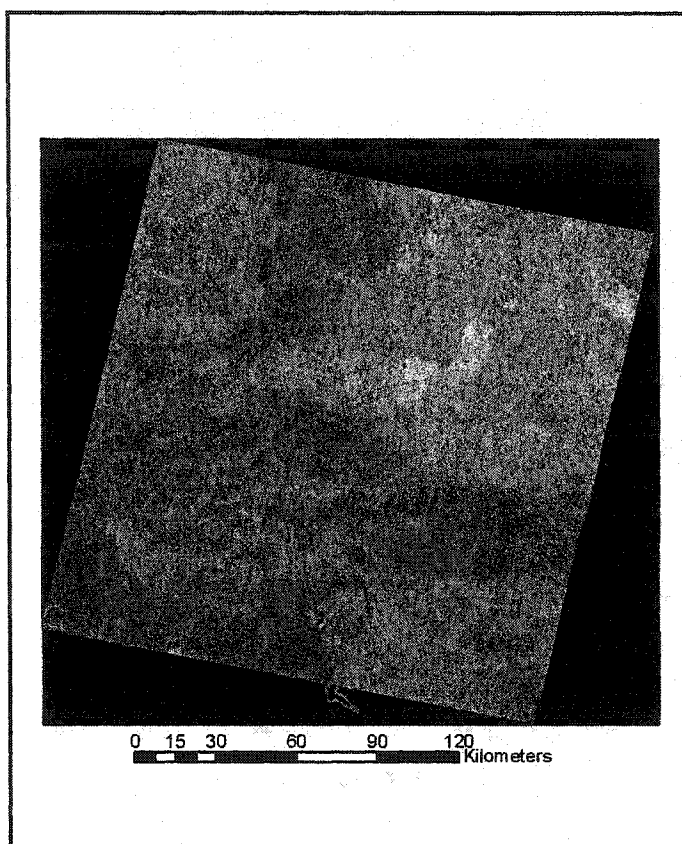


Figure 3
Thematic Mapper Raw Image (Band 1)



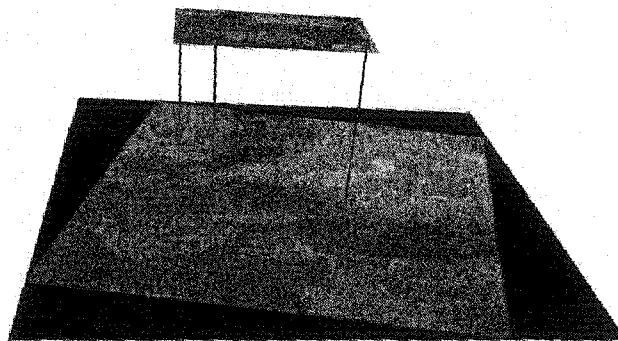


Figure 4
A Clipped Subset of the Raw Image

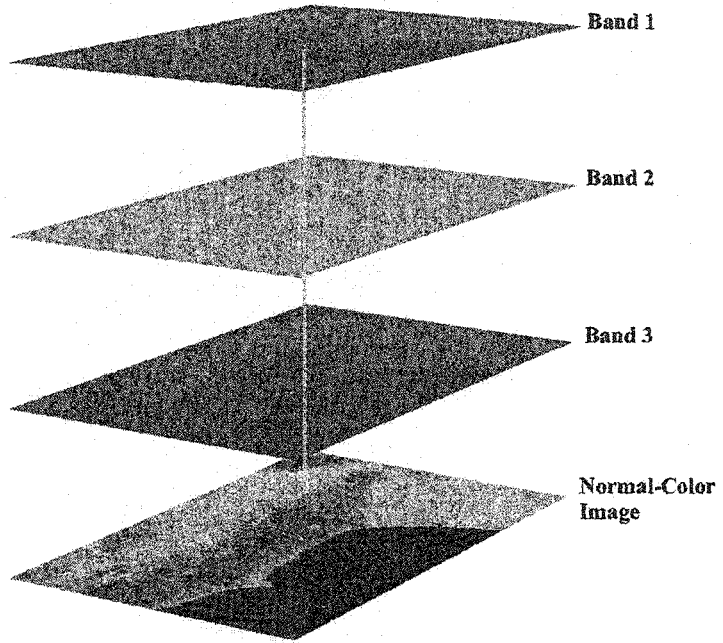


Figure 5
Formation of Normal-color Multi-band Image

Figure 6
Normal-color Composite Image of SMB

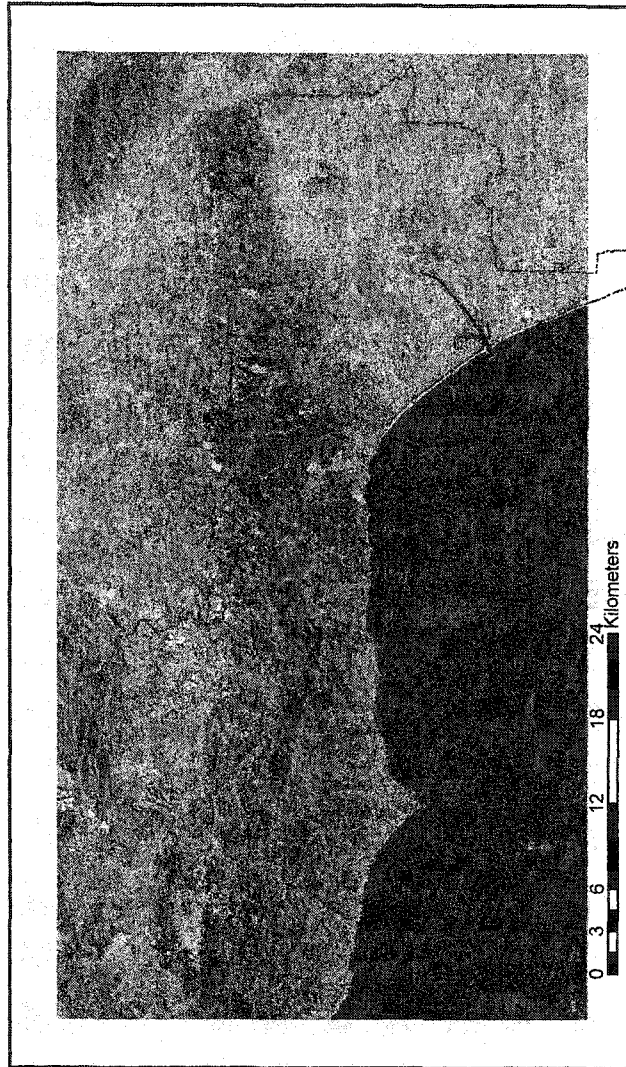


Figure 7
Infrared Composite Image of SMB

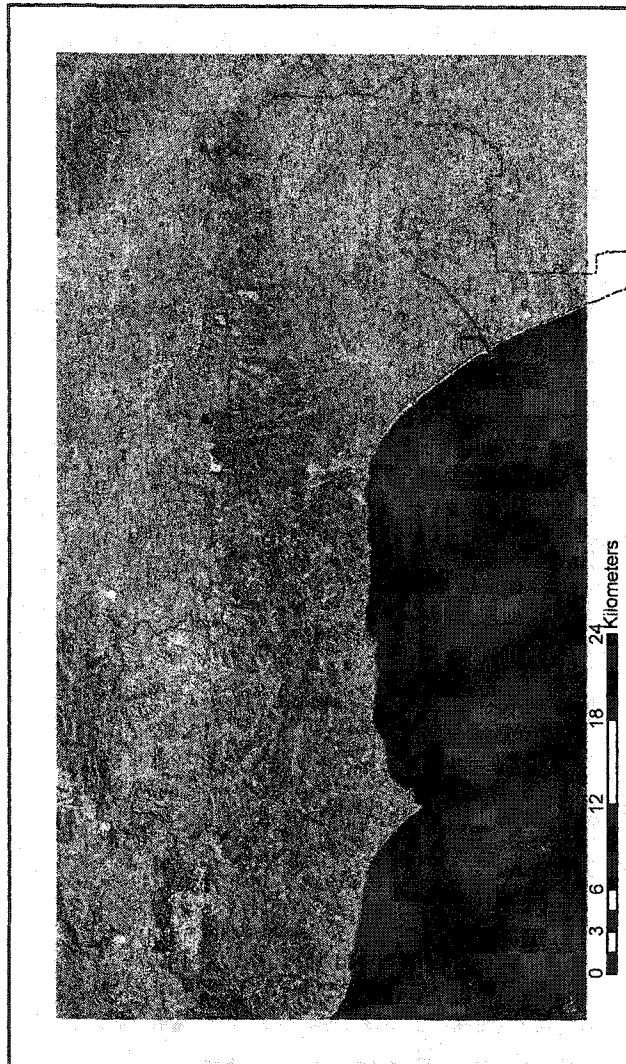


Figure 8
Infrared Composite Image of West Los Angeles



0 1.5 3 6 9 12 Kilometers

Figure 9
SMB Land Use Polygon Layer

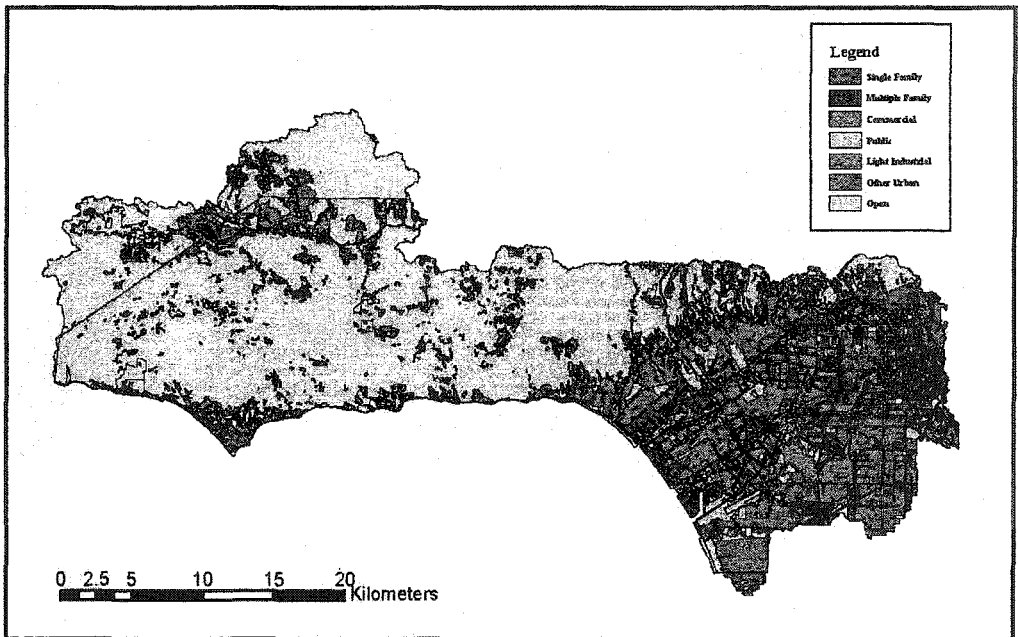


Figure 10 Average DN for Band 1 and Band 2

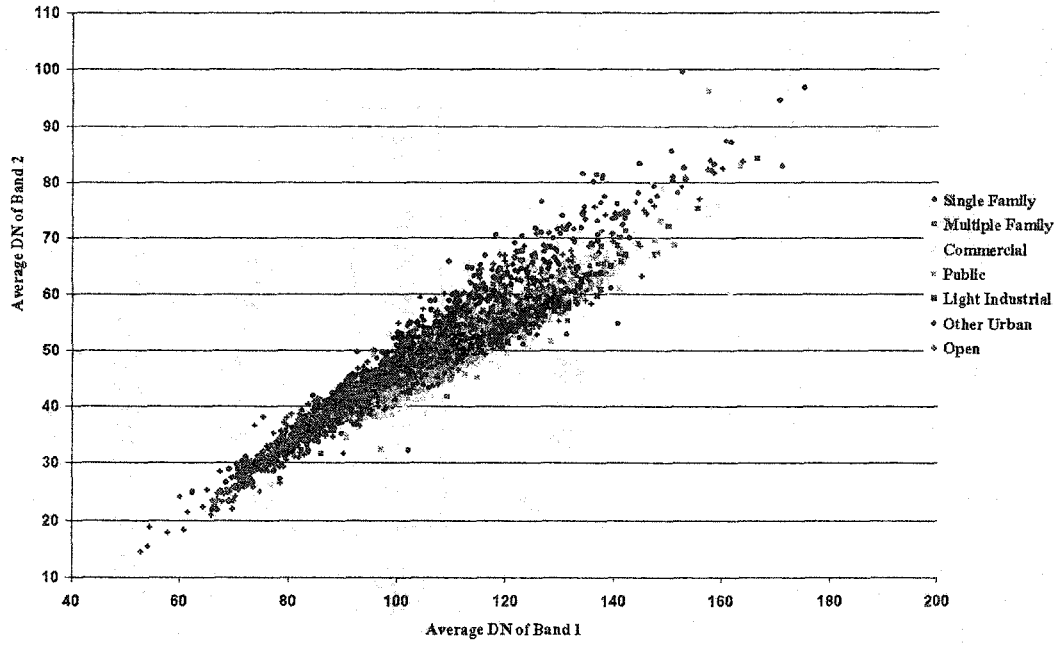


Figure 11 Average DNs of Single and Multiple Family Land Use for Band 1 and Band 2

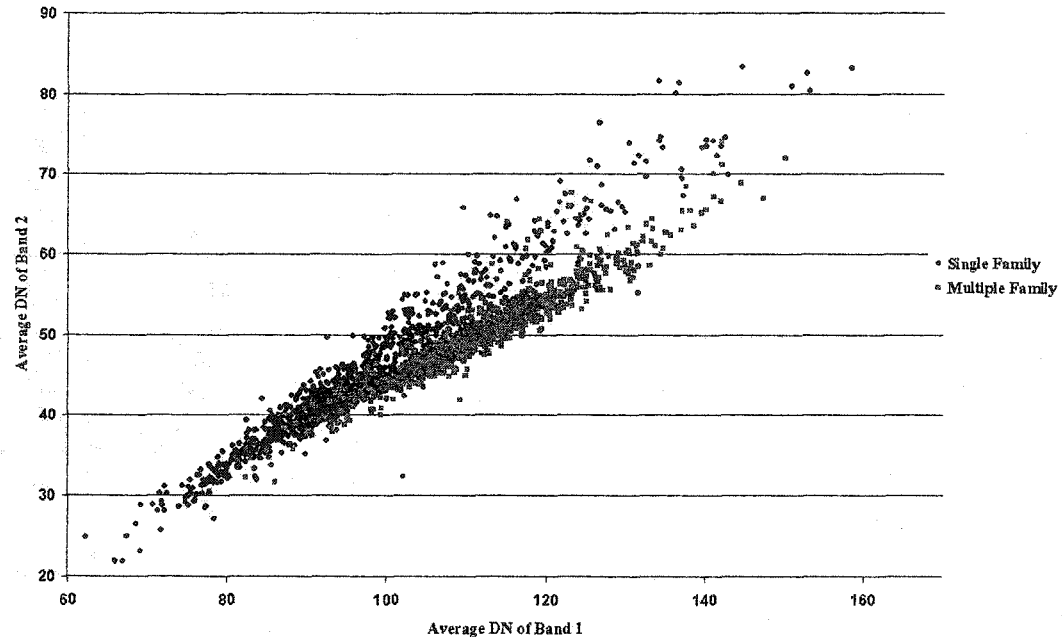


Figure 12 Average DN of Single and Multiple Family Land Use for Band 3 and Band 4

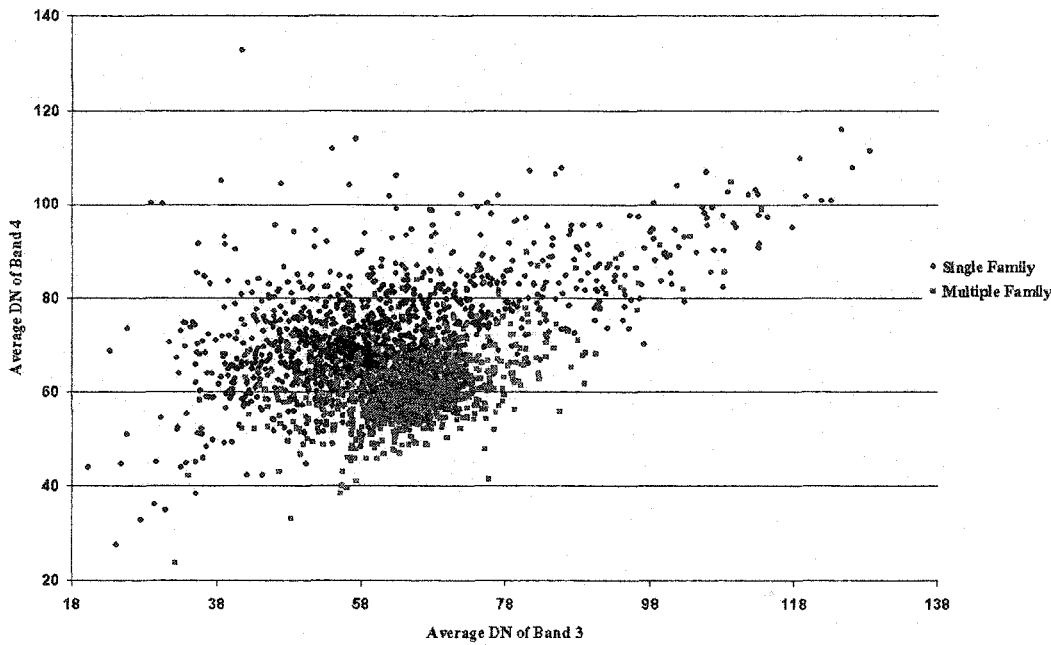
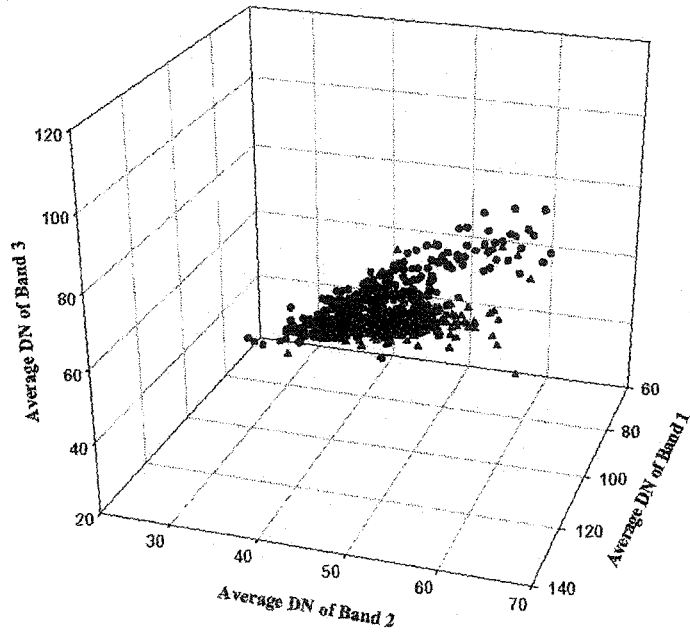


Figure 13
Average DN of Single and Multiple Family Polygons on Band 1, 2, and 3



- Single Family Land Use
- ▲ Multiple Family Land Use

Figure 14
Average DN's of Single and Multiple Family Polygons for Band 2, 3, and 4

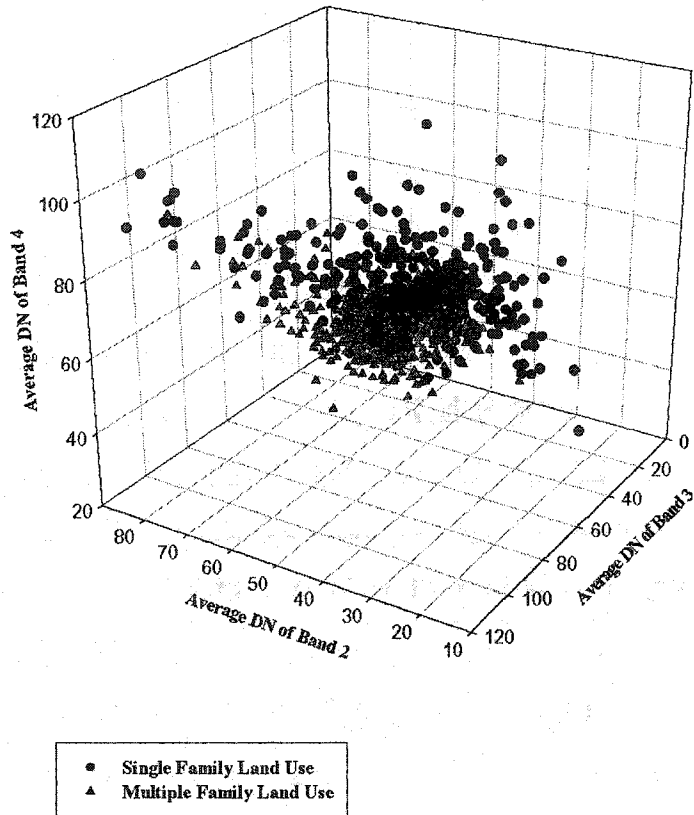


Figure 15 Average DN of Commercial and Open Land Use for Band 1 and Band 2

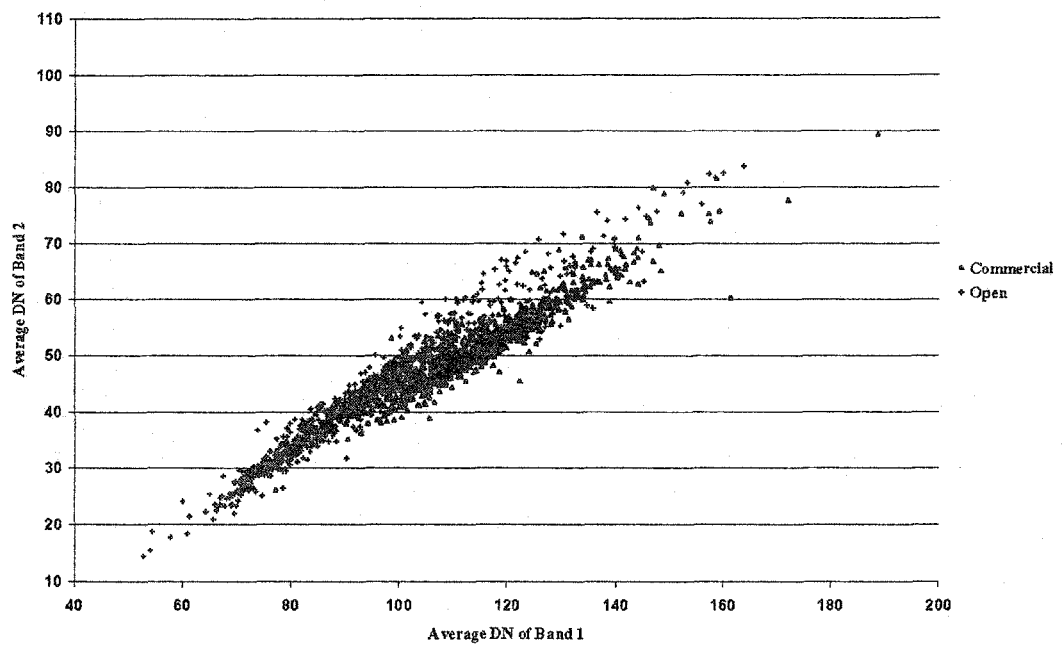


Figure 16 Average DN of Commercial and Open Land Use for Band 3 and Band 4

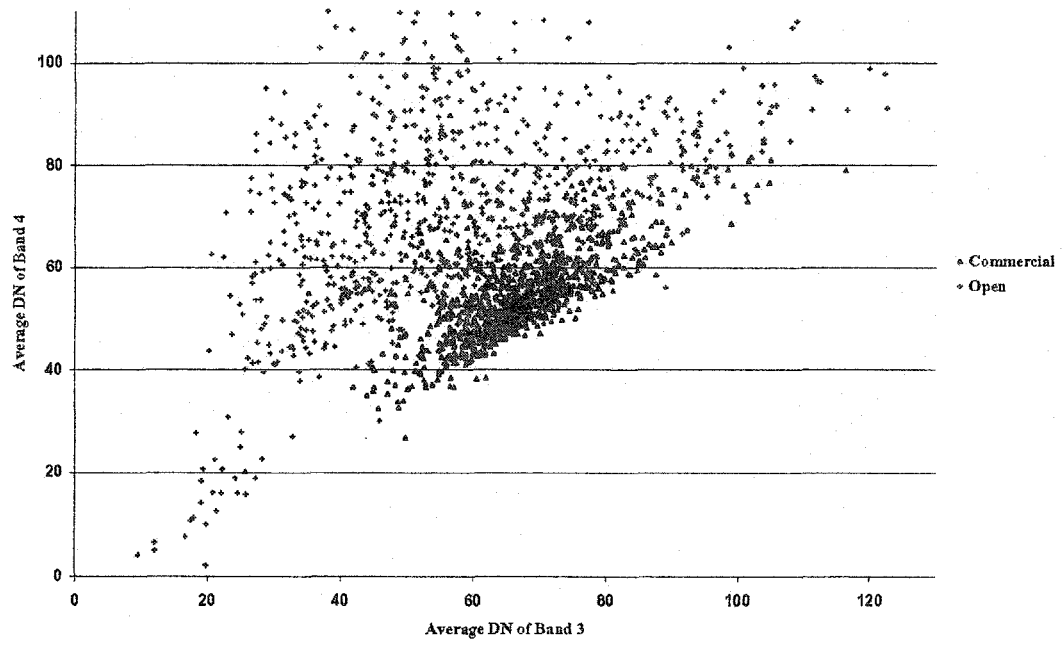
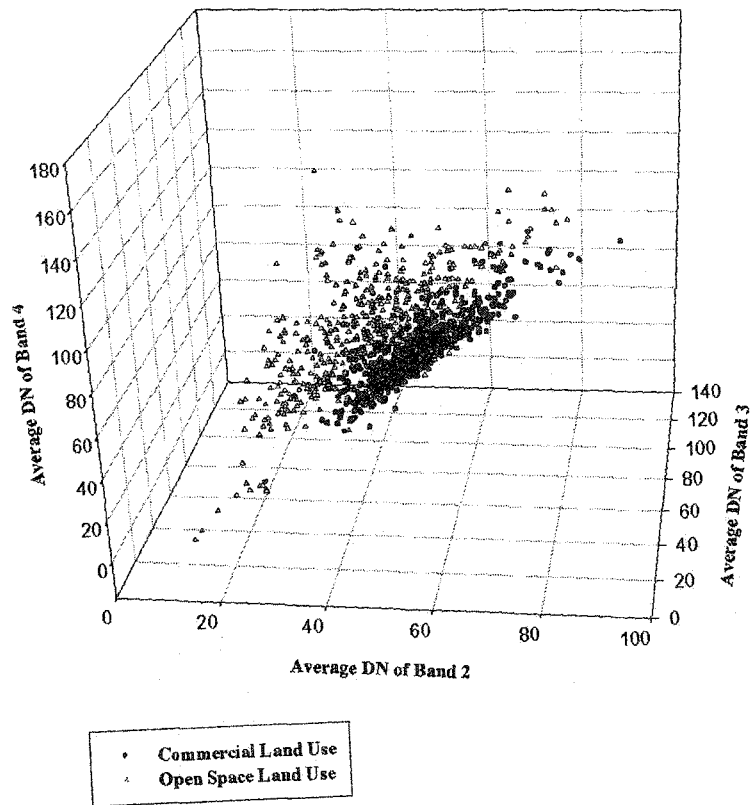


Figure 17
Average DNs of Commercial and Open Space Land Use for Band 2, 3, and 4



3. Fuzzy Neural Networks and GIS for Multi-Spectral Land Use Classification

Hsueh-hwa Lee, Michael K. Stenstrom, and Jiun-shiu Ma

Abstract

Until the appearance of geographic information system (GIS), the conversion of the spatially distributed remote sensed data into artificial neural network (ANN) input was an arduous, even insurmountable task. Characterizing the heterogeneous and spectrally complex systems required a level of technological sophistication that most researchers could not overcome; others remained daunted by the complexity of the data characterization process.

Now, after decades of research stymied by the ineffectuality of traditional approaches, the urban land use patterns of the Santa Monica Bay (SMB) can at last be classified. The use of fuzzy neural networks together with a GIS circumvents the challenges mentioned above in two steps. First, GIS spatially transforms remote sensed data into proper network input representation. Then the network was trained to learn land use patterns from various space images and geomorphic data. This classification is then quantitatively evaluated in terms of mean square error, percent accuracy, and correlation coefficient. An input data sensitivity analysis was conducted to evaluate the significance of each of these input parameters to the output.

The results elicited by these approaches support the main goal of this study: to develop a methodology for classifying SMB land use polygons with both Landsat Thematic Mapper images and USGS DEM (Digital Elevation Model) data. Thematic

Mapper spectral Band 4 was found to be the most significant input parameter for most land use categories, and the Multiple Layer Perceptron network with a fuzzy logic pre-processor outperformed all other classifiers.

3.1 Introduction

Over the past few years, ANN has gained enormous attention as a technique for classifying remote sensed data. These include: images acquired by the Landsat Multispectral Scanner (MSS) [Benediktsson et al., 1990; Lee et al., 1990], Landsat Thematic Mapper [Hepner et al., 1990; Civco, 1993; Yoshida and Omatu, 1994; Moody et al., 1996], and AVHRR [Gopal et al., 1994; 1997] as well as other data sets. The majority of studies indicate that ANN classifiers perform better than conventional statistic algorithms, for they can incorporate ancillary data easily and their architecture is more flexible and thus can be easily optimized for best performance. Additionally, they are distribution-free, so that no prior knowledge about the statistical distributions of the classes of data sources is required [Benediktsson et al., 1990].

Regarding classification, most ANN applications use a supervised and feedforward network structure with a back propagation algorithm [Key et al., 1989; Hepner et al., 1990; Benediktsson et al., 1990; Kanellopoulos et al., 1992]. While yielding a higher classification accuracy than other techniques, the computationally complex ANN classifiers can be very complex computationally and requires a lot of representative samples to elicit their performance.

Most investigators have constructed ANN applications without sufficient quantity or quality of data sets, and as well have lacked modern database management tools. In contrast, the approach adopted in this paper is the integration of the ANN model with spatial and spectral data stored and analyzed in GIS.

3.2 Methodology

3.2.1 Description of Raw Data

In addition to Thematic Mapper image and land use data as previously mentioned in Chapter 2, USGS 7.5-min DEM (Digital Elevation Model) data were used as ancillary data. The USGS 7.5-minute DEM data correspond to the USGS 1:24,000 and 1:25,000 scale topographic quadrangle map series for all of the United States. The 7.5-min. DEM is composed of 30 x 30-meter pixels, each of which stores an elevation value. Figure 3.1 presents the DEM of Santa Monica Bay. The average slope data (Figure 3.2) can be derived from DEM with GIS.

3.2.2 Description of the Artificial Neural Network Algorithm

Based on biological nervous systems, ANNs are parallel systems built from massive processing elements (PEs). These PEs are interconnected by massive sets of weights, and it is this interconnectivity that defines the ANN's topography. The signals flowing on the connections are scaled by a weight matrix. The sum of all these connections produces an output, which is a nonlinear function of that sum [Principe et al., 1999]. Because the PE is a processor with many different input connections and only one

output sent to other PEs, this output becomes either the system output or input of other PEs.

The Multiple Layer Perceptron (MLP) network with a backpropagation learning algorithm was used in this study [Bruzzone et. al., 1997; Serpico and Roli, 1995; Hertz et al., 1991; Rao et al., 1995]. The MLP network topology consists of multiple layers of PEs where only adjacent layers of PEs are connected. Input is processed from the input layer, and travels through successive hidden layers, eventually reaching the output layer.

MLP is a feedforward network, in which the information flows in forward direction only. Figure 3.3 shows the topology of an MLP with one hidden layer. The PE is the sum of products through nonlinear activation function f . Equation 1 describes the input (x) – output (y) relationship.

$$y = f(net) = f\left(\sum_{i=1}^D w_i x_i + b\right) \quad (1)$$

where D is the number of input PEs, x_i is input, w_i is weights, and b is a bias term. The activation function used is the hyperbolic tangent (\tanh) function.

$$f(net) = \tanh(net) \quad (2)$$

The nonlinearity of the Tanh function is smooth. This is very important for network learning, because it means that its derivative exists.

Back propagation, a learning algorithm based on training examples, was coupled with MLP at this point in the study. Thus, as is customary, it underwent supervised training with a finite number of pattern pairs consisting of both input and desired output patterns. The desired and computed outputs were compared first and subsequently a

function of errors was calculated. Adjustment of connection weights between was performed next. This procedure was repeated with each sample data set assigned for training the network until the error is confined within a desired tolerance. Each cycle through all the training samples is called an epoch.

The mean square error (MSE) is the sum of the square difference between the desired response and the computed output. It (J) can be calculated from Equation 3.

$$J = \frac{1}{2N} \sum_{p=1}^N \sum_{i=1}^M \varepsilon_{pi}^2 = \frac{1}{2N} \sum_{p=1}^N \sum_{i=1}^M (d_{pi} - y_{pi})^2 \quad (3)$$

where N is the number of observations, ε is an error between the desired response (d) and the computed output (y), p is the index over the patterns, and i is the index over the output PEs.

The goal of the ANN training process is to present a sufficient number N of unique input-output training samples to search for the connection weights to minimize cost function (Mean Square Error), J . The back propagation algorithm minimizes J by gradient decent rules and updates the weight values according to the following equation:

$$w_{ij}(n+1) = w_{ij}(n) + \eta \delta_i x_j(n) \quad (4)$$

where w_{ij} is the weight connecting the j th PE of the input layer and the i th PE of the output layer, η is the learning rate, and δ_i is the local error of the i th PE. δ_i is defined as

$$\delta_i(n) = -\varepsilon(n) y'_i(n) \quad (5)$$

if PE is at the output. For all other layers, the local error is computed by summing all the contributions of the local errors in the output layer, scaled by the corresponding weights

$$\delta_i(n) = y'_i(n) \sum_k \delta_k w_{ki}(n) \quad (6)$$

Momentum learning is an improvement to gradient-descent search. In momentum learning, the following equation is used for weight update:

$$w_{ij}(n+1) = w_{ij}(n) + \eta \delta_i(n) x_j(n) + \alpha (w_{ij}(n) - w_{ij}(n-1)) \quad (7)$$

where α is the momentum constant, usually between 0.5 and 0.9. The momentum term can speed up the learning process, keep the weight update process moving, and thereby not get stuck in a local minimum.

The network training process should be stopped when the network has learned the patterns of all of the training samples. Recent developments in learning theory [Vapnik, 1995] indicate that after a critical point, the MLP with backpropagation learning algorithm will continue to perform better in the training data sets, but the performance of testing data sets will begin to deteriorate. This phenomenon, called overtraining, can be tempered by the *cross validation* method. The training data sets are divided into two data sets: training and cross validation sets. The training process should stop when the error of the cross validation set begin to increase. Utilization of cross validation can facilitate MLP in reaching maximum generalization: how well the network performs with data not belonging to the training set.

3.2.3 Input Sensitivity Study

The sensitivity analysis of MLP input parameters can be performed by fixing all the weights, and, while keeping the other inputs at their mean value, randomly perturbing each channel of the input around its mean value, and then measuring the change in the output (Principe et al., 2000). The sensitivity for input k is can be expressed as

$$S_k = \frac{\sum_{p=1}^P \sum_{i=1}^o (y_{ip} - \bar{y}_{ip})^2}{\sigma_k^2} \quad (8)$$

where \bar{y}_{ip} is the i th output obtained with the fixed weights for the p th pattern, o is the number of network outputs, P is the number of patterns, and σ_k^2 is the variance of the input perturbation. Input parameters with large sensitivities have a significant impact on the results and should be kept. In contrast, input parameters with smaller sensitivities have less impact on the result and can be discarded. This is critical for conserving computing resources and reducing the cost of data collection.

3.2.4 Fuzzy Logic

Because of the natural synergy between neural networks and fuzzy logic, the latter has become a major instrument in processing uncertainty in geographical databases [Bigus, 1996; Canters, 1997; Rao et al., 1995]. Fuzzy logic, best known in the context of set membership, is concerned with imprecision, while neural networks deal with learning. First proposed by Dr. Lotfi Zadeh at the University of California at Berkley in 1965, fuzzy sets are data sets in which members are presented as ordered pairs, including information on degree of membership [Zadeh, 1994].

In this project, triangle fuzzy functions were used to map raw input data onto a range of 0.0 and 1.0. The function f was shaped by calculating the global average and the standard deviation of each input parameter (Figure 3.4). Fuzzy sets for input parameters of all Thematic Mapper spectral bands in the Santa Monica Bay are listed in Figure 3.5(a-g). Figure 3.6, for example, demonstrates how the memberships of seven land use categories can be obtained from a single pixel value input through the fuzzy function of Band 1.

The most effective application of fuzzy logic in artificial neural networks is as a data pre-processor (Figure 3.7). This results in a multi-layer feedforward network with fuzzy logic in either the processing units or in the connection weight representations [Buckley and Hayashi, 1994]. By using intuitive fuzzy rules to represent knowledge and converting them into feedforward neural networks, it creates a way of imparting explicit domain knowledge to neural networks, without the need for training [Okada et al., 1992]. The combination of neural networks and fuzzy logic results not only in better initial performance by the network, but also in faster learning.

3.3 Results and Discussion

3.3.1 Network Construction

In the next stage of the project, a MLP with a back propagation learning algorithm was constructed (Figure 3.8). The input was first normalized from -1 to 1 , and then it was summed up and sent to hidden layer through a hyperbolic tangent function. The output of the hidden layer was again summed up and delivered to an output layer through

another hyperbolic tangent function. The output of that layer was then compared with the desired output of the training samples and the errors were calculated. The weights connecting the layers were adjusted by the back propagations algorithms until the training stop criteria were satisfied. Table 3.1 lists all the input parameters used this study.

When fuzzifier was used as an MLP pre-processor, each of the spectral band data set was fuzzified to one of seven land use fuzzy memberships. The coordinate data of the polygon centroid calculated geo-spatially with GIS were used as ancillary input. Four MLP networks and one Fuzzy Neural Network of different input were constructed; their characteristics are listed in Table 3.2. In the Fuzzy Neural Network (NeuroFuzzy1), the input of the seven spectral bands was fuzzified into 49 memberships (seven memberships based on land use categories for each of the seven spectral bands). 524 (10%), 1,310 (25%), and 3,407 (65%) among a total of 5,241 land use polygons were randomly chosen as cross validation, testing, and training data (Figure 3.9).

3.3.2 Classification Results

The MSEs of the networks are listed in Table 3.3. The training results showed that the MSEs of both training and cross validation data decreased as the input increased. The cross validation MSEs of all except the NeuroFuzzy1 networks reached their minimum at the end of training (1,000th epoch). For NeuroFuzzy1, the MLP Network with fuzzy logic pre-processor, the cross validation MSE minimized at the 498th epoch and started to increase slightly as the training process continued and as the training MSE continued to decrease. This observation is proof that the addition of fuzzy logic

processor can accelerate the training process when knowledge is introduced to the network explicitly through the fuzzifier function.

However, due to the larger network size, the neural fuzzy network can be very costly in terms of computing resources. The MSE is a valuable indicator, but there is no direct relationship between it and classification accuracy. The performance of a network is based on the accuracy of classification. The classification results (percent accuracy) for each land use categories of training, cross validation, and testing are listed in Table 3.4(a), (b) and (c).

Correlation coefficient, r , is another parameter to evaluate the network performance. Table 3.5 lists the r^2 of training, cross validation, and testing data. As with the inverse ratio of MSEs to input, in general, classification accuracy improves as the networks receive more training information--more input. The classification of training data outperformed those of cross validation and testing data. MLP1 performed poorly with limited input information--Bands 1, 2, and 3. With the addition of spectral Bands 4, 5, 6, and 7, MLP2 significantly increased the classification accuracy. MLP3, which incorporated geomorphic information (DEM and Slope) into the network, performed only slightly better than MLP2. The addition of coordinate information (X-Cen and Y-Cen) again enhanced the accuracy in MLP4. Finally, NeuroFuzzy1, equipped with the fuzzy pre-processor, increased the classification accuracy significantly.

The true land use classification is compared with the output from the neural network in a table entitled Confusion Matrix. A perfect classification gives a confusion matrix with only the diagonal populated and all other entries zero. Table 3.6 presents the

Confusion Matrix of NeuroFuzzy1 network. The classification results of Public, Light Industrial, and Other Urban land use polygons were less satisfying than other categories. This might have resulted from the fact that fewer training samples were selected in the training process for these three categories, and thus, the network weights were trained favorably to match other categories.

Further complicating things, land use definitions themselves can be ambiguous. For instance, two land use polygons with similar physical and geomorphic features might be classified into two categories for political, economic, or taxational purposes. In Table 3.6, most misclassified Multiple Family polygons were classified as Single Family land use, due to their similar spectral feature. The majority of misclassified Light Industrial land use polygons were classified as Commercial or Open land use. Most wrongfully identified Other Urban land use polygons were categorized to Open land use. The observation from the confusion matrix was consistent with the common sense used in land use definition.

3.3.3 Sensitivity Analysis

A sensitivity analysis of input parameters was performed based on MLP4 training data. The change of output was monitored while specific input was varied and all others were fixed at their mean values (Figure 3.10). Table 3.7 calculates the sensitivity of all input parameters for each land use category. In three of seven land use categories, Spectral Band 4 was found to be the most significant parameter as well as having a major impact on the others. Spectral Band 4, the so-called "Infrared Band," has proven very

useful in urban feature identification in many other studies. The results of the sensitivity analysis from this study corroborate traditional spectral band classification theories. They also indicate that coordinate information had little impact on the output based on the sensitivity analysis, though experimental observation did demonstrate the performance improvement

3.4 Conclusions and Future Research

The main purpose of this investigation was the assessment of the effectiveness of multi-source data for land use characterization and the capabilities of fuzzy neural networks with GIS to efficiently exploit remote sensed data. The results not only attest to the ability of MLPs with Fuzzy Logic pre-processors to classify urban land use patterns, they also reinforce certain research features, e.g., the importance of spectral data as well as other ancillary geomorphic and coordinate data in complex land use classification. Better classification performance could have been elicited with more complete input information.

The results also indicate that the incorporation of fuzzy logic into neural networks greatly enhances the overall performance and yields satisfactory accuracy. The sensitivity analysis recognizes the most significant input parameters impacting the classification results. The clumsiness of the network to classify certain land use polygons is undesirable, but not surprising, given the complex nature of land use definitions.

Increased network accuracy can be achieved by the introduction of economic and political input parameters. It should also be noted that, though many environmental and

engineering studies rely heavily on current land use definitions to estimate imperviousness, concentrations of pollutants, and other parameters, it does not mean that these definitions are ideal for environmental or engineering purposes.

This study focused on classification of existing land use polygons and the input variables were averaged in a polygon, but there are several areas in which it could be expanded. Future studies could adopt the Thematic Mapper image pixel (30 meters by 30 meters) as operation units for study of the interesting features for each pixel. Pixel-level operations require much more computing resources. For example, the Santa Monica Bay study area is composed of only 5,241 land use polygons, but approximately 1.1 million image pixels. The pixel level operation could be a great challenge to neurocomputing.

Future studies should include other important physical parameters as well: imperviousness, pollutant concentration, canopy cover, soil type, vegetation type, etc. The network should perform regression analysis in addition to classification tasks in order to estimate those interesting parameters with quantitative magnitude. It is very difficult to find proper training data sets for neural network regression analyses. Extensive field survey and lab analysis need to be conducted to gather sufficient data sets.

Another issue mentioned earlier is related to the proper definition of land use. Current definitions might not be suitable for certain research interests. Unsupervised neural networks could be developed to discover valuable patterns hidden in multi-dimensional space composed by input vectors. These new patterns are strictly related to only input data without any outside or pre-determined influence.

The methodology in this study can be applied to higher resolution data, such as aerial photo or IKONOS Satellite images. The resolution of the IKONOS Satellite (launched in 1999) is one meter and the resolution of aerial photo could be less than an inch with a much higher cost. The higher spectral resolution images, coupled with other ancillary data and appropriate neural networks, will greatly improve the accuracy of land use classification.

3.5 References

- AIS (1996). *Southern California 1990 Aerial Land Use Study: Land Use Level III/IV Classification*. Aerial Information Systems, Redlands, California.
- Anderson, J. R., Hardy, E. T., and Witmer, R. E. (1976). "A land use and land cover classification system for use with remote sensor data", U.S. Geological Survey Professional Paper 964.
- Benediktsson, J. A., Swain, P. H., and Ersoy, O. K. (1990). "Neural Network Approaches versus Statistical Methods in Classification of Multisource Remote Sensing Data." *IEEE Trans. Geosci. Remote Sens.* Vol. 28, pp. 540-552.
- Bigus, Joseph. (1996). *Data Mining with Neural Networks*. McGraw Hill, Inc., New York, NY.
- Bruzzone, L., Conese, C., Maselli, F., and Roli, F. (1997). "Multisource Classification of Complex Rural Areas by Statistical and Neural-Network Approaches." *Photogrammetric Engineering & Remote Sensing*, Vol. 63, No. 5, pp. 523-533.

- Buckley, J. J. and Hayashi, Y. (1994) Fuzzy Neural Networks: A Survey, Fuzzy Sets and Systems, Vol. 66, No. 1.
- Canters, F. (1997). "Evaluating the Uncertainty of Area Estimates Derived from Fuzzy Land-Cover Classification." *Photogrammetric Engineering & Remote Sensing*, Vol. 63, No. 4, pp. 403-414.
- Civco, D. L. (1993). "Artificial Neural Networks for Land-cover and Mapping." *International Journal of Geographical Information Systems*. Vol. 7, pp. 173 – 186.
- Gopal, Sucharita, Woodcock, C. E., and Strahler, A. H. (1999). "Fuzzy Neural Network Classification of Global Land Cover from a 10 AVHRR Data Set", *Remote Sens. Environ.* Vol. 67, pp. 230-243.
- Gopal, S., Sklarew, D. M. and Lambin, E. (1994). "Fuzz-neural Networks in Multi-temporal Classification of Land-cover Change in the Sahel." In *Proceedings of the DOSES Workshop on New Tools for Spatial Analysis*, Lisbon, Portugal, DOSES, EUROSTAT, ECSC-EC-EAEC, Brussels, Luxembourg, pp. 55 - 68.
- 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." *Photogramm. Eng. Remote Sens.* Vol. 56, pp. 469 – 473.
- Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*, Addison Wesley Pub. Co., The Advance Book Program.
- Kanellopoulos, I., Varfis, A., Wilkinson, G., and Megier, J. (1992). "Land-cover Discrimination in SPOT HRV Imagery Using an Artificial Neural Network – a 20

- class experiment.” *International Journal of Remote Sensing*. Vol. 13, pp. 917 – 924.
- Key, J., Maslanik, J. A., and Schweiger, A. J. (1989). “Classification of Merged ANHRR and SMMR Arctic Data with Neural Networks, *Photogramm. Engr. Remote Sens.* Vol. 55, pp. 1331 – 1338.
- Lee, J., Weger, W. C., Sengupta, S. K., and Welch, R. M. (1990). “A Neural Network Approach to Cloud Classification.” *IEEE Trans. Geosci. Remote Sens.* Vol. 28, pp. 846 – 855.
- Moody, A., Gopal, S. and Strahler, A. H. (1996). “Artificial Neural Network Response to Mixed Pixels in Coarse-Resolution Satellite Data.” *IEEE Trans. Geosci. Remote Sens.* Vol. 58, pp. 329 – 343.
- Okada, H., Watanabe, N., Kawamura, A., Asakawa, K., Taira, T., Isida, K., Kaji, T., and Narita, M. (1992). “Initializing multilayer neural networks with fuzzy logic”, *Proceedings of the IEEE International Conference on Neural Networks*, Vol. 1.
- Principe, C. P., Euliano, N. R., and Lefebvre, W. C. (2000). *Neural and Adaptive Systems: Fundamentals Through Simulations*. John Wiley & Sons, Inc., New York.
- Rao, V. B. and Rao, H. V. (1995). *C++ Neural Networks & Fuzzy Logic*. MIS Press. New York, NY.
- Serpico, S. B., and Roli, F. (1995). “Classification of multisensor remote sensing images by structured neural networks, *IEEE Transactions on Geoscience and Remote Sensing*, 33(3): 562 –578.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

Yoshida, T., and Omatu, S. (1994). "Neural Network Applications to Land-cover Mapping." *IEEE Trans. Geosci. Remote Sens.* Vol. 32, pp. 1103 – 1109.

Zadeh, L. A. (1994). "Fuzzy logic, neural networks, and soft computing" in *Communications of the ACM*, 3, pp. 77 – 84.

Table 3.1 Summary of Input Data Used

Input Code	Description	Unit	Range
Band1	Average Pixel Value of TM* Spectral Band 1 in a Land Use Polygon	Dimensionless	53 - 188
Band2	Average Pixel Value of TM Spectral Band 2 in a Land Use Polygon	Dimensionless	14 - 100
Band3	Average Pixel Value of TM Spectral Band 3 in a Land Use Polygon	Dimensionless	10 - 143
Band4	Average Pixel Value of TM Spectral Band 4 in a Land Use Polygon	Dimensionless	2 - 159
Band5	Average Pixel Value of TM Spectral Band 5 in a Land Use Polygon	Dimensionless	0 - 217
Band6	Average Pixel Value of TM Spectral Band 6 in a Land Use Polygon	Dimensionless	136 - 198
Band7	Average Pixel Value of TM Spectral Band 7 in a Land Use Polygon	Dimensionless	2 - 140
DEM	Average DEM in a Land Use Polygon	Meter	0 - 838
Slope	Average Slope in a Land Use Polygon	Degree	0 - 48
X-Cen	The X Coordinate of Centroid in a Land Use Polygon**	Meter	320,602 - 384,436
Y-Cen	The Y Coordinate of Centroid in a Land Use Polygon**	Meter	3,753,925 - 3,786,702

Table 3.2 Lists of Networks Constructed

Network	No. of Input Layer Pes	No. of Hidden Layer PEs	Network Inputs
MLP1	3	7	Band1, Band2, Band3
MLP2	7	14	Band1, Band2, Band3, Band4, Band5, Band6, Band7
MLP3	9	18	Band1, Band2, Band3, Band4, Band5, Band6, Band7, DEM, Slope
MLP4	11	20	Band1, Band2, Band3, Band4, Band5, Band6, Band7, DEM, Slope, X-Cen, Y-Cen
NeuroFuzzy1	53	41	Memberships of Band 1 to 7 (7 x 7), DEM, Slope, X-Cen, Y-Cen

Table 3.3 Network Mean Square Errors

Network	Training Data			Cross Validation		
	Epoc No.*	Minimum MSE	Final MSE	Epoc No.*	Minimum MSE	Final MSE
MLP1	1000	0.199	0.199	1000	0.196	0.196
MLP2	1000	0.171	0.171	1000	0.168	0.168
MLP3	1000	0.166	0.166	1000	0.167	0.167
MLP4	1000	0.156	0.156	1000	0.160	0.160
NeuroFuzzy1	1000	0.134	0.134	498	0.152	0.187

*Epoch when MSE is at its minimum

Table 3.4(a) Testing Results on Training Data

Network	Land Use Categories (% Accuracy of Training Data)						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
MLP1	70	58	56	26	21	16	68
MLP2	77	71	73	52	58	49	83
MLP3	80	75	73	55	60	55	86
MLP4	83	77	78	60	65	60	88
NeuroFuzzy1	91	80	86	70	71	66	91

Table 3.4(b) Testing Results on Cross Validation Data

Network	Land Use Categories (% Accuracy of Cross Validation Data)						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
MLP1	69	64	56	20	15	15	56
MLP2	69	79	67	40	46	44	75
MLP3	75	79	70	54	54	52	80
MLP4	77	82	75	58	62	52	85
NeuroFuzzy1	88	83	80	66	62	59	89

Table 3.4(c) Testing Results on Testing Data

Network	Land Use Categories (% Accuracy of Testing Data)						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
MLP1	68	66	51	15	13	15	57
MLP2	70	70	65	40	50	46	77
MLP3	78	77	71	55	60	55	82
MLP4	82	80	77	57	59	58	84
NeuroFuzzy1	90	82	80	67	67	60	90

Table 3.5(a) Correlation of Training Data

Network	r^2						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
MLP1	0.46	0.35	0.40	0.14	0.19	0.28	0.46
MLP2	0.71	0.71	0.54	0.41	0.43	0.41	0.75
MLP3	0.74	0.76	0.66	0.52	0.57	0.51	0.76
MLP4	0.81	0.82	0.79	0.55	0.61	0.55	0.79
NeuroFuzzy1	0.87	0.85	0.82	0.61	0.65	0.66	0.89

Table 3.5(b) Correlation of Cross Validation Data

Network	r^2						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
MLP1	0.48	0.35	0.44	0.19	0.18	0.19	0.50
MLP2	0.71	0.69	0.48	0.40	0.40	0.38	0.77
MLP3	0.73	0.72	0.62	0.53	0.61	0.55	0.71
MLP4	0.75	0.76	0.71	0.54	0.62	0.57	0.76
NeuroFuzzy1	0.82	0.81	0.77	0.60	0.65	0.61	0.85

Table 3.5(c) Correlation of Testing Data

Network	r^2						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
MLP1	0.43	0.32	0.42	0.17	0.24	0.20	0.48
MLP2	0.73	0.68	0.47	0.42	0.43	0.41	0.77
MLP3	0.78	0.71	0.50	0.43	0.51	0.44	0.79
MLP4	0.81	0.79	0.69	0.52	0.61	0.55	0.79
NeuroFuzzy1	0.85	0.82	0.79	0.59	0.63	0.59	0.85

Table 3.6(a) Confusion Matrix of Training Data (NeuroFuzzy1)

Predict / Desired	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open	Total
LU11-Single Family	777	73	18	5	3	10	29	915
LU12-Multiple Family	21	561	40	10	1	8	7	648
LU20-Commercial	5	19	554	20	15	2	10	625
LU30-Public	19	34	22	275	5	3	4	362
LU40-Light Industrial	0	0	0	0	70	0	0	70
LU50-Other Urban	0	0	0	0	0	106	1	107
LU60-Open	30	11	12	83	5	31	508	680
Total	852	698	646	393	99	160	559	3,407

Table 3.6(b) Confusion Matrix of Cross Validation Data (NeuroFuzzy1)

Predict / Desired	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open	Total
LU11-Single Family	117	12	8	3	0	1	2	143
LU12-Multiple Family	10	92	5	2	1	1	2	113
LU20-Commercial	3	1	73	10	3	1	2	93
LU30-Public	1	4	3	44	0	0	3	55
LU40-Light Industrial	0	0	0	0	8	0	0	8
LU50-Other Urban	0	0	0	0	0	16	0	16
LU60-Open	2	2	2	8	1	8	73	96
Total	133	111	91	67	13	27	82	524

Table 3.6(c) Confusion Matrix of Testing Data (NeuroFuzzy1)

Predict / Desired	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open	Total
LU11-Single Family	281	24	9	15	1	2	17	349
LU12-Multiple Family	13	207	5	11	1	2	4	243
LU20-Commercial	3	10	195	10	9	1	2	230
LU30-Public	4	11	27	95	1	9	1	148
LU40-Light Industrial	0	0	0	0	26	0	0	26
LU50-Other Urban	1	0	0	0	0	50	0	51
LU60-Open	10	1	8	11	1	20	212	263
Total	312	253	244	142	39	84	236	1,310

Table 3.7 Sensitivity of Input Parameters

Sensitivity	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
Band1	0.0116	0.0078	0.0025	0.0006	0.0002	0.0003	0.0007
Band2	0.0129	0.0116	0.0018	0.0004	0.0000	0.0004	0.0005
Band3	0.0074	0.0088	0.0017	0.0005	0.0001	0.0003	0.0007
Band4	0.0138	0.0057	0.0097	0.0010	0.0001	0.0008	0.0009
Band5	0.0043	0.0091	0.0017	0.0007	0.0001	0.0001	0.0024
Band6	0.0044	0.0093	0.0005	0.0004	0.0003	0.0002	0.0008
Band7	0.0007	0.0114	0.0006	0.0029	0.0000	0.0003	0.0002
X-Cen	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Y-Cen	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM	0.0002	0.0009	0.0001	0.0001	0.0000	0.0000	0.0000
Slope	0.0079	0.0092	0.0023	0.0115	0.0001	0.0005	0.0003
Most Sensible Input	Band 4	Band 2	Band 4	Slope	Band 6	Band 4	Band 5

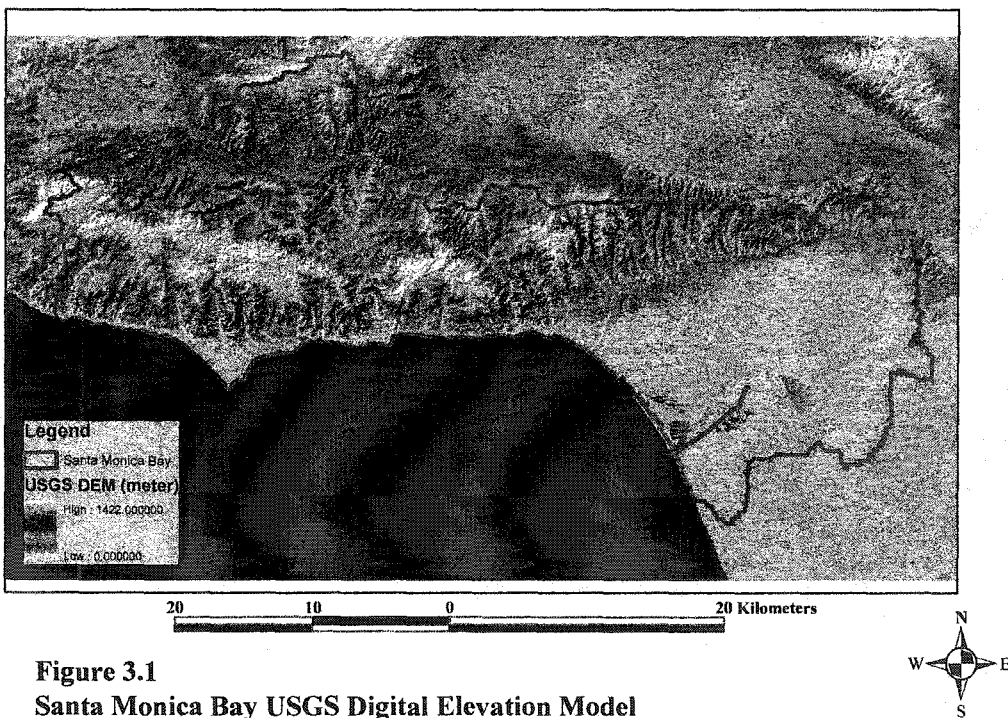


Figure 3.1
Santa Monica Bay USGS Digital Elevation Model

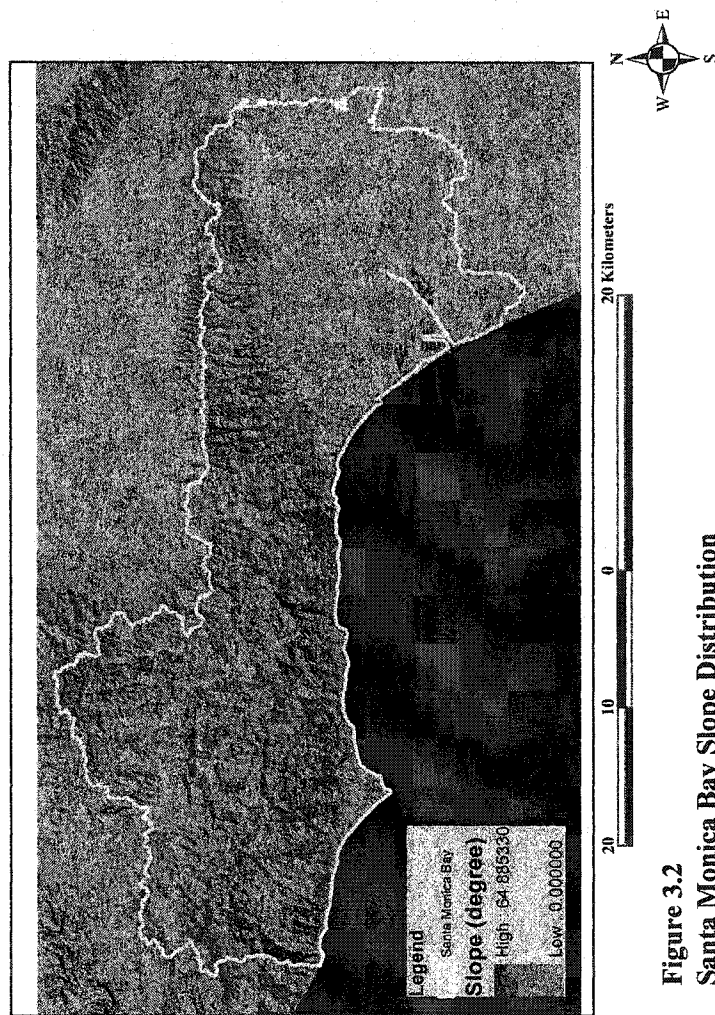


Figure 3.2
Santa Monica Bay Slope Distribution

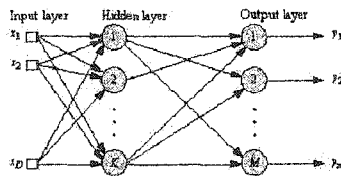


Figure 3.3 MLP of One Hidden Layer with D Inputs and M Outputs

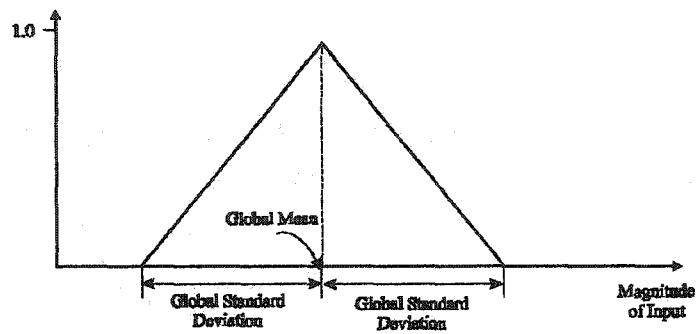


Figure 3.4 Triangular Fuzzy Function

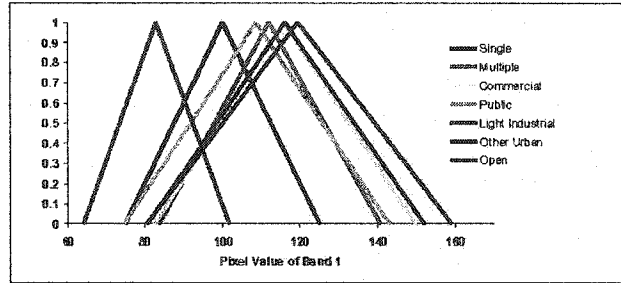


Figure 3.5(a) Fuzzy Sets for Land Use Categories Based on Average Pixel Value of Spectral Band 1

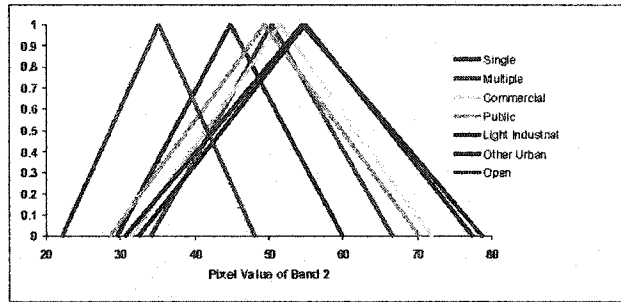


Figure 3.5(b) Fuzzy Sets for Land Use Categories Based on Average Pixel Value of Spectral Band 2

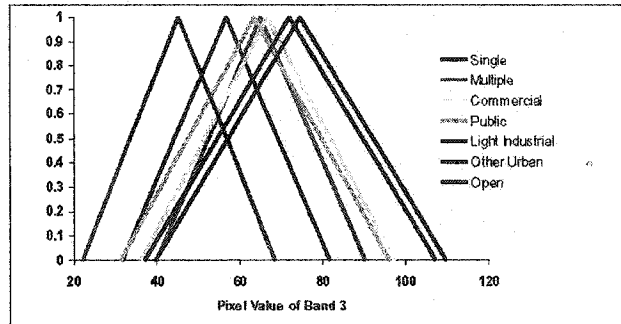


Figure 3.5(c) Fuzzy Sets for Land Use Categories Based on Average Pixel Value of Spectral Band 3

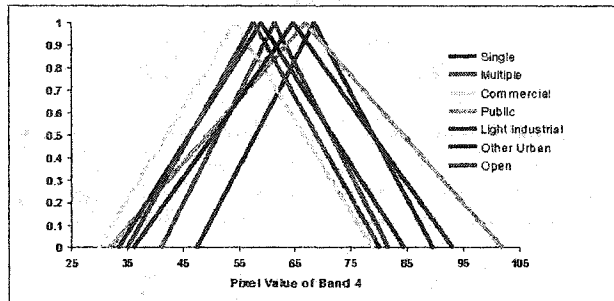


Figure 3.5(d) Fuzzy Sets for Land Use Categories Based on Average Pixel Value of Spectral Band 4

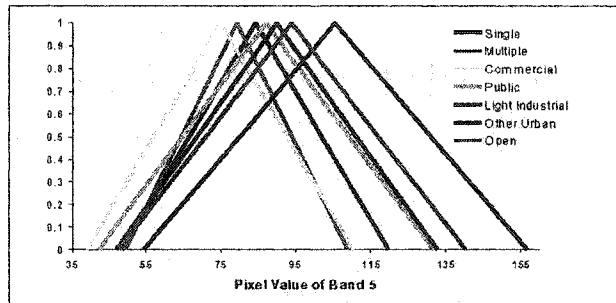


Figure 3.5(e) Fuzzy Sets for Land Use Categories Based on Average Pixel Value of Spectral Band 5

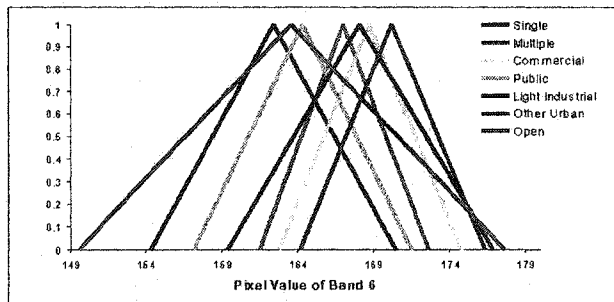


Figure 3.5(f) Fuzzy Sets for Land Use Categories Based on Average Pixel Value of Spectral Band 6

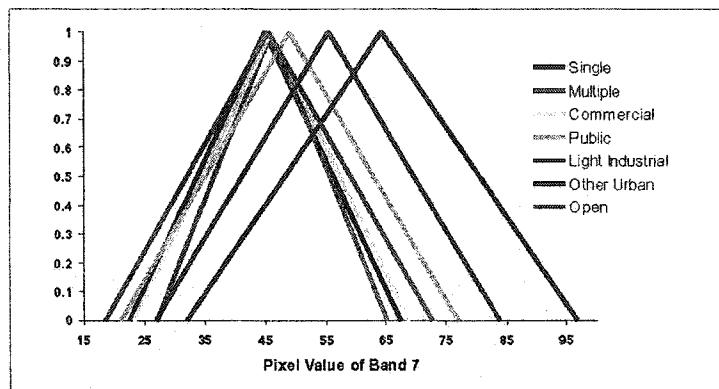


Figure 3.5(g) Fuzzy Sets for Land Use Categories Based on Average Pixel Value of Spectral Band 7

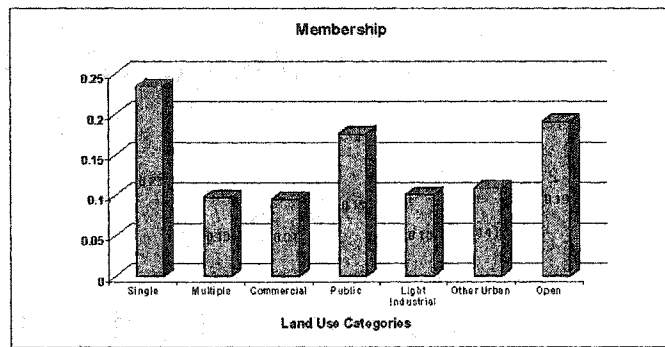
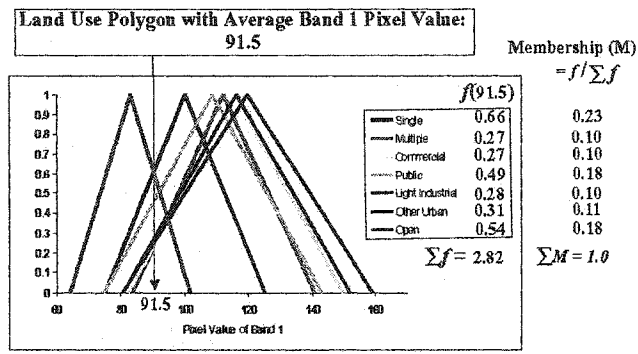


Figure 3.6 Derivation of Memberships from Single Input Through Fuzzifier Function

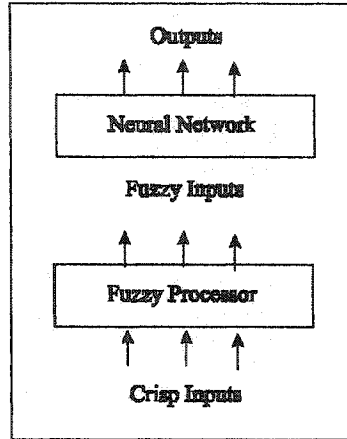


Figure 3.7 A Neural Network with Fuzzy Pre-processor

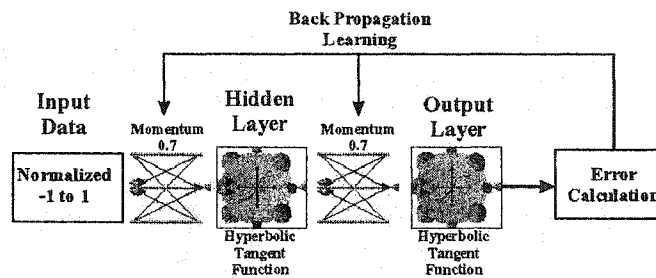


Figure 3.8 MLP with Back Propagation Learning

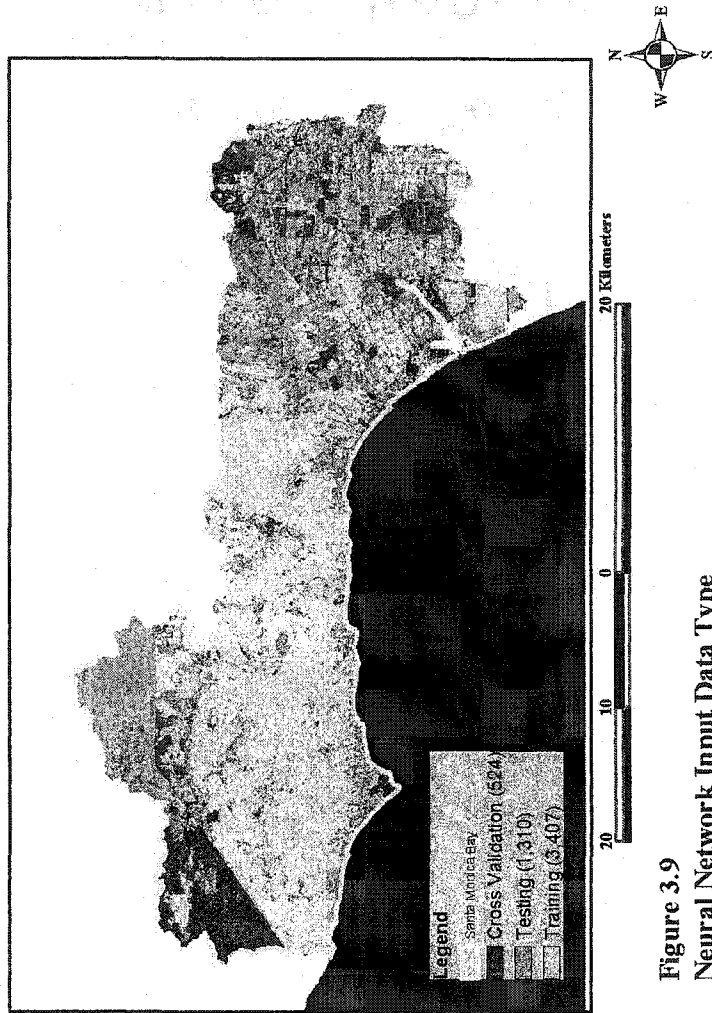


Figure 3.9
Neural Network Input Data Type

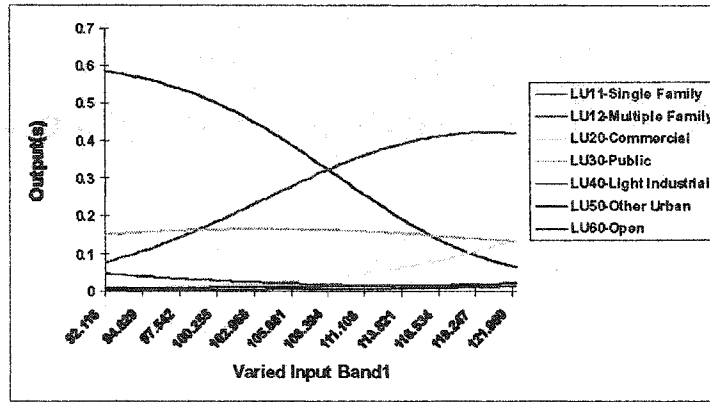


Figure 3.10(a) Network Outputs with Varied Band1

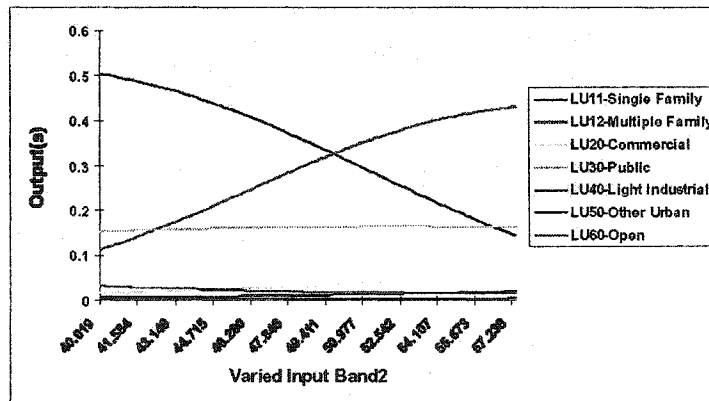


Figure 3.10(b) Network Outputs with Varied Band2

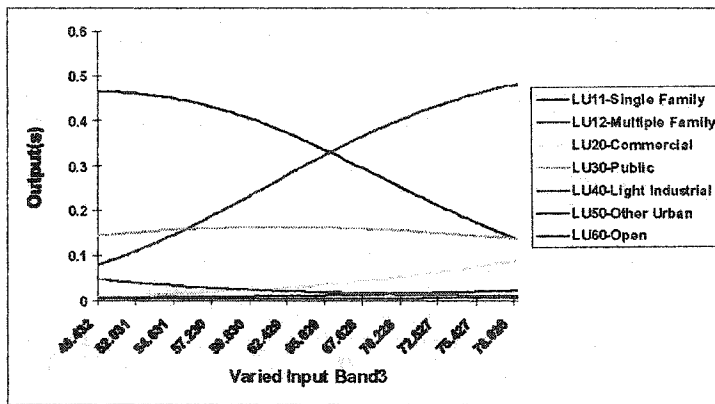


Figure 3.10(c) Network Outputs with Varied Band3

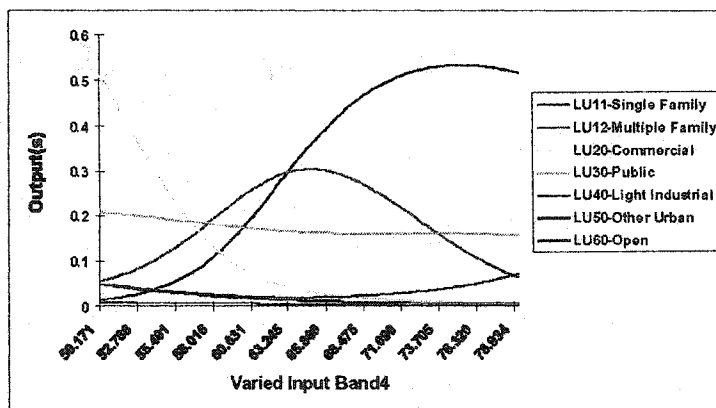


Figure 3.10(d) Network Outputs with Varied Band4

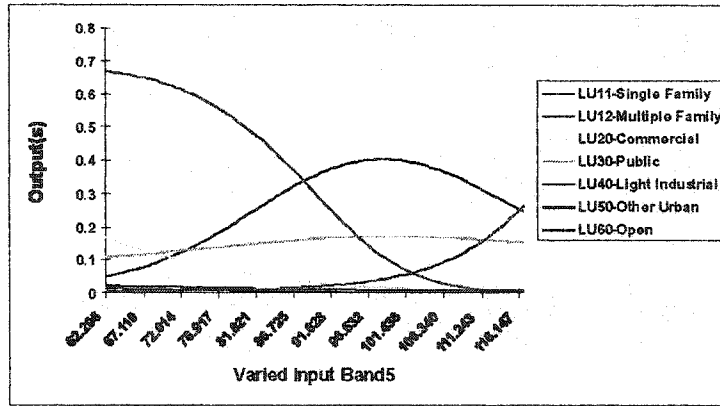


Figure 3.10(e) Network Outputs with Varied Band5

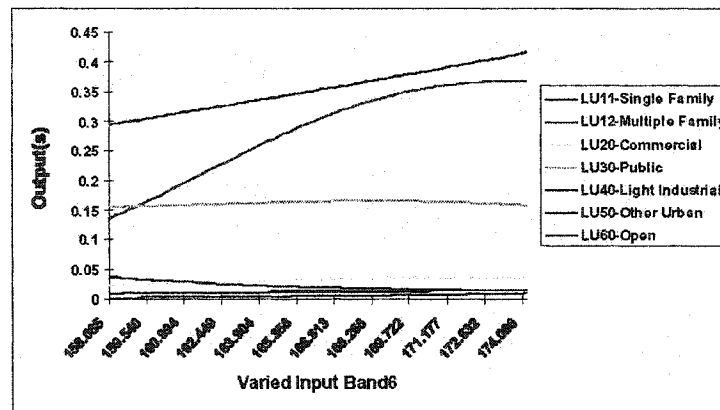


Figure 3.10(f) Network Outputs with Varied Band6

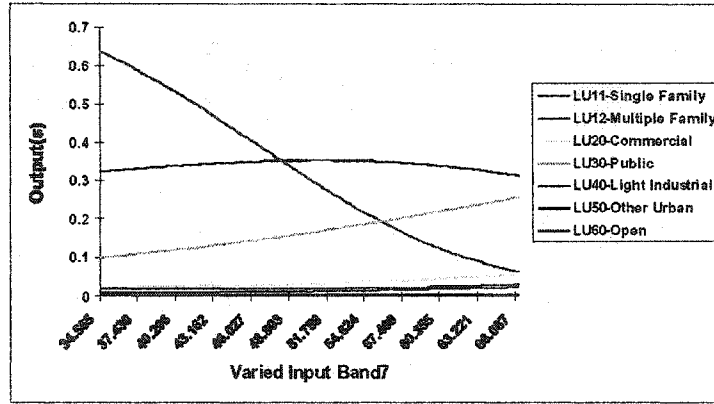


Figure 3.10(g) Network Outputs with Varied Band7

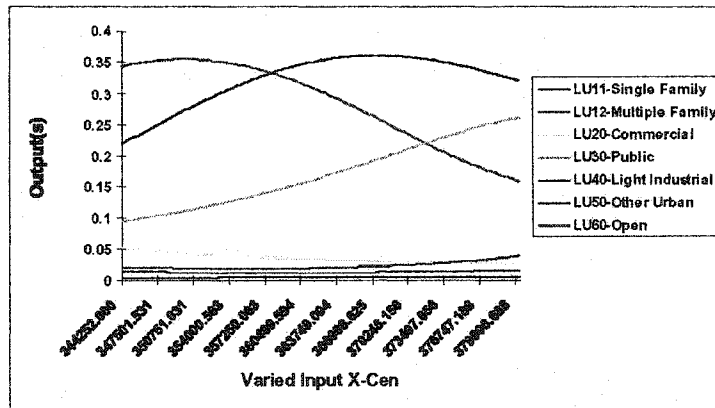


Figure 3.10(h) Network Outputs with Varied X-Cen

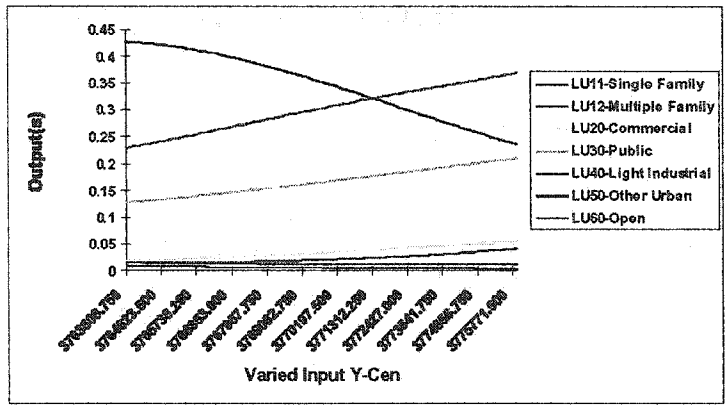


Figure 3.10(i) Network Outputs with Varied Y-Cen

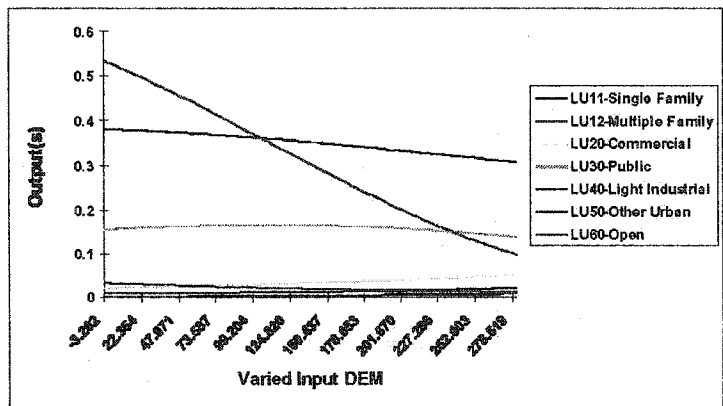


Figure 3.10(j) Network Outputs with Varied DEM

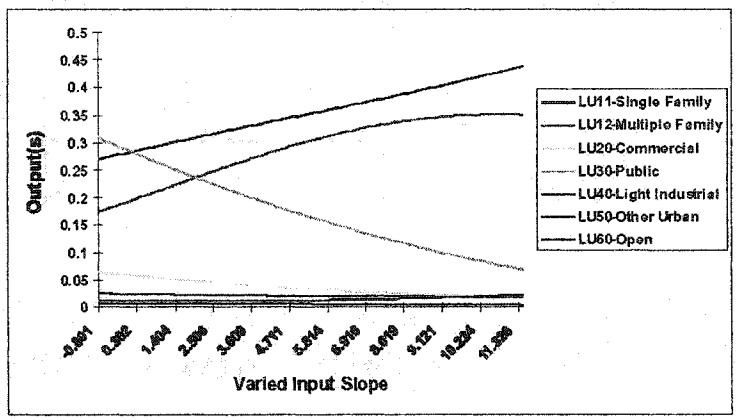


Figure 3.10(k) Network Outputs with Varied Slope

4. Neural Network and GIS to Determine Pixel-level Urban Land

Use for Thematic Mapper Imagery

Hsueh-hwa Lee and Michael K. Stenstrom

Abstract

A study of pixel-level land use classification was performed on the Ballona Wetlands and the surrounding areas in Southern California. The objective was to classify and characterize land use pixels with Landsat Thematic Mapper imagery and other remote sensed data at the pixel level, using both supervised and unsupervised artificial neural network algorithms.

The study area contained 26,614 pixels, each with a 25 x 25 meter resolution. Among them, 2,661 (10%), 6,654 (25%), and 17,299 (65%) pixels were randomly selected as cross validation, testing, and training data for the supervised networks. The results indicate that increasing input data types greatly enhances classification accuracy; the best overall classification accuracy reached 90%, 86%, and 85% for training, cross validation, and testing data sets respectively.

Three unsupervised networks were constructed with four, seven, and nine clusters. These results showed few correlations between clustering and existing land use categories; however, when viewed side by side with high resolution aerial imagery, a different picture emerged: the clusters closely resembled the major land features they were intended to imitate.

4.1 Study Area and Land Use Data Processing

The Ballona Wetlands and the vicinity land use pixels, located between Marina Del Rey and Westchester Bluffs, were the focus of this study (Figure 1). The Wetlands, formed over the last several thousand years, and once encompassing an area of over 2,000 acres, have been degraded to less than 190 acres in the past century, mainly due to urban development. However, as a filter of toxic wastes and pollutants from stormwater runoff reaching the Santa Monica Bay (SMB), the area is still considered the biggest ecosystem of Los Angeles County, and is invaluable to environmental research.

The land use data came from the Southern California Association of Governments (SCAG) [AIS, 1996] and consisted of 271 land use polygons (Table 1) in seven pre-defined land use categories (Figure 2). It was necessary to first break down the current land use polygons into 25 x 25 meter land use pixels (Figure 3), and then spatially overlay the latter with TM pixels of the same resolution. Some of the land use pixels with boundaries of different land use polygons can reflect more than one land use pattern (Figure 3). This heterogeneity can detract from the validity of the experiment. In fact, previous research [Canter, 1997] shows that these non-homogeneous pixels can significantly degrade the performance of supervised neural networks. Thus, data from training pixel statistics, on which network characterizations of each land use category are predicated, must be as homogeneous as possible in order to keep the resultant data consistently valid.

Degradation of the networks arises in two ways: in the probable event that one pixel contains more than one land use pattern, and secondly, in inaccuracies of the representative training data, critical to network learning.

Traditional methods have been unable to remove the heterogeneous pixels and pick up representative training pixels efficiently. In this study, however, all heterogeneous pixels were removed by a GIS spatial operation, and only the remaining homogeneous pixels were used in classification (Figure 4). This is just one example of the ways in which the application of GIS can greatly improve research results. A total of 26,614 homogeneous land use pixels were used their statistics are listed in Table 2. Seven Thematic Mapper spectral bands were used for classification (Figure 5). Each land use pixel can be associated with seven unique pixel values from spectral bands one to seven. These seven pixel values constitute the “spectral signature” of the land use pixels.

4.2 Methodology

This study made use of both supervised and unsupervised learning algorithms. It is important to understand the difference between supervised classification and unsupervised clustering: clustering is the process of grouping sets of input that are spatial neighbors. Classification involves labeling the input data via external criteria [Principe et al., 2000].

In supervised learning, these external samples were used as desired output for specific input. The network was given a learning algorithm to follow and calculate new connection weights that would elicit the desired output. In the unsupervised case, a

learning algorithm was sometimes given, but desired outputs were never given. Because similar units of input cause similar responses, data fed into the unsupervised network got clustered together [Rao et al., 1995]. Both the Multiple Layer Perceptron (MLP) with supervised learning (Figure 6) and the Kohonen Self-Organization Map (SOM) with unsupervised learning algorithms were used here. MLP was mentioned previously in Chapter 3 and SOM is discussed below.

Kohonen SOMs are feed-forward networks that use an unsupervised training algorithm. This process is called self-organization, meaning self-adaptation of a network. The closest possible response to a given input signal is generated and other input then cluster together. The connection weights are modified through different iterations of the network operation, and the network creates the closest possible set of output for the given input [Rao et al., 1995].

The Kohonen SOM output generally is organized in a one- or two-dimensional neighborhood of PEs (Figure 7). The weights between the input and output perform an association between themselves and the input. The PE whose weight vector is closest to the present input wins the competition. This is called competitive learning, which is unsupervised, and extracts information from the input patterns alone, without the need for a desired response. The change in weight vectors can be presented in the following equation.

$$w_i(n+1) = w_i(n) + \Lambda_{i,j^*}(n)\eta(n)(x(n) - w_i(n)) \quad (1)$$

where Λ_{i,i^*} is a neighborhood function centered at the winning PE. Typically, both the neighborhood and the step size change with the iteration number. The neighborhood function Λ is normally a Gaussian function:

$$\Lambda_{i,i^*}(n) = \exp\left(\frac{-d_{i,i^*}^2}{2\sigma^2(n)}\right) \quad (2)$$

The Kohonen SOM is able to preserve the structure of the input space relatively well. The number of PEs is chosen experimentally. The number of output PEs affects the accuracy of the mapping and the training time. Increasing the number of PEs increases the resolution of the map, but also dramatically increases the training time.

4.3 Results and Discussion

4.3.1 Network Construction

This study constructed three MLPs with a back-propagation learning algorithm and three Kohonen SOM networks. Table 3 lists all the input parameters. The coordinate data calculated geo-spatially with GIS were used as ancillary input. The characteristics of each MLP network with different input data are listed in Table 4. A total of 2,661 (10%), 6,654 (25%), and 17,299 (65%) among 26,614 land use pixels were randomly chosen as cross validation, testing, and training data (Figure 8). For the unsupervised learning, 3 SOMs with various numbers of output clusters were built (Table 5).

4.3.2 Classification Results

The MSEs of training and cross validation data are listed in Table 6. The training results show that both training and cross validation MSEs decreased as more input information was provided to the MLP. However, the real performance of a MLP network is based on the accuracy of classification, not only the MSEs. The classification results (percent accuracy) for each land use categories of training, cross validation, and testing are listed in Table 7(a), (b) and (c). Correlation coefficients, r , another parameter to evaluate the network performance, are listed in Table 8. As noted above, the classification accuracy increases--from MLP1 to MLP3--as the networks receive more input data. Thus, the incorporation of coordinate information with pixel values as inputs to MLP3 improves the network performance significantly.

Confusion Matrices, which compare expected classification results with the network output, are presented in Table 9. Figure 9 depicts the confusion matrix of P-MLP3 network training data classification in percentages. A perfect classification gives a confusion matrix with only the diagonal populated and all other entries zero. The classifications of Commercial, Public, and Light Industrial land use pixels underperformed that in other categories. These three land use categories also happened to have the least number of pixels included in the study area. Thus the network weights are trained favorably to match other categories due to the lack of training pixels of these three.

It should also be noted that the introduction of coordinate information to MLP3 greatly enhanced the classification accuracy. This information, though not directly

related to the SCAG land use categories, provided neighborhood information to the network. The neighborhood information improved the classification process by adding the land use information of adjacent pixels to the network, and it can be rationalized that a pixel tends to fall into the same land use categories with those of its surrounding training pixels.

Figures 10, 11, and 12 present the unsupervised clustering of P-SOM1 (four clusters), P-SOM2 (seven clusters), and P-SOM3 (nine clusters), together with the existing SCAG land use polygons. Clearly, the clustering processes have reached a certain degree of homogeneity inside each land use polygon. Table 10 lists the confusion matrices of P-SOM1, 2, and 3 for the SCAG land use pixels. In Figure 13, a comparison of P-SOM1 clustering and an aerial photo are presented side-by-side. Though no obvious correlations could be drawn from Table 10's confusion matrices, almost all major land features in the aerial photo were somewhat represented in the clustering process from landsat pixels.

4.4 Conclusions and Future Research

The results in this paper confirm that ANNs and GIS together with remote sensed data can be valuable tools for enhancing pixel-level land use classifications. The MLP performance improved as more input data types and samples were provided to the networks. Once again, GIS proved a great tool in the analysis of spatial data as well as in the visual presentation of the results.

As noted in Chapter 3, there is some inherent uncertainty and fuzziness in the current SCAG land use definition. The definition of land use is very generic in its nature. For instance, two spectrally identical pixels might be categorized into two different categories due to socioeconomic variance. Thus, the spectral signature concept might be useful in classifying static natural or artificial land features. However, it cannot describe the underlying human activities used to define some of the land use categories. This paper also performed unsupervised SOM analysis, which, although it appeared to have few associations with the SCAG land use categories, depicted the land features vividly.

The planned future expansion of this study involves methodology described in this paper: better imagery, and a different approach to defining the existing land use definition. Recent improvements in sensor technology, computation speed, and processing algorithms have drastically increased the availability of the imagery. More affordable and higher resolution datasets are now accessible to the general public due to a very competitive remote sensing marketplace. For example, 1-meter resolution aerial imagery for the whole Los Angeles County is available for around \$10,000 --a very affordable cost. Better resolution definitely leads to better recognition. However, more computing resources are required as 1-meter imagery requires 625 times as much storage space as the 25-meter landsat TM imagery does.

Future study should also involve more than just correlating the input to the generic SCAG land use; the spectral signature information, along with other socioeconomic information should be used to quantitatively define a new land use. The current land use categories and boundaries are significantly influenced by census

tracts/blocks, and the assessor's parcel layer. It is believed that a new spectral pixel based land use definition will better serve hydrological modeling and environmental monitoring. However, more research activities of integrating spectral signatures with hydrological and environmental models need to be conducted in order to evaluate the correlations of spectral signatures and desired model outputs.

4.5 References

- AIS (1996). *Southern California 1990 Aerial Land Use Study: Land Use Level III/IV Classification*. Aerial Information Systems, Redlands, California.
- Principe, C. P., Euliano, N. R., and Lefebvre, W. C. (2000). *Neural and Adaptive Systems: Fundamentals Through Simulations*. John Wiley & Sons, Inc., New York.
- Rao, V. B. and Rao, H. V. (1995). *C++ Neural Networks & Fuzzy Logic*. MIS Press. New York, NY. *IEEE Transactions on Geoscience and Remote Sensing*, 33(3): 562–578.

Table 1 Summary of Land Use Polygons

Land Use Category	Code	Number of Polygons
Single Family	11	23
Multiple Family	12	63
Commercial	20	52
Public	30	31
Light Industrial	40	12
Other Urban	50	16
Open	60	74
Total		271

Table 2 Summary of Land Use Pixels

Land Use Category	Code	Number of Pixels
Single Family	11	7,589
Multiple Family	12	2,560
Commercial	20	1,733
Public	30	1,436
Light Industrial	40	1,213
Other Urban	50	2,955
Open	60	9,128
Total		26,614

Table-3 Summary of Input Data Used

Input Code	Description	Unit	Range
Band1	Pixel Value of TM* Spectral Band 1	Dimensionless	5 – 255
Band2	Pixel Value of TM Spectral Band 2	Dimensionless	0-255
Band3	Pixel Value of TM Spectral Band 3	Dimensionless	0-255
Band4	Pixel Value of TM Spectral Band 4	Dimensionless	0-255
Band5	Pixel Value of TM Spectral Band 5	Dimensionless	0-255
Band6	Pixel Value of TM Spectral Band 6	Dimensionless	134-190
Band7	Pixel Value of TM Spectral Band 7	Dimensionless	0-255
X-Cen	The X Coordinate of Centroid ** in a Pixel	Meter	365,234 – 368,534
Y-Cen	The Y Coordinate of Centroid ** in a Pixel	Meter	3,756,968 – 3,762,625

* Landsat Thematic Mapper

** Based on NAD1983 UTM (Zone 11N) Projection with Units (Meters)

Table-4 Lists of MLP Networks

Network	No. of Input Layer PEs	No. of Hidden Layer PEs	Network Inputs
P-MLP1	3	7	Band1, Band2, Band3
P-MLP2	7	14	Band1, Band2, Band3, Band4, Band5, Band6, Band7
P-MLP3	9	18	Band1, Band2, Band3, Band4, Band5, Band6, Band7, X-Cen, Y-Cen

Table-5 Lists of SOM Networks

Network	No. of Input Layer PEs	No. of Output Clusters	Network Inputs
P-SOM1	3	7	Band1, Band2, Band3, Band4, Band5, Band6, Band7
P-SOM2	7	14	Band1, Band2, Band3, Band4, Band5, Band6, Band7
P-SOM3	9	18	Band1, Band2, Band3, Band4, Band5, Band6, Band7

Table-6 MLP Network Mean Square Errors

Network	Training Data			Cross Validation		
	Epoc No.*	Minimum MSE	Final MSE	Epoc No.*	Minimum MSE	Final MSE
P-MLP1	1000	0.208	0.208	1000	0.208	0.208
P-MLP2	1000	0.125	0.125	1000	0.127	0.127
P-MLP3	1000	0.088	0.088	1000	0.085	0.085

*Epoch when MSE is at its minimum

Table-7(a) Testing Results on Training Data

Network	Land Use Categories (% Accuracy of Training Data)						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
P-MLP1	73	65	70	73	70	75	75
P-MLP2	85	78	76	76	75	82	92
P-MLP3	92	85	81	82	80	88	96

Table-7(b) Testing Results on Cross Validation Data

Network	Land Use Categories (% Accuracy of Training Data)						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
P-MLP1	70	64	70	69	69	73	75
P-MLP2	82	72	73	73	75	80	85
P-MLP3	90	80	78	80	76	85	90

Table-7(c) Testing Results on Testing Data

Network	Land Use Categories (% Accuracy of Training Data)						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
P-MLP1	69	67	70	71	63	70	72
P-MLP2	81	73	71	76	76	78	86
P-MLP3	91	85	79	68	78	87	92

Table-8(a) Correlation of Training Data

Network	r^2						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
P-MLP1	0.55	0.48	0.47	0.31	0.35	0.51	0.58
P-MLP2	0.72	0.70	0.68	0.78	0.75	0.69	0.79
P-MLP3	0.77	0.77	0.82	0.88	0.91	0.83	0.92

Table-8(b) Correlation of Cross Validation Data

Network	r^2						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
P-MLP1	0.53	0.44	0.42	0.32	0.38	0.60	0.55
P-MLP2	0.75	0.73	0.65	0.75	0.77	0.70	0.75
P-MLP3	0.75	0.75	0.80	0.87	0.88	0.81	0.89

Table-8(c) Correlation of Testing Data

Network	r^2						
	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open
P-MLP1	0.54	0.46	0.41	0.35	0.35	0.56	0.58
P-MLP2	0.69	0.70	0.66	0.70	0.70	0.65	0.78
P-MLP3	0.73	0.72	0.80	0.84	0.88	0.78	0.86

Table-9(a) Confusion Matrix of Training Data (P-MLP3)

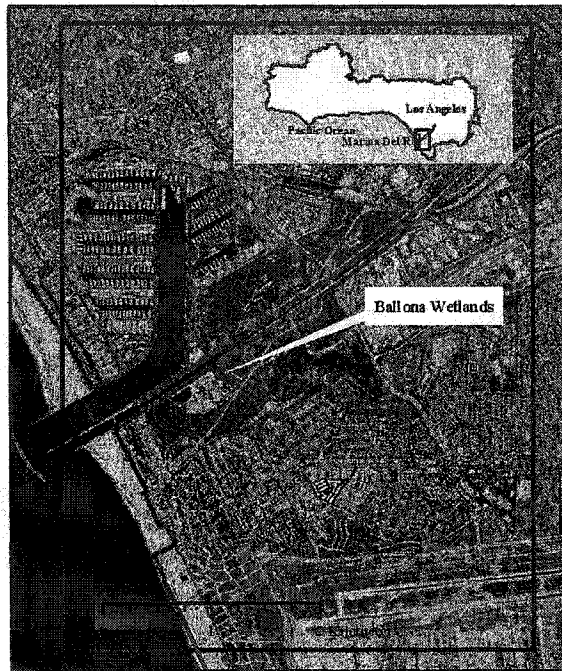
Predict / Desired	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open	Total
LU11-Single Family	4,558	142	40	30	84	59	117	5,030
LU12-Multiple Family	205	1,406	20	19	2	10	11	1,673
LU20-Commercial	36	30	902	32	39	27	12	1,078
LU30-Public	50	37	30	769	13	7	20	926
LU40-Light Industrial	8	19	81	28	632	27	18	813
LU50-Other Urban	0	0	20	4	0	1,673	60	1,757
LU60-Open	97	20	20	56	20	98	5,711	6,022
Total	4,954	1,654	1,113	938	790	1,901	5,949	17,299

Table-9(b) Confusion Matrix of Cross Validation Data (P-MLP3)

Predict / Desired	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open	Total
LU11-Single Family	722	27	10	11	12	12	20	814
LU12-Multiple Family	55	213	8	4	1	1	15	297
LU20-Commercial	10	14	126	4	6	4	14	178
LU30-Public	6	7	7	106	1	1	15	143
LU40-Light Industrial	0	2	4	4	94	5	1	110
LU50-Other Urban	0	0	0	1	0	263	22	286
LU60-Open	9	2	7	3	10	23	779	833
Total	802	265	162	133	124	309	866	2,661

Table-9(c) Confusion Matrix of Testing Data (P-MLP3)

Predict / Desired	LU11- Single Family	LU12- Multiple Family	LU20- Commercial	LU30- Public	LU40- Light Industrial	LU50- Other Urban	LU60- Open	Total
LU11-Single Family	1,668	39	19	145	13	21	42	1,947
LU12-Multiple Family	80	545	3	7	1	2	35	673
LU20-Commercial	18	17	362	10	4	17	33	461
LU30-Public	27	16	13	137	6	2	40	241
LU40-Light Industrial	0	11	29	22	233	10	2	307
LU50-Other Urban	0	0	0	2	0	646	25	673
LU60-Open	40	13	32	42	42	47	2,136	2,352
Total	1,833	641	458	365	299	745	2,313	6,654



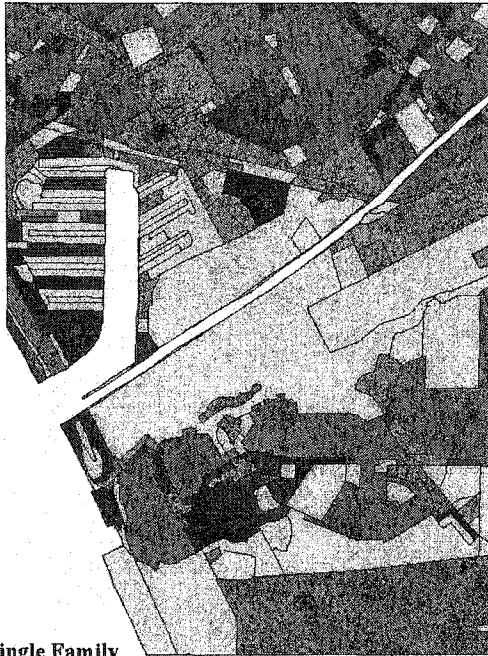
Legend

- Santa Monica Bay
- Study Area



Figure 1
Marina Del Rey and Santa Monica Bay

0 0.5 1 2 Kilometers










-  Single Family
-  Multiple Family
-  Commercial
-  Public
-  Light Industrial
-  Other Urban
-  Open



Figure 2
Land Use Polygons

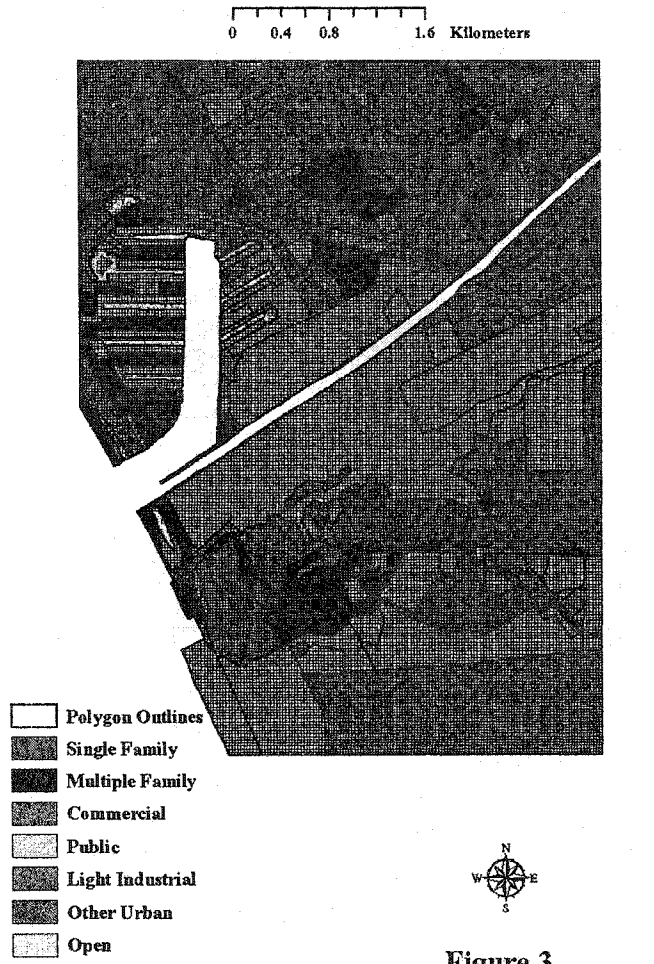
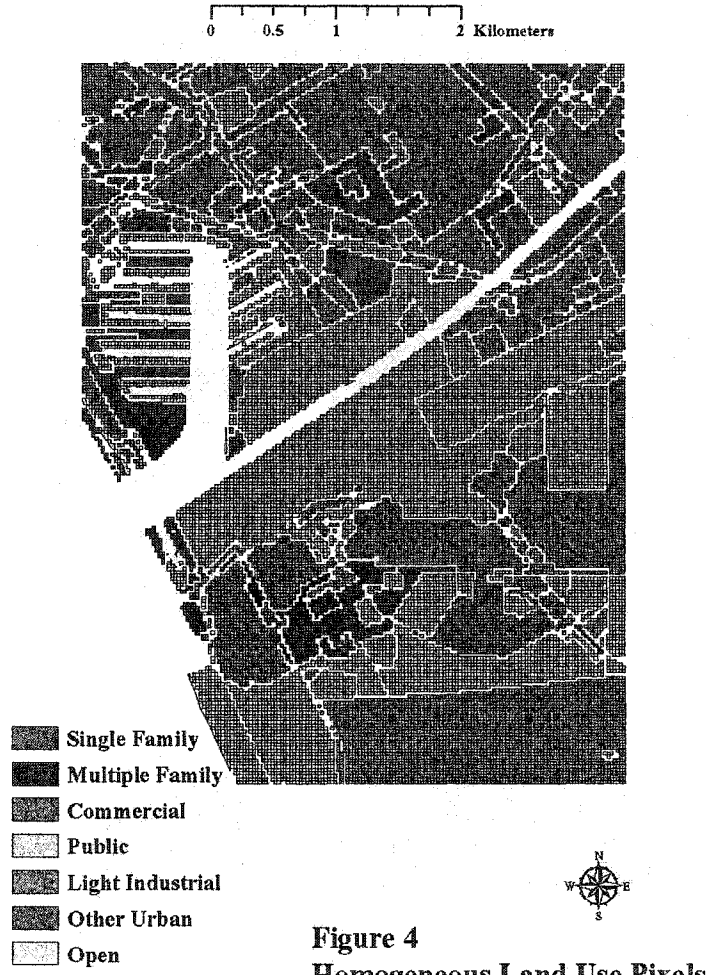


Figure 3
Land Use Pixels



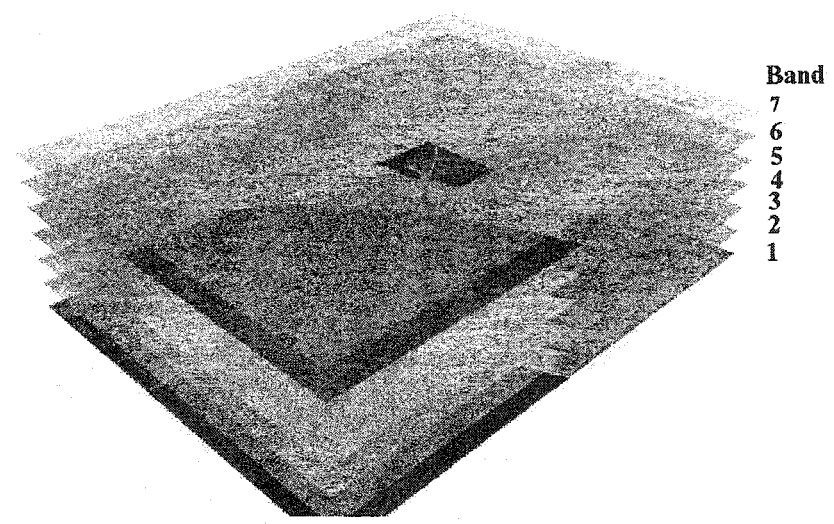


Figure 5
Study Area and Seven Thematic Mapper Spectral Bands

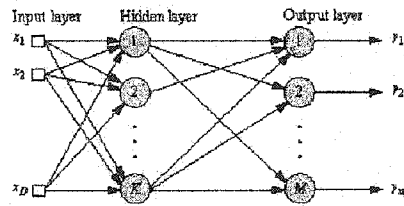


Figure 6 Multiple Layer Perceptron Network

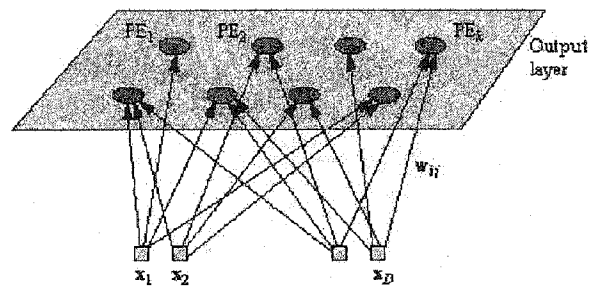
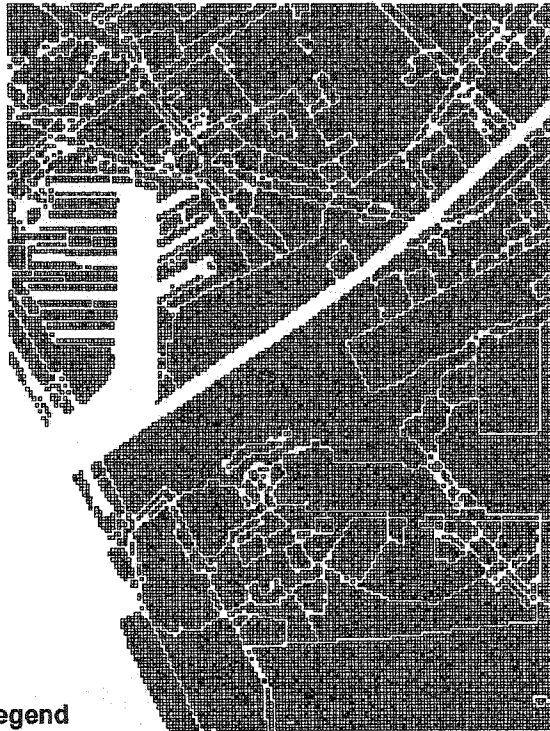


Figure 7 Kohonen Self Organization Map Network

0 0.45 0.9 1.8 Kilometers



Legend




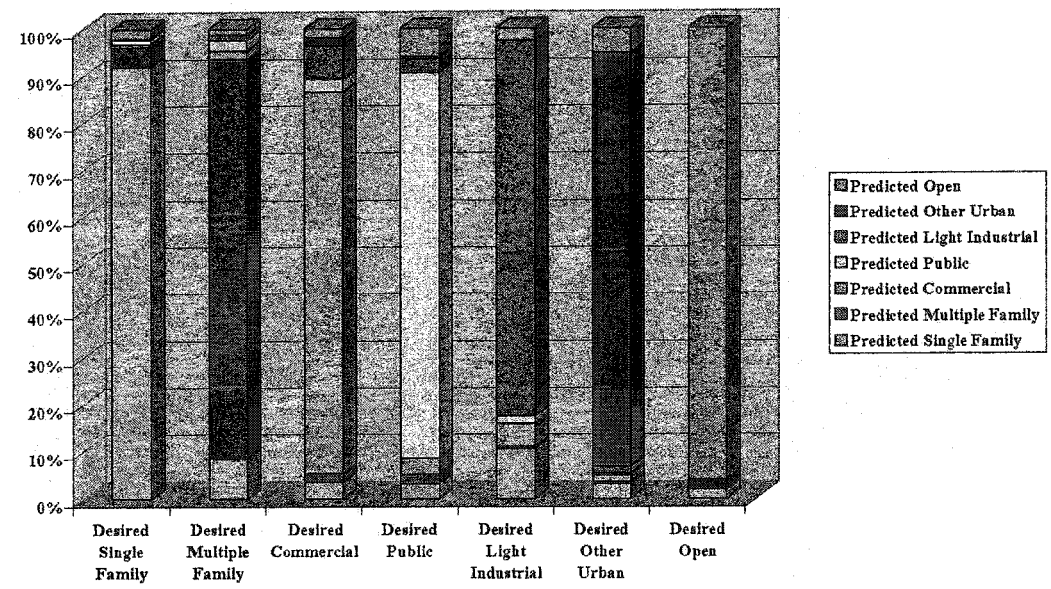
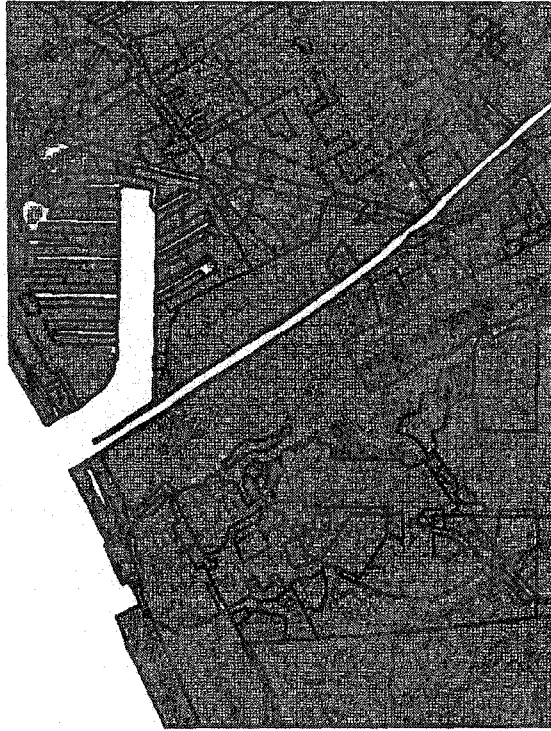
-  Cross Validation (2,661)
-  Testing (6,654)
-  Training (17,299)



Figure 8
MLP Networks Input Data Type

Figure 9 ANN Predicted Outputs vs. Desired Results for P-MLP3 Training Data





Legend

 Land Use Polygon Outline

Land Use Pixel Clustering

 A (7,191)

 B (5,467)

 C (6,985)

 D (6,971)



Figure 10
SOM-1 Network: Four Clusters



Legend

 Land Use Polygon Outline

Land Use Pixel Clustering








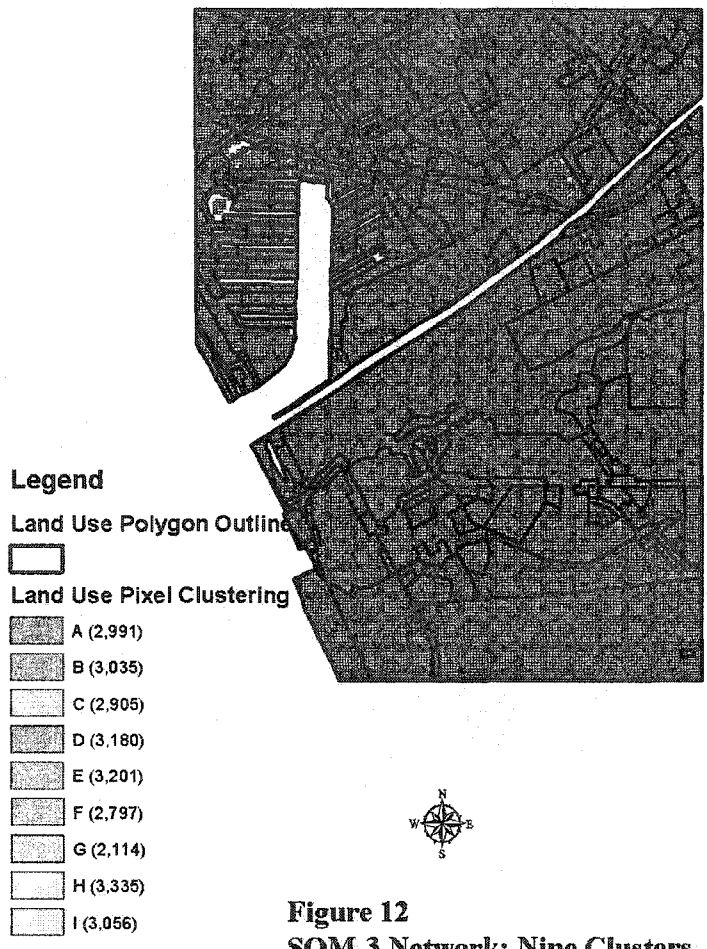
-  A (4,274)
-  B (3,984)
-  C (3,849)
-  D (2,834)
-  E (4,043)
-  F (3,756)
-  G (3,874)



Figure 11
SOM-2 Network: Seven Clusters



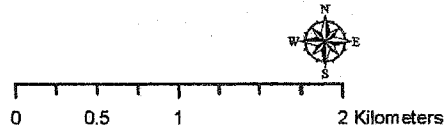
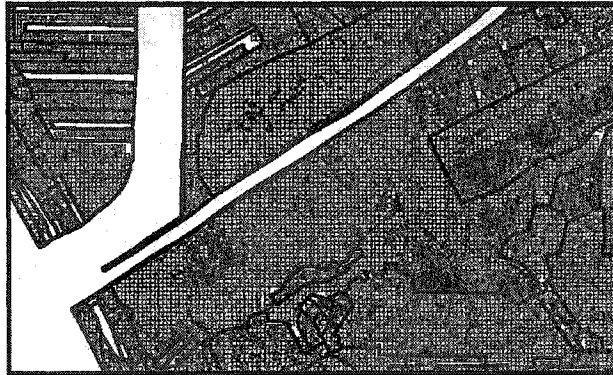
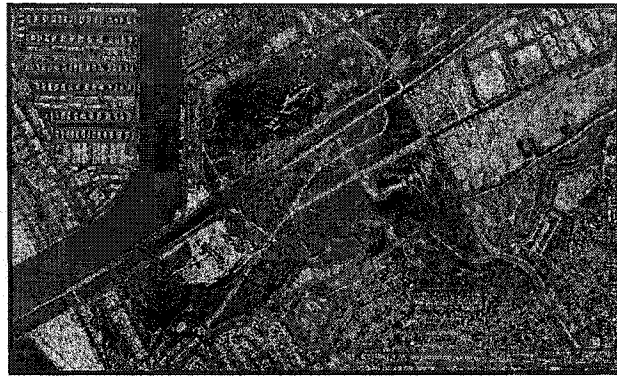


Figure 13
Aerial Photo and SOM-1 Clustering

5. Stormwater Runoff Simulation in Malibu Creek Watershed Using a Deterministic Hydrological Model and Artificial Neural Networks

Hsueh-hwa Lee and Michael K. Stenstrom

Abstract

This paper makes use of a deterministic hydrologic model and artificial neural networks (ANN) to simulate the average daily rainfall-runoff processes of the Malibu Creek Watershed (MCW). The introduction of GIS as a spatial data management and analyzing tool had the effect of making the spatially distributed parameters required by this model more accurate and representative than ever before. The result was a successful prediction of the general trend of daily discharge hydrographs after the model was calibrated with observed discharge gauging data from Los Angeles County Department of Public Works (LADPW).

Additionally, with only precipitation data as input, several ANNs with different architectures were constructed and trained to simulate solely wet-weather daily discharges. This simulation demonstrated a high correlation on training data sets, and an acceptable correlation on testing data sets. Though limited to wet-weather conditions, it presented a performance comparable to that of the deterministic model.

The data layers collected at the site and the results they elicited offer insight into the complexities of MCW. The computing methodology and data management models developed in this study could be readily implemented in any other major watersheds.

5.1 Introduction

Modeling stormwater runoff quantity is integral to watershed management, the designing of flood control facilities, and regional planning. The response of a watershed to precipitation is influenced by many heterogeneously distributed parameters [Zhang, 2000], such as geomorphic features (topology, vegetation, and soil type) and climatic parameters (precipitation, temperature, etc.). There have been a great number of successful applications in the field of hydrological modeling in recent times, including both traditional hydrological modeling and soft computing methodologies. Most traditional models relate temporal and spatial precipitation history to stream flowrate data at a specific point of interest, and have been performed on different scales.

Different scales require different types of models since it is well known that the nature of hydrological process varies greatly along these scales. They include *micro-scale*, ranging from one cm^2 to one km^2 , *meso-scale* from one km^2 to 100,000 km^2 , macro-scale from 100,000 km^2 up to global scale [Becker et al., 1987]. Most watershed-wide models are in meso-scale and are deterministic. A *deterministic* hydrological model describes the cause-effect relations stemming from the known features of physical system under study, and thus the modeling processes are considered free of random variation. The structure of a scientifically based deterministic hydrological model should be independent with spatial variance, and only the model parameters should be calibrated according to different study areas [Schultz, 1994].

Since the 1990s, the geographic information system (GIS) has become a powerful tool for hydrological modeling. GIS is particularly suited for meso-scale models because

of the spatial resolutions as well as its consistency in stream and watershed delineation using USGS Digital Elevation Models (DEMs) [Maidment, 2002]. Furthermore, GIS can be used to assemble the crucial water resource information required by deterministic models: land use, soil cover, gauging station, and other climatic variables. Likewise, soft computing using artificial neural networks (ANN) has also emerged as a significant tool for hydrological modeling in the past decade since is best suited to describe the complex and nonlinear nature of precipitation-runoff processes.

There is a large number of publications recently concerning ANN applications of rainfall-runoff processes [Abrahart et al., 1997, 2000; Dawson et al., 1998; Lachtermacher et al., 1994; Mason et al., 1996; Smith et al., 1995; Zhang et al., 2000]. Most of these studies, however, assume a uniform distribution of precipitation over the study area based on insufficient data sets of watershed studies, and use them as network input. Conversely, the objective of this project is to collect, quality-check, and integrate extensive spatial data sets with GIS, and then perform both traditional deterministic hydrological modeling and ANN simulation on the Malibu Creek Watershed (MCW).

After brief discussions of the background theories, the procedures used to preprocess the deterministic model inputs are discussed in detail. Afterward, a comparison of the results of both approaches is presented. Finally, based on the model output, conclusions are made along with the implications for the future study.

5.2 Watershed Description

The Malibu Creek Watershed (MCW), selected as the study area, is significant as one of the largest discrete watersheds draining to the Santa Monica Bay. MCW encompasses approximately 109 square miles, and is located in the northwestern end of Los Angeles County and the southeastern end of Ventura County (Figure 5.1). It has a Mediterranean-type climate: dry summer and moist winter periods [PCR, 2001]. The watershed is comprised of all or parts of the cities of Agoura Hills, Calabasas, Malibu, Thousand Oaks, Westlake Village, unincorporated Los Angeles County, and Ventura County.

5.3 Deterministic Hydrological Modeling

This study used the HEC Hydrologic Modeling System (HEC-HMS), developed by the U.S. Army Corps of Engineers Hydrologic Engineering Center (HEC). HEC-HMS is a computer program that simulates both natural and controlled precipitation-runoff and routing processes [USACE(a), 2000]. The followings are some of the HEC-HMS components used to simulated rainfall-runoff process of MWC in this study:

- HEC Geospatial Hydrologic Modeling Extension (HEC-GeoHMS) was used to preprocess GIS data and delineate the watershed.
- Meteorologic components were used to store precipitation and discharge data, and to properly represent spatial variations of precipitation.
- Loss models were used to estimate the volume of runoff.

- Direct runoff models were used to estimate stream daily discharge.
- Calibration component was used to calibrate model output based on observed discharge gauging data.

All the models used in HMS were deterministic. Figure 5.2 presents the schematics of the HEC-HMS representation and Table 5.1 lists the major parameters used in this study.

5.3.1 HEC Geo-HMS Processing

HEC-GeoHMS is a GIS module used to develop a number of hydrologic modeling input. HEC-GeoHMS analyzes digital terrain information and transforms the drainage paths and watershed boundaries into a hydrologic data structure that represents the watershed response to precipitation. This GIS module featured terrain-preprocessing capabilities to construct a hydrologic schematic of the watershed at stream gages, hydraulic structures, and other control points. The hydrologic results from HEC-GeoHMS were then imported by the Hydrologic Modeling System, HEC-HMS, at which point the simulation is performed [USACE(b), 2000].

In order to reach maximum modeling accuracy, the USGS 10-meter Digital Elevation Model (DEM) was used in this paper as Geo-HMS terrain input data. The USGS DEM data files are digital representations of cartographic information in a raster format, and consist of a sampled array of elevations for a number of ground positions at regularly spaced intervals. The 7.5- and 15-minute DEMs were included in the large-

scale category while 2-arc-second DEMs fell within the intermediate scale category, and 1-degree DEMs fell within the small-scale category.

A master DEM required for this study was created by spatially merging several units of USGS 24K quads from the topographic quadrangle map series for all of the United States and its territories (Figure 5.3). It contains all of ten 24K quads: Calabasas, Camarillo, Canoga Park, Malibu Beach, Newbury Park, Point Dume, Point Mugu, Thousand Oaks, Topanga, and Triunfo Pass. The DEM terrain data was analyzed and processed to determine flow path and watershed delineation. Figure 5.4 schematically presents the HEC-GeoHMS processing of MCW.

The HEC-GeoHMS processing has defined 17 sub-watersheds using GeoHMS processing (Figure 5.5). Table 5.2 lists the major parameters calculated for each sub-watershed.

5.3.2 Meteorologic Model

This study used historical daily precipitation data from Los Angeles County Department of Public Works (LADPW) to simulate MCW runoff of five seasons, 1997 to 2001. It should be noted that in this study a specific season is defined as October 1 to September 30. Thus, Season 1997 starts on October 1, 1996, and ends on September 30, 1997. Five precipitation stations (Table 5.3) were selected to provide model input because they are geographically closer to the study area and they contain sufficient data sets to cover the time frame.

This paper adopted *Inverse-distance-square method*, which relies on the notation of “nodes” positioned within a watershed to calculate adequate spatial resolution of precipitation. The node of each sub-watershed can be represented by the centroid. Weight factors were computed and assigned to the gauges in inverse proportion to the square of the distance from the centroid to the gauge. The weighting factors could be calculated by:

$$w_{ij} = \frac{1}{d_{ij}^2} \bigg/ \sum_{k=1}^N \frac{1}{d_{ik}^2} \quad (\text{Equation 1})$$

in which w_{ij} = the weighting factor of the j th gauge to the centroid of the i th sub-watershed; d_{ij} = the distance of the j th precipitation gauge to the centroid of the i th sub-watershed; N = total number of precipitation gauges used in the calculation. The node hyetograph at time t could then be calculated as:

$$P_i(t) = \sum_{k=1}^N w_{ik} P_k(t) \quad (\text{Equation 2})$$

where $P_i(t)$ = the precipitation of sub-watershed i at time t ; w_{ik} = the weighting factor of the k th gauge to the i th sub-watershed; $P_k(t)$ = precipitation measured at gauge k at time t .

There are a total of five precipitation gauges and seventeen centroids (Figure 5.6) for all sub-watersheds. Therefore, 75 separate distances between precipitation stations and sub-watershed centroids of Equation 5.1 were calculated first (Table 5.4) in order to determine the weight factor of each station to each specific sub-watershed. Figure 5.7

shows the schematics of deriving a synthetic hyetograph of sub-watershed No. 17 from hyetographs of all precipitation gauges using the Inverse-distance-square method.

5.3.3 Loss Model

The *Initial and Constant-rate Loss Model* was used in this study. In this model, it is assumed that the maximum potential precipitation rate, f_c (inch/hour), is constant throughout an event, and initial loss, I_a (inch), is added to the model in order to reflect interception and depression loss. The model considers that only precipitation on the pervious surface area is subject to loss. Thus, the model requires three sets of model input: f_c , I_a , and imperviousness (*Imp* in %).

The constant loss rate, f_c , can be considered the ultimate infiltration capacity of the soil. The U.S. Department of Agriculture Natural Resources Conservation Service (NRCS), (formerly the Soil Conservation Service (SCS)), classifies soil based on infiltration capacity [SCS, 1986]; Table 5.5 lists the estimated infiltration corresponding to SCS categories [Skaggs et al., 1982]. The State Soil Geographic (STATSGO) Data Base of NRCS [NRCS, 1994] was used to calculate f_c .

Figure 5.8 shows the STATSGO GIS soil map labeled with *Map Unit ID* (MUID) of MCW. In STATSGO, each map unit can have multiple components and each component can have multiple layers. In order to calculate the average f_c of each sub-watershed, the SCS soil categories in Table 5.5 must be quantified. The STATSGO layer was thus rasterized to a surface with the constant loss rate as the pixel value; a surface calculation was subsequently performed to calculate the average constant loss rate of

each sub-watershed (Figure 5.9). Similarly, a surface of the average imperviousness (Figure 5.10) for each sub-watershed can be calculated by using an imperviousness GIS layer from Los Angeles County Department of Public Works [LADPW, 1991].

I_a , initial loss, represents the maximum precipitation depth falling on the watershed without generating runoff. It is influenced by the degree of saturation of the watershed, watershed terrain, land use, soil type, and soil treatment. USACE (1994) suggests that 0.1-0.2 inches be used. This study adopted 0.1 inch initial loss rate, and later used it as a calibration parameter.

5.3.4 Direct Runoff Model

This paper employed *kinematic-wave* model to simulate overland flow. The kinematic-wave model represents the watershed with an open channel, and can be applied to the equations that simulate unsteady shallow flow in an open channel [USACE, 1970]. A kinematic-wave watershed is comprised of the following four components: overland flow planes, sub-collector channels, collector channels, and the main channel. This study installed the minimum configuration, which included one overland flow plane and the main channel for each sub-watershed. The parameters required for main channel were calculated by HEC-GeoHMS.

In overland flow planes, all three required parameters, *typical length*, *representative slope*, and *overland-flow roughness coefficients*, can be elicited by GIS functions. In order to calculate *typical length*, first a distance surface to the longest flow channel in GIS grid format was generated, and GIS zonal functions were used afterward

to calculate the average distance to longest flow channel of each sub-watershed (Figure 5.11).

Representative slope can be calculated by averaging the slope surface derived from the USGS 10-meter DEM for each sub-watershed (Figure 5.12). The overland-flow roughness coefficient is influenced solely by over-land surface types, and USACE (1998) suggests that it be estimated by using Table 5.6. However, due to the unavailability of geographic data using the USACE surface definitions in Table 5.6, a correlation between the existing land use layer definition from the Southern California Association of Governments and the surface description of Table 5.6 had to be established to calculate the average roughness of each sub-watershed. Table 5.7 assumes the correlation between the corresponding USACE overland flow surface categories and overland roughness for each SCAG land use definition. Again, a surface of roughness was created and the average roughness of each sub-watershed was then calculated (Figure 5.13).

5.3.5 Model Calibration

The goal of HEC-HMS calibration is to identify reasonable model parameters that generate the best fit with observed hydro-meteorological data. Only one discharge gauge, LADPW stream gauging station F-130R, named “Malibu Creek below Cold Creek,” is available in the study area. This gauge is located 0.2 miles downstream of Cold Creek, and 4.5 miles upstream of the MCW outlet (Figure 5.14). The total drainage area to this station is 104.96 square miles--approximately 96% of the whole watershed tributary to

the MCW outlet. Therefore, the observed discharge data of this station was used to calibrate the outflow simulated by HEC-HMS at the outlet.

The HEC-HMS calibration model computes an index of the goodness-of-fit to compare a computed hydrograph to the observed hydrograph. HEC-HMS searches for calibration parameters that yield the best values of an *objective function*. The goal is to minimize the objective functions with reasonable calibration parameters. The objective function used in this study is the *Peak-weighted Mean Square Error* method [USACE, 1998]. The *Peak-weighted Mean Square Error* objective function can be represented by:

$$Z_{MSE} = \left\{ \frac{1}{NQ} \left[\sum_{i=1}^{NQ} (q_o(i) - q_s(i))^2 \left(\frac{q_o(i) + q_o(mean)}{2q_o(mean)} \right) \right] \right\}^{1/2} \quad (\text{Equation 3})$$

in which Z_{MSE} = objective function of the *Peak-weighted Mean Square Error* method; NQ = number of computed hydrograph ordinates; $q_o(t)$ = observed flows; $q_s(t)$ = calculated flows, computed with a selected set of model parameters; $q_o(mean)$ = mean of observed flows. This function compares all ordinates, squaring differences, and it weights the squared differences. This method is an implicit measure of comparison of the peak discharges, runoff volumes, and times of peak of the simulated and observed hydrographs.

HEC-HMS uses the *trial-and-error* approach to find the set(s) of parameters that will minimize the objective function. Initial and Constant loss rates were selected as the calibration parameters in this study. A universal scale factor of each parameter was used to calculate the error. HEC-HMS changes the parameters and reiterates the process until the error becomes acceptable. *Univariate-gradient search algorithm* was used in this

study. Detailed information of the search algorithm is described in HEC-HMS Technical Reference Manual [USACE, 2000]

5.4 Artificial Neural Network Simulation

The daily runoff hydrograph is affected by many parameters, including precipitation, temperature, land features, but precipitation is the primary influencing variable. In order to obtain the maximum forecasting potential from a limited number of data sets, lags of rainfall data were used to provide short-term recollection of previous events and antecedent conditions. Thus, the daily runoff hydrograph of MCW in this ANN simulation can be symbolically expressed as:

$$Q(t) = f(R_t, R_{t-1}, \dots) \quad (\text{Equation 4})$$

where R_t is the daily rainfall of day “t”, and R_{t-1} is the daily rainfall of day “t-1”.

There were five feedforward multilayer perception (MLP) networks constructed with the backpropagation learning algorithm. Table 5.8 lists the input, output, and data sets assigned to each one of the five MLP networks. The centroid of the whole MCW was calculated and, again, *Inverse-distance-square method* was used to incorporate the five precipitation stations to calculate the effective precipitation for the whole watershed. In total, there were 1,825 (from Oct 1, 1996 to Sep 30, 2001) observed daily discharges from the gauging data. However, due the great variance of dry- and wet-weather flow patterns, only wet-weather daily discharges were studied in the ANN simulation. Figure 5.15 presents all of the 549 daily discharge points selected in the ANN simulation.

5.5 Results

5.5.1 HEC-HMS Simulation

Figure 5.16 presents the hyetograph and hydrographs of the HEC-HMS simulated, calibrated, and observed hydrographs. The simulated (without calibration) runoff discharge hydrograph seemed to follow the general trend of the observed hydrograph from gauging data. However, it is obvious that the simulated peak discharges of precipitation events were off the observed ones quite a bit.

The calibration process greatly enhanced the prediction. As an example, Figure 5.17 plots the hydrograph of the 1998 wet season. It should also be noted that in Figures 5.16 and 5.17, both calibrated and non-calibrated (simulated) HEC-HMS results over-calculated the peak discharge and under-estimated the dry weather base flow. Figure 5.18 plots a time series of residuals, differences between HEC-HMS calibrated and observed flows. Although the HEC-HMS calibrated hydrograph effectively simulated the base flow early in the season, it under-estimated the base flow later in the season. Table 5.9 summarizes the residuals by years and Table 5.10 lists the optimized scale factors of loss rates and the values of final objective functions from the calibration process.

5.5.2 ANN Simulation

Table 5.11 lists the minimum Mean Square Error (MSE) and correlation coefficients, r , of each network. Almost all networks had minimal MSEs, and the training data sets generated reasonable correlation coefficients. However, not all the

networks presented convincing correlations on the testing data sets. In Network Q-MLP1 to Q-MLP4, only one season of daily discharge data was selected as testing data, and the rest of the (four seasonal discharge) data sets were used for network training. In Network Q-MLP5, all training, testing, and cross validation data sets were selected randomly. The correlation coefficient of testing data of Q-MLP5 was the highest. The lower correlations of Q-MLP1 to Q-MLP4 testing data suggest insufficient training data sets and that the networks had not been generalized enough to simulate any specific season of daily discharges.

On the other hand, for Network Q-MLP5, the randomized training data sets vividly generalized network connection weights, and yielded a higher correlation of the testing data sets. Figures 5.19(a) to 5.19(e) compare the simulated and observed daily discharges of all the networks. It should be noted that though randomizing data sets led to a higher correlation coefficient in Q-MLP5, they did not predict peak flow as well as the other networks did. The additions of more antecedent precipitation data to Q-MLP3 and Q-MLP4 did not really improve the network correlation coefficients. In Table 5.12, a sensitivity study conducted on Q-MLP4 training data sets indicated that R_{t-3} , R_{t-4} , and R_{t-5} have little impact in predicting the discharges.

The observed discharge data show that from November 25, 1997 to December 9, 1997, almost all precipitation gauges recorded significant rainfalls, but no corresponding reasonable discharge hydrograph was recorded (Figure 5.20). In order to maintain the integrity of the LADPW monitoring data, the discharge data have not been manually rationalized or altered in any of the HEC-HMS and ANN installations. However, it

should be pointed out that by removing daily discharge points of this potentially problematic period from the training data sets, the correlation coefficient of Q-MLP3 was improved from 0.90 to 0.92, and the correlation coefficient of Q-MLP4 increased from 0.84 to 0.88.

5.5.3 HEC-HMS and ANN Comparisons

Figure 5.21(a) illustrates a comparison of HEC-HMS calibrated simulation and Q-MLP4 network output for all five seasons, 1997 to 2001. Both simulations generally followed the trend of observed discharges. Results for other networks are similar and are not included here for the sake of brevity. Figures 5.21(b) and 5.21(c) show the comparisons of the wet seasons of 1997 and 1998. In Figure 5.21(b), both HEC-HMS and Q-MLP4 simulations predicted the first several major precipitation events before January 1997. The calibrated HEC-HMS simulation outperformed Q-MLP4 in predicting peak discharges of events. However, both simulations underestimated the discharges beginning mid-January 1997. In Figure 5.21(c), while HEC-HMS overestimated several eventful peak discharges, Q-MLP4 provided a better overall replication of the observed hydrograph.

5.6 Conclusions

This study developed a GIS for the Malibu Creek Watershed with multiple layers, once again proving the effectiveness of the GIS in managing hydrological data and leveraging traditional modeling. A full-scale HEC-HMS simulation was conducted using

data collected by GIS, and spatial analyses were performed to delineate and calculate several major model input parameters. 17 distinct sub-watersheds were delineated after HEC-GeoHMS terrain processing. Five seasons of precipitation data from five LADPW precipitation gauges were used to calculate the daily discharge for the entire watershed. One set of LADPW discharge gauge data near Malibu Creek outlet was used for calibration, and detailed hydrological analysis results were obtained for all for all of the 17 sub-watersheds. The established GIS database and methodology of GIS grid (raster)-based calculation provide a great resource for future studies of MCW.

Five Artificial Neural Networks with the backpropagation learning algorithm were constructed as alternatives of HEC-HMS to simulate only wet-weather daily discharges. Instead of using all the parameters required by HEC-HMS, only precipitation data were used for ANN simulation. The results indicated that for certain rainfall events, ANN simulation could perform at least as well as the HEC-HMS simulation.

This study also addressed the critical issue of source data quality and how it influences project results. Faulty observations have the potential to deteriorate the calibration and training process. More historical meteorological data will greatly enhance the HEC-HMS calibration and ANN training. The modeling effort itself will not lead to a better understanding of the MCW precipitation-runoff process; it is the quality of input data that establish the bottom line of performance. Thus, in addition to improvement in computing methodologies, future researches should incorporate even more observation data obtained from an ever better monitoring approach.

5.7 References

- Abrahart, R. J., and Kneale, P. E. (1997). "Exploring Neural Network Rainfall-Runoff Modeling", *Proceedings Sixth National Hydrology Symposium, University of Salford*, 15 – 18 September 197, 9.35 – 9.44.
- Abrahart, R. J., and Kneale, P. E. (2000). "Comparing Neural Network and Autoregressive Moving Average Techniques for the Provision of Continuous River Flow Forecasts in Two Contrasting Catchments", *Hydrological Processes*, Vol. 14, pp. 2157-2172.
- Becker, A. and Nemeč, J. (1987). "Macroscale Hydrologic Models in Support to Climate Study." In: *The Influence of Climate Change and Climatic Variability on the Hydrologic Regime and Water Resources* ed. S. I. Solomon, M. Bevan, and W. Hogg. 431-445. IAHS Publ. no. 168.
- Dawson C. W., and Wilby, R. L. (1998). "An Artificial Neural Network Approach to Rainfall-runoff Modeling". *Hydrological Sciences Journal*, Vol. 43, pp. 47 – 66.
- Lachtermacher, G., and Fuller, J. D. (1994). "Backpropagation in Hydrological Time Seriesforecasting". In: Hipel, K. W. et al. (eds) *Stochastic and Statistical Methods in Hydrology and Environmental Engineering*, Vol. 3, pp. 229 – 242.
- LADPW (1991). *Hydrology Manual*. Alhambra, CA.
- Maidment, D. (2002). *Arc Hydro: GIS for Water Resources*. ESRI, Redlands, CA.

- Mason, J. C., Price, R. K., and Tem'ME, A. (1996). "A Neural Network Model of Rainfall-runoff Using Radial Basis Functions". *Journal of Hydraulic Research*, Vol. 34, 1996, No. 4, pp. 537 – 548.
- NRCS (1994). *State Soil Geographic (STATSGO) Data Base: Date Use Information, Publication Number 1492*. Fort Worth, TX.
- PCR Services Corporation (2001). "Watershed Management Area Plan for the Malibu Creek Watershed". Report prepared for Las Virgenes Malibu Conejo Council of Government.
- Schultz, G. A. (1994). "Meso-scale Modeling of Runoff and Water Balances Using Remote Sensing and Other GIS Data". *Hydrological Sciences Journal*, Vol. 39, No.2, April 1994, pp. 121—142.
- Skaggs, R. W., and Khaleel, R. (1982). *Infiltration, Hydrological Modeling of Small Watersheds*. American society of Agriculture Engineers, St. Joseph, MI.
- Smith, J., and Eli, R. N. (1995). "Neural-Network Models of Rainfall-Runoff Process" *Journal of Water Resources Planning and Management*, Vol. 121, No. 6, pp. 499 – 507.
- Soil Conservation Service (1986). *Urban Hydrology for Small Watersheds, Technical Release 55*. USDA, Springfield, VA.
- U.S. Army Corps of Engineers (1979). *Introduction and Application of Kinematic Wave Routing Techniques Using HEC-1, Training Document 10*. Hydrologic Engineering Center, Davis, CA.

- U.S. Army Corps of Engineers(a) (1994). *Flood-runoff Analysis, EM 1110-2-1417*.
Office of Chief of Engineers, Washington, DC.
- U.S. Army Corps of Engineers(a) (1998). *HEC-1 Flood Hydrograph Package User's Manual*. Hydrologic Engineering Center, Davis, CA.
- U.S. Army Corps of Engineers(a) (2000). *HEC-HMS Technical Reference Manual*.
Hydrologic Engineering Center, Davis, CA.
- U.S. Army Corps of Engineers(b) (2000). *HEC-GeoHMS User's Manual*. Hydrologic
Engineering Center, Davis, CA.
- Zhang, B., and Govindaraju, R. S. (2000). "Prediction of Watershed Runoff Using
Bayesian concepts and Modular Neural Networks". *Water Resources Research*,
Vol. 36, No. 3, pp. 753 – 762.

Table 5.1 Major Parameters Used in HEC-HMS

Parameters	Usage	Data Source
Precipitation Data	Model Input	LADPW Precipitation Gauges
Constant Loss Rate	Runoff Volume Calculation	NRCS STATSGO Soil Database
Imperviousness		LADPW Imperviousness GIS Layer
Typical Length	Direct Runoff Calculation	Processing USGS DEM
Roughness Coefficients		SCAG Land Use Layer and USACE Tables
Slope		Processing USGS DEM
Longest Flow Length		Processing USGS DEM
Runoff Discharge Data	Model Calibration	LADPW Malibu Canyon Runoff Gauge

Table 5.2 Major GeoHMS Parameters of Sub-watersheds

Watershed Id	Name	Longest Flow Length (ft)	Downstream Elevation (ft)	Upstream Elevation (ft)	Slope at End Point	Area (mi ²)
1	R10W10	62,917	505	2207	2.7%	19.67
2	R20W20	27,540	830	2362	5.6%	6.20
3	R30W30	42,093	830	2260	3.4%	8.58
4	R40W40	39,560	803	1794	2.5%	6.67
5	R50W50	28,464	846	2349	5.3%	6.07
6	R60W60	28,597	951	2522	5.5%	7.41
7	R70W70	3,584	803	879	2.1%	0.09
8	R80W80	40,592	951	2972	5.0%	9.23
9	R90W90	27,427	846	1604	2.8%	5.51
10	R100W100	21,322	715	2020	6.1%	3.44
11	R110W110	39,837	715	2099	3.5%	11.27
12	R120W120	27,892	505	1581	3.9%	4.60
13	R130W130	31,070	485	2782	7.4%	5.68
14	R140W140	11,719	433	1683	10.7%	2.31
15	R150W150	3,013	485	882	13.2%	0.07
16	R160W160	34,966	433	1633	3.4%	8.15
17	R170W170	26,952	0	1482	5.5%	4.61

Table 5.3 Los Angeles County Department of Public Works Precipitation Gauges

Station ID	Station Location	Latitude	Longitude	Elevation (ft)	Type
434	AGOURA	34-08-08	118-45-08	800	Automatic
735H	BELL CANYON	34-11-40	118-39-23	895	Automatic
435	MONTE NIDO	34-04-41	118-41-35	600	Automatic
1264	CALABASAS LANDFILL	34-08-25	118-42-35	800	Manual
306H	ZUMA BEACH	34-01-15	118-49-42	15	Manual

Table 5.4 Distance (ft)-matrix between Precipitation Gauges and Centroids of Sub-watersheds

Precipitation Gauge Watershed ID	Agoura 434	Bell Canyon 735H	Calabasas Landfill 1264	Monte Nnido 435	Zuma Beach 306H
1	16,387	19,692	7,490	30,409	62,709
2	14,825	33,478	20,853	41,383	59,303
3	14,656	27,759	16,708	38,795	61,598
4	15,466	39,624	24,721	43,001	55,560
5	27,525	60,226	40,395	50,268	42,543
6	45,339	75,873	58,020	68,081	51,575
7	3,175	36,278	14,839	30,716	48,995
8	45,050	78,494	58,025	64,454	42,085
9	21,869	50,213	33,505	48,343	51,199
10	4,373	39,155	14,656	24,761	43,663
11	16,885	52,837	29,352	35,085	34,551
12	21,048	31,124	11,855	12,838	54,111
13	12,662	44,824	19,364	19,825	37,028
14	22,831	43,241	20,470	5,348	42,777
15	17,839	40,960	16,795	10,049	42,567
16	28,742	35,533	19,873	10,103	55,679
17	33,970	52,093	31,320	8,408	42,121

Table 5.5 SCS Soil Groups and Corresponding Loss Rates [SCS, 1986]

Soil Group	Description	Range of Loss Rates (in/hr)
A	Deep sand, deep loess, aggregated silts	0.30-0.45
B	Shallow loess, sandy loam	0.15-0.30
C	Clay loams, shallow sandy loam, soils low in organic contents, and soils usually high in clay	0.05-0.15
D	Soils that swell significantly when wet, heavy plastic clays, and certain saline soils	0.00-0.05

Table 5.6 Overland-flow Roughness Coefficients [USACE, 1998]

Code	Surface Description	Roughness Coefficient
01	Smooth surfaces (concrete, asphalt, gravel, or bare soil)	0.011
02	Fallow (no residue)	0.050
03	Cultivated soils (1): Residue cover less than or equal to 20%	0.060
04	Cultivated soils (2): Residue cover greater than 20%	0.170
05	Grass (1): Short grass prairie	0.150
06	Grass (2): Dense grasses, including species such as weeping love grass, bluegrass, buffalo grass, and native grass mixture	0.240
07	Grass (3): Bermudagrass	0.410
08	Range	0.130
09	Woods (1): Light underbrush	0.400
10	Woods (2): Dense underbrush	0.800

Table 5.7 Overland-flow Roughness and SCAG Land Use Correlation Table

SCAG Land Use Code	Land Use Description	Corresponding Surface	Roughness Coefficient
1111	High Density Single Family Residential	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1112	Low Density Single Family Residential	Range	0.130
1121	Mixed Multi-Family Residential	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1122	Duplexes and Triplexes	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1123	Low-Rise Apartments Condominiums and Townhouses	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1124	Medium-Rise Apartments and Condominiums	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1125	High-Rise Apartments and Condominiums	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1131	Trailer Parks and Mobile Home Courts Low Density	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1132	Mobile Home Courts and Subdivisions Low Density	Range	0.130
1140	Mixed Residential	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1151	Rural Residential High Density	Range	0.130
1152	Rural Residential Low Density	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1211	Low- and Medium-Rise Major Office Use	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1212	High-Rise Major Office Use	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1221	Regional Shopping Mall	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1222	Retail Centers	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1223	Modern Strip Development	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1224	Older Strip Development	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1231	Commercial	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1232	Commercial Recreation	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1233	Hotels and Motels	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1234	Attended Pay Public Parking Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1241	Government Offices	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1242	Police and Sheriff Stations	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1243	Fire Stations	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1244	Major Medical Health Care Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1245	Religious Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1246	Other Public Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1247	Non-Attended Public Parking Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1251	Correctional Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1252	Special Care Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1253	Other Special Use Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1261	Pre-Schools/Day Care Centers	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1262	Elementary Schools	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1263	Junior or Intermediate High Schools	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1264	Senior High Schools	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1265	Colleges and Universities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1266	Trade Schools	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1271	Military Base (Built-up Area)	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1272	Military Vacant Area	Range	0.130
1311	Manufacturing and Assembly	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1312	Motion Picture	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1314	Research and Development	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1323	Open Storage	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1325	Chemical Processing	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1331	Mineral Extration - Other Than Oil and Gas	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1340	Wholesaling and Warehousing	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011

1411	Airports	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1412	Railroads	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1413	Freeways and Major Roads	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1414	Park and Ride Lots	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1415	Bus Terminals and Yards	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1416	Truck Terminals	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1420	Communication Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1431	Electrical Power Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1432	Solid Waste Disposal Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1433	Liquid Waste Disposal Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1434	Water Storage Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1435	Natural Gas and Petroleum Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1436	Water Transfer Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1437	Improved Flood Waterways and Structures	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1440	Maintenance Yards	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1450	Mixed Transportation	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1460	Mixed Transportation and Utility	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1500	Mixed Commercial and Industrial	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1600	Mixed Urban	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1700	Under Construction	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
1810	golf Courses	Grass: Bermudagrass	0.410
1820	Local Parks and Recreation	Grass: Bermudagrass	0.410
1830	Regional Parks and Recreation	Grass: short grass prairie	0.150
1832	Regional Park Undeveloped	Grass: dense grasses	0.240
1840	Cemeteries	Grass: short grass prairie	0.150
1860	Specimen Gardens and Arboreta	Grass: dense grasses	0.240
1870	Beach Parks	Grass: short grass prairie	0.150
1880	Other Open Space and Recreation	Grass: short grass prairie	0.150
1900	Urban Vacant	Grass: dense grasses	0.240
2110	Irrigated Cropland and Improved Pasture Land	Cultivated soils (residue cover > 20%)	0.170
2120	Non-Irrigated Cropland and Improved Pasture Land	Fallow (no residue)	0.050
2200	Orchards and Vineyards	Woods: Dense underbrush	0.800
2300	Nurseries	Cultivated soils (residue cover <= 20%)	0.060
2500	Poultry Operations	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
2600	Other Agriculture	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
2700	Horse Ranches	Range	0.130
3100	Vacant Undifferentiated	Fallow (no residue)	0.050
3200	Abandoned Orchards and Vineyards	Woods: Light underbrush	0.400
3300	Vacant With Limited Improvements	Fallow (no residue)	0.050
4100	Water Undifferentiated	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011
4300	Marina Water Facilities	Smooth Surface (concrete, asphalt, gravel, or base soil)	0.011

Table 5.8 MLP Networks Used in This Study

Network	Input Data	Training Data	Testing Data
Q-MLP1	R_t, R_{t-1}, R_{t-2}	Seasons 1997, 1998, 1999, 2000	Season 2001
Q-MLP2	R_t, R_{t-1}, R_{t-2}	Seasons 1998, 1999, 2000, 2001	Season 1997
Q-MLP3	$R_t, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}$	Seasons 1997, 1998, 1999, 2000	Season 2001
Q-MLP4	$R_t, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}$	Seasons 1998, 1999, 2000, 2001	Season 1997
Q-MLP5	$R_t, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}$	Data Sets have been randomized. 10% for cross validation, 25% for testing, and 65% for training	

Table 5.9 Residuals of the Seasonal Runoff Volume and Peak Discharge

Season	Volume (ac.-ft)			Peak Discharge (cfs)		
	Calibrated	Observed	Difference (%)	Calibrated	Observed	Difference (%)
1997	8,664	31,102	-72.14%	602.40	807.00	-25.35%
1998	54,786	81,494	-32.77%	4,425.90	4,420.00	0.13%
1999	5,544	7,418	-25.26%	292.06	134.00	117.96%
2000	8,553	16,401	-47.85%	938.47	701.00	33.88%
2001	18,914	38,824	-51.28%	3,508.30	3,950.00	-11.18%

Table 5.10 Optimized Scale Factors and Values of Objective Function

Season	Optimized Scale Factor		Objective Function
	Initial Loss Rate	Constant Loss Rate	
1997	0.97	6.68	163.80
1998	0.88	1.72	555.90
1999	0.98	3.75	38.70
2000	1.50	9.70	13.50
2001	1.30	3.58	1,285.10

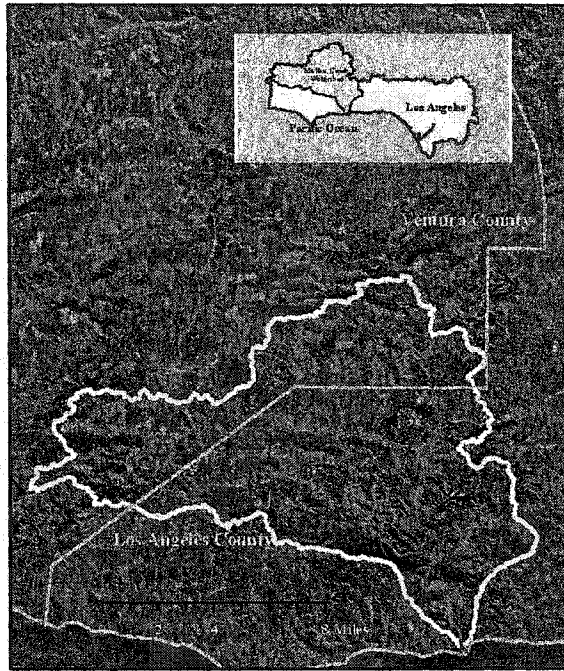
Table 5.11 ANN Simulation Results

Network	MSE	r		
		Training	Testing	Cross Validation
Q-MLP1	0.0046	0.82	0.67	N/A
Q-MLP2	0.0058	0.84	0.54	N/A
Q-MLP3	0.0027	0.90	0.51	N/A
Q-MLP4	0.0060	0.84	0.62	N/A
Q-MLP5	0.0043	0.79	0.84	0.84

MSE: Minimum Mean Square Error; r: correlation coefficient

Table 5.12 Sensitivity Analysis of Q-MLP4 Network

Parameter	Sensitivity
R_t	57.13
R_{t-1}	24.24
R_{t-2}	8.84
R_{t-3}	6.30
R_{t-4}	5.35
R_{t-5}	3.06



Legend
 [White Box] Santa Monica Bay
 [Stippled Box] Study Area



Figure 5.1
Malibu Creek Watershed Vicinity Map

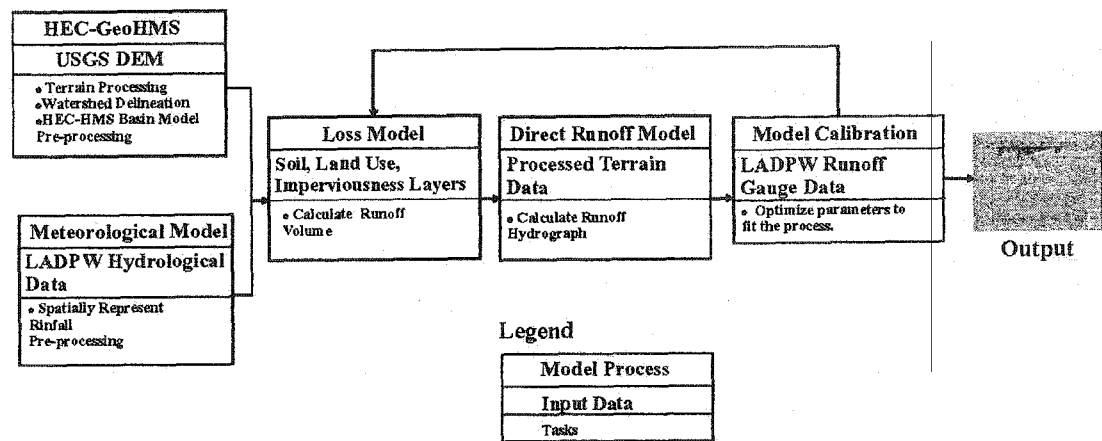


Figure 5.2
HEC-HMS Schematics

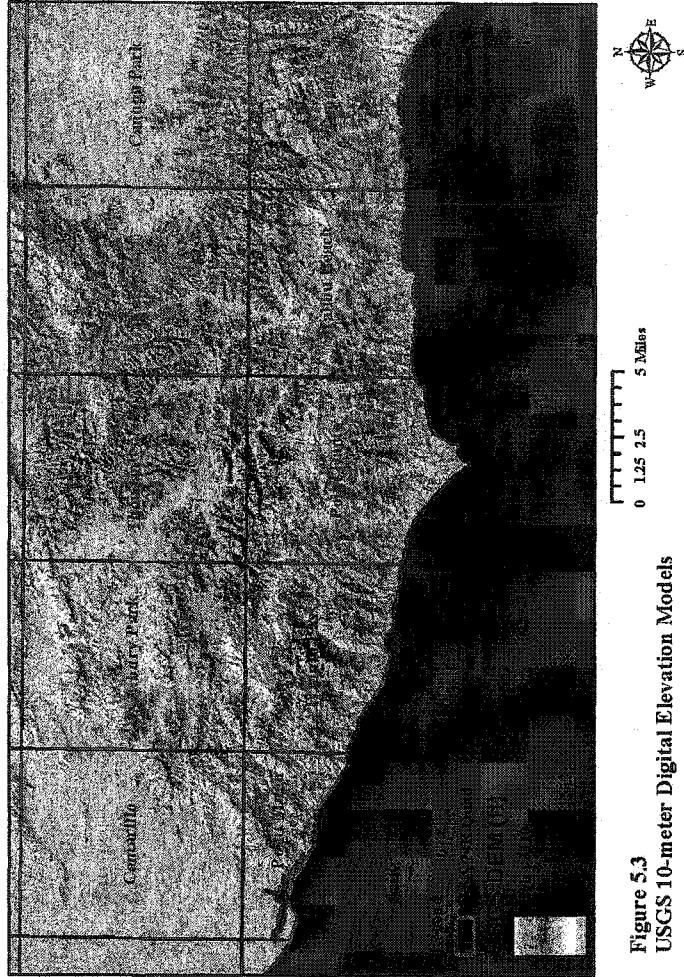


Figure 5.3
USGS 10-meter Digital Elevation Models

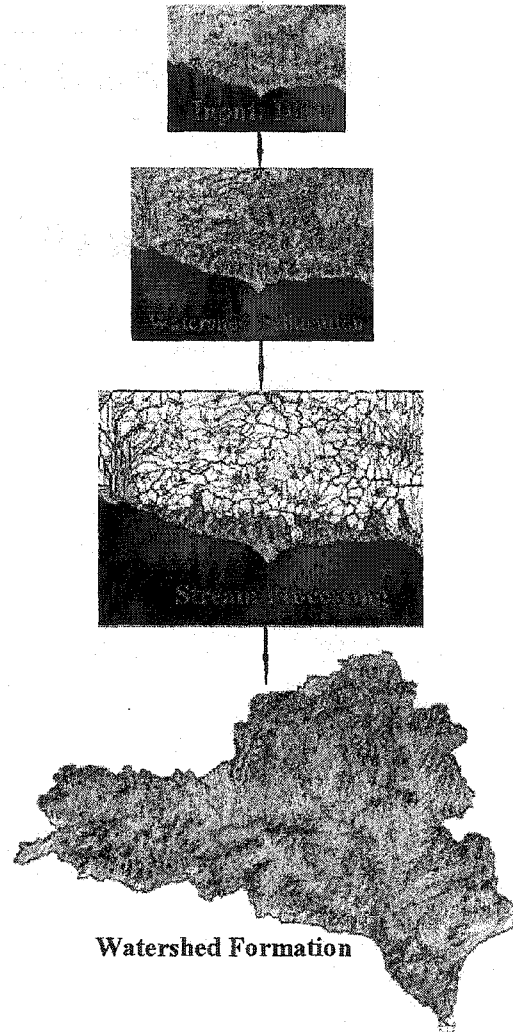


Figure 5.4
HEC-GeoHMS Processing

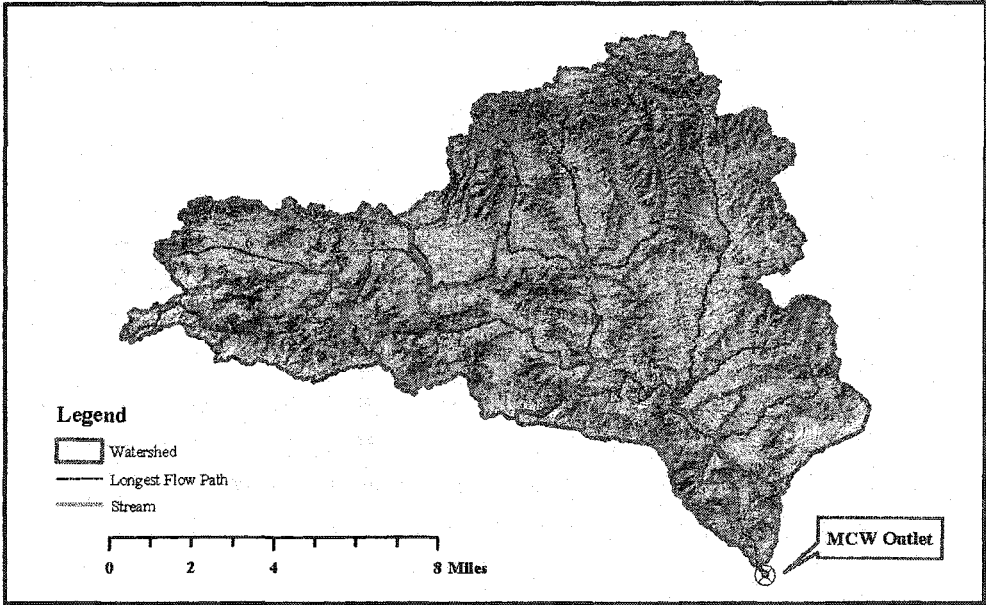


Figure 5.5
Sub-watersheds after Delineation



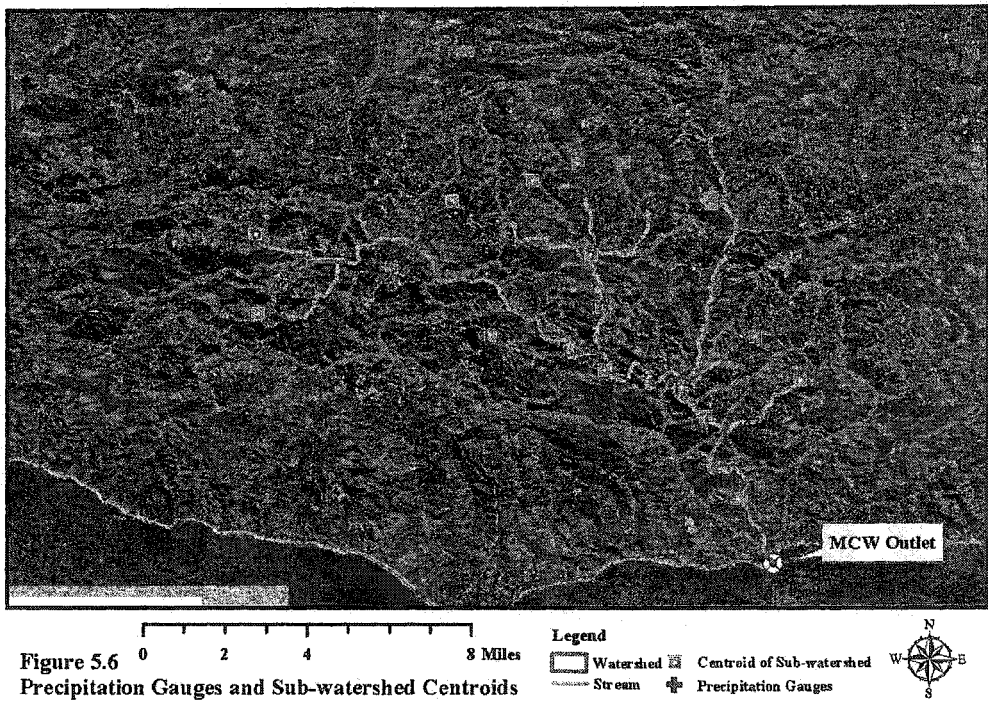


Figure 5.6
Precipitation Gauges and Sub-watershed Centroids

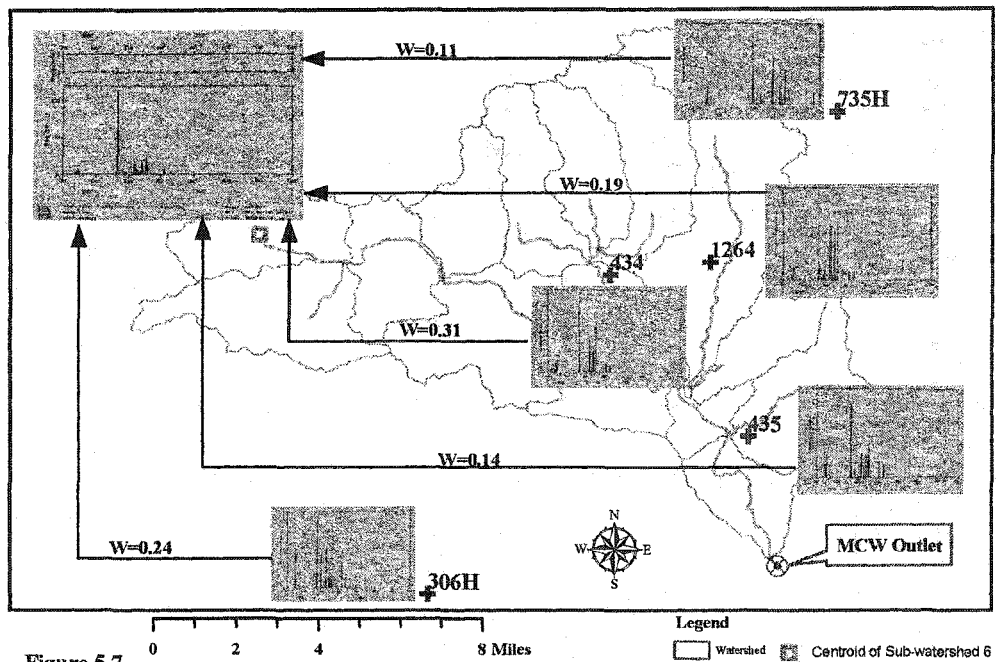


Figure 5.7
Inverse-distance-square Method to Calculate Sub-watershed Hyetograph

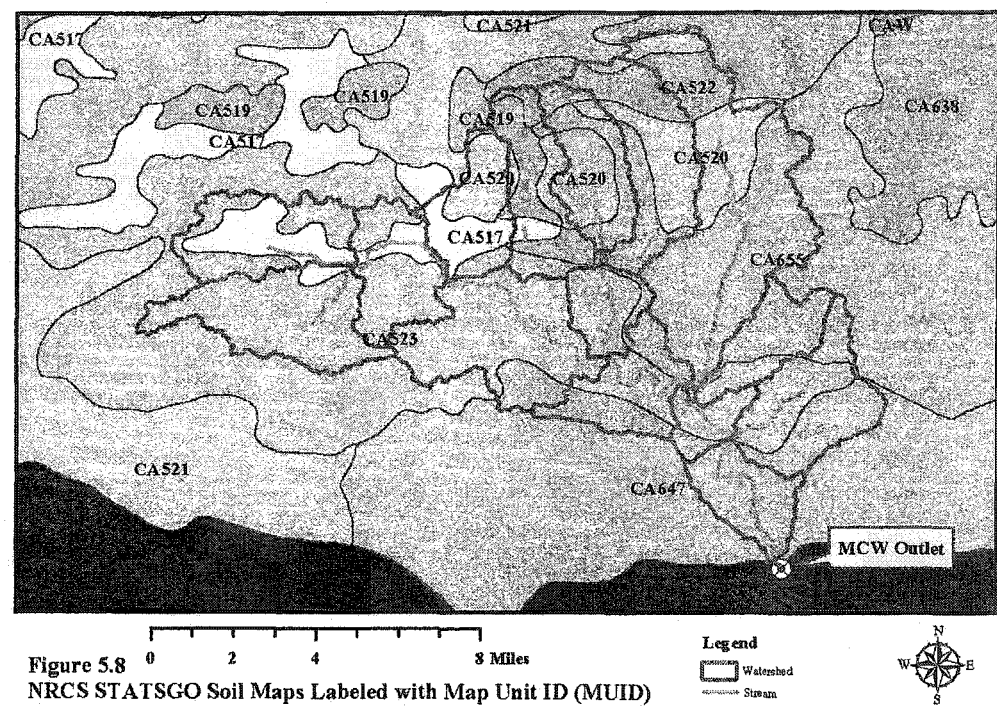


Figure 5.8
NRCS STATSGO Soil Maps Labeled with Map Unit ID (MUID)

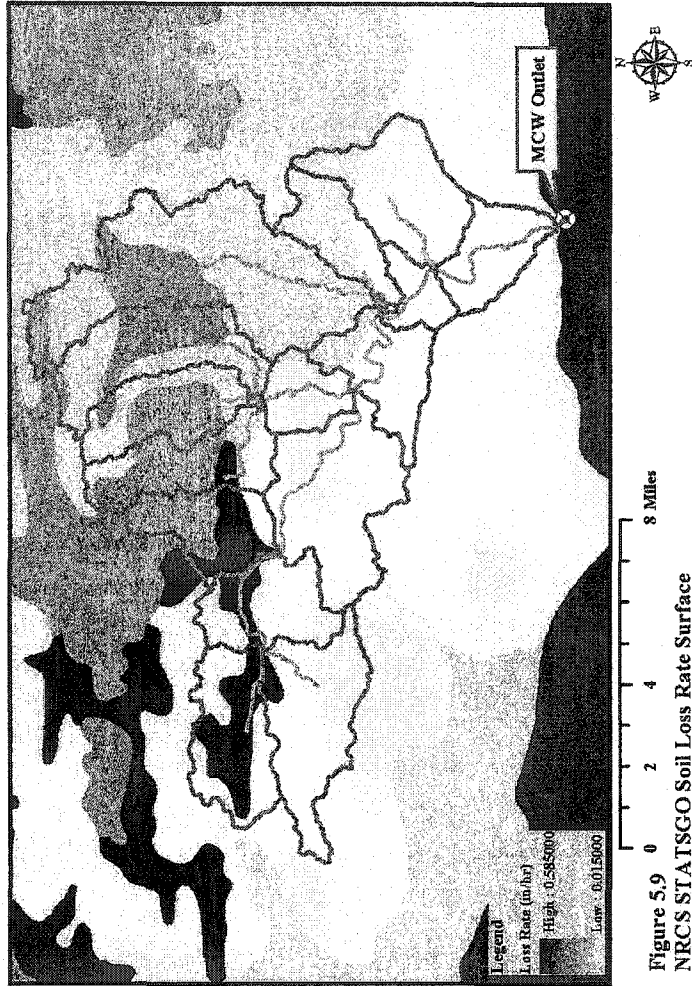
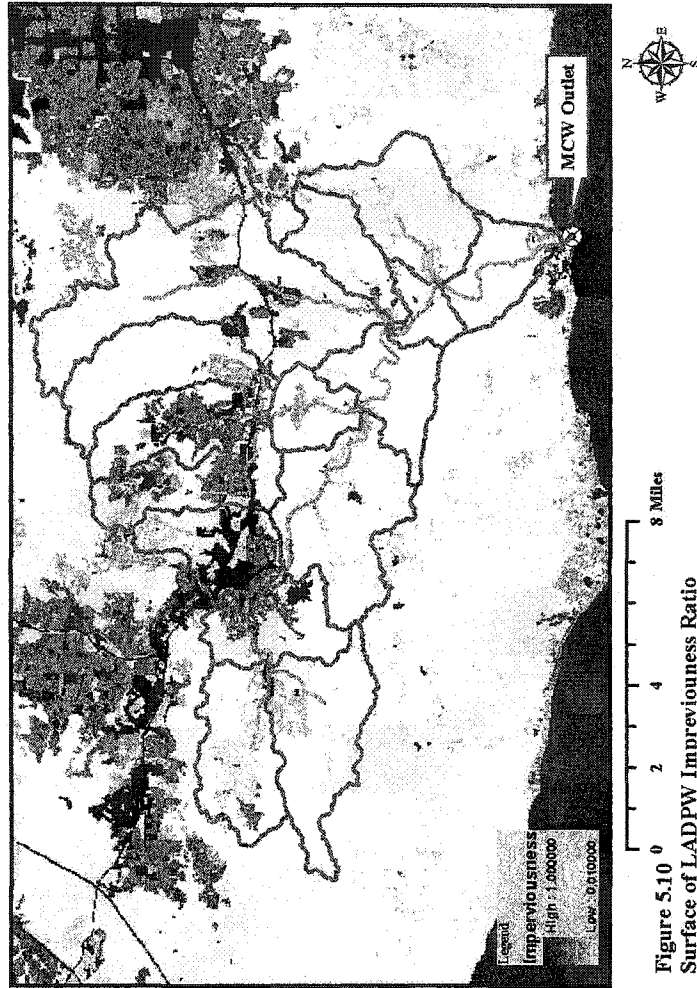


Figure 5.9
NRCS STATSGO Soil Loss Rate Surface



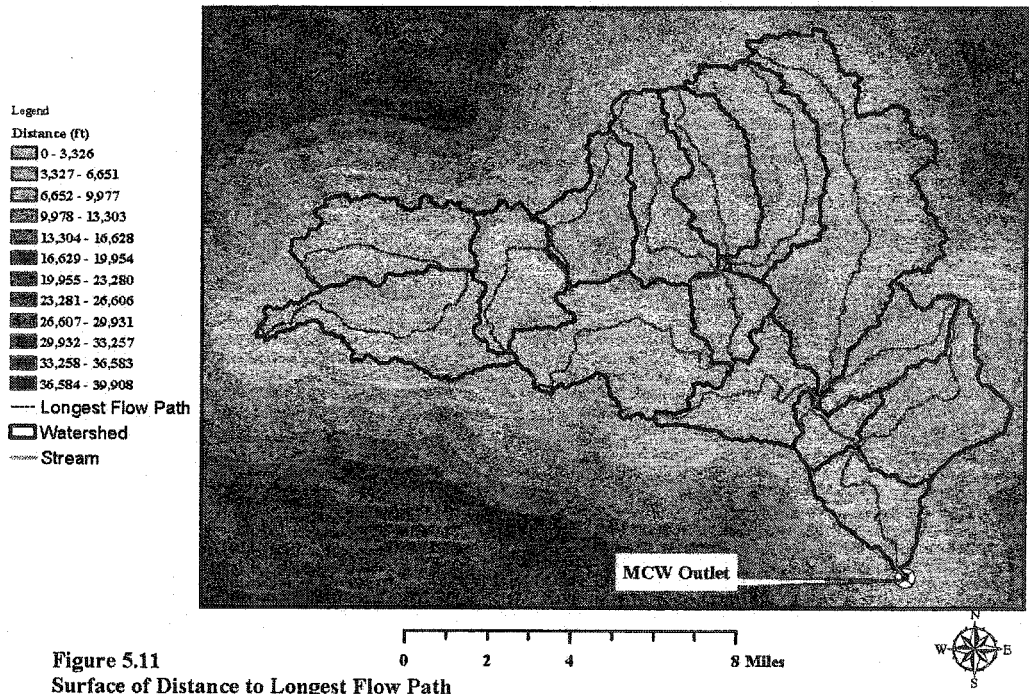


Figure 5.11
Surface of Distance to Longest Flow Path

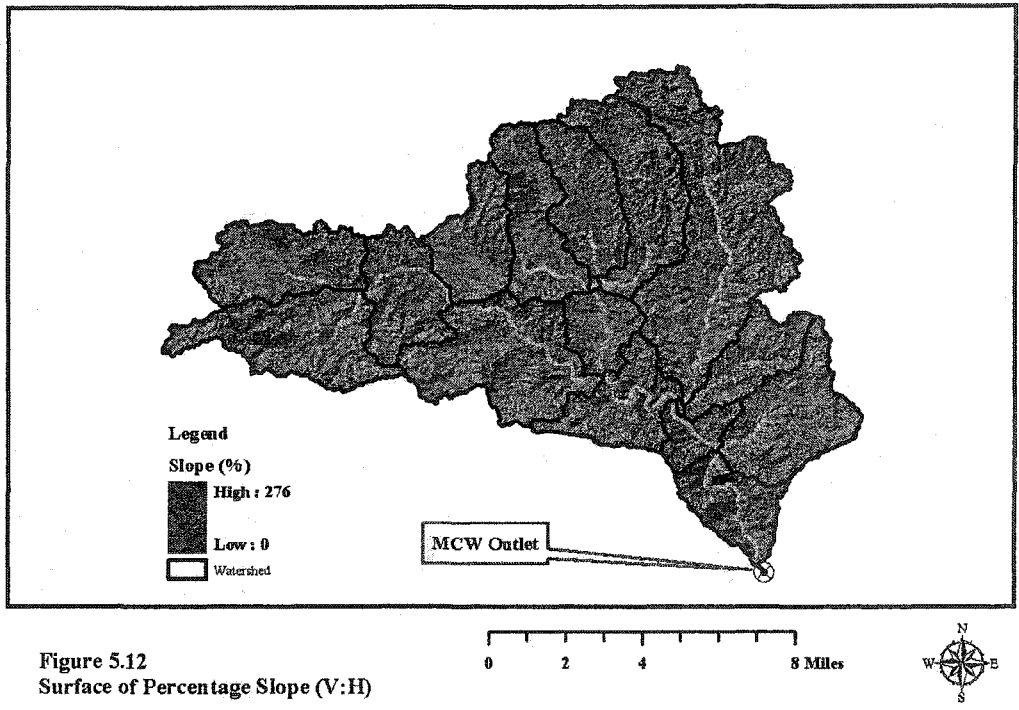


Figure 5.12
Surface of Percentage Slope (V:H)

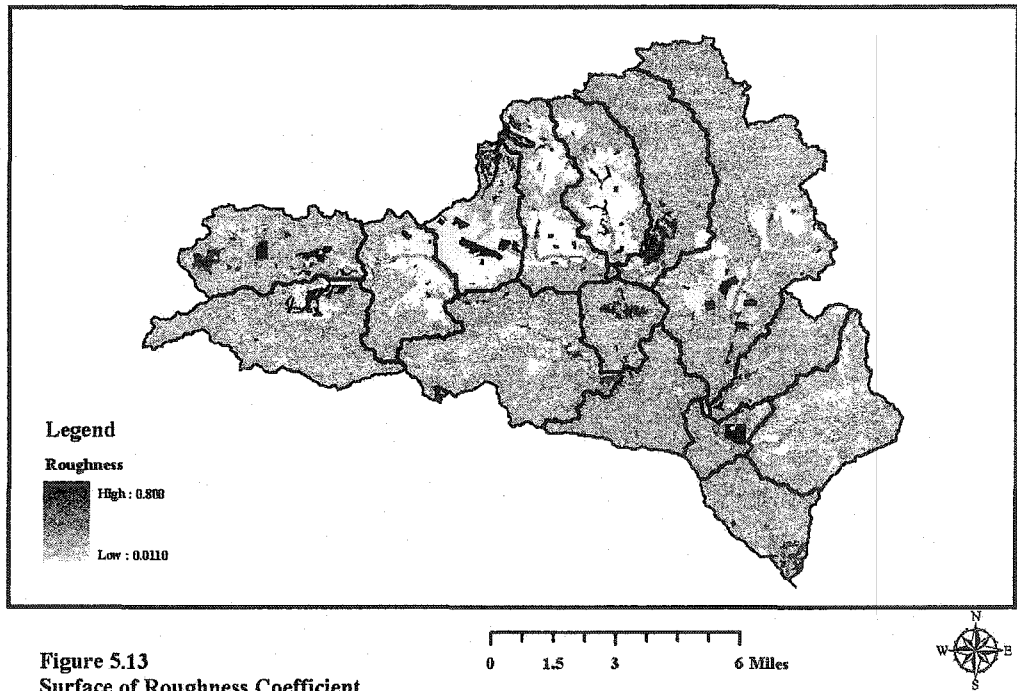


Figure 5.13
Surface of Roughness Coefficient

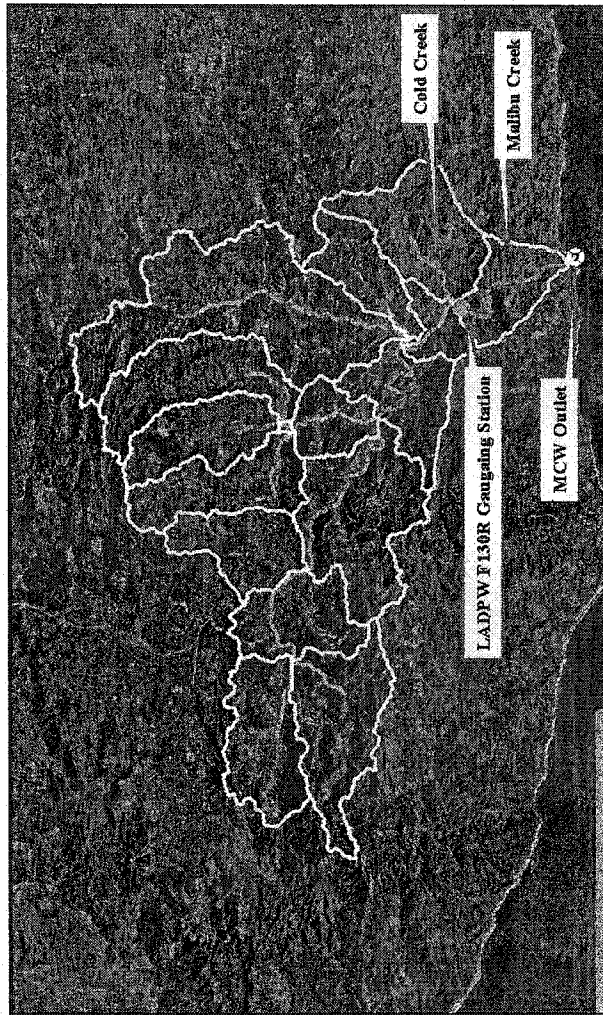
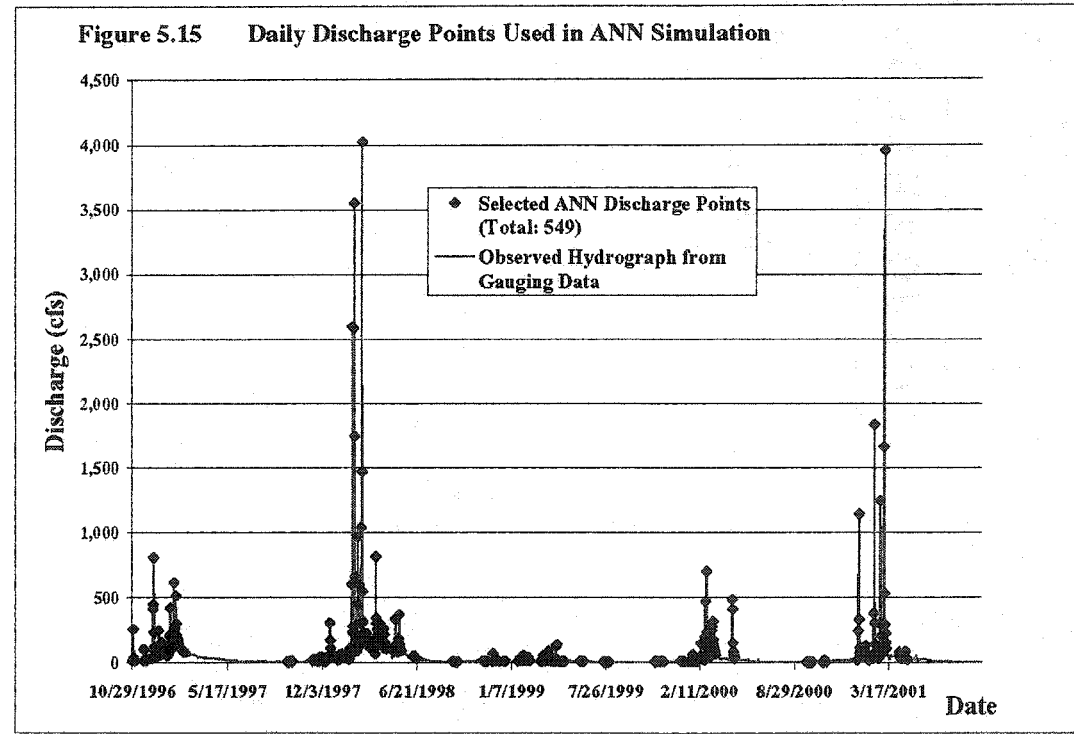
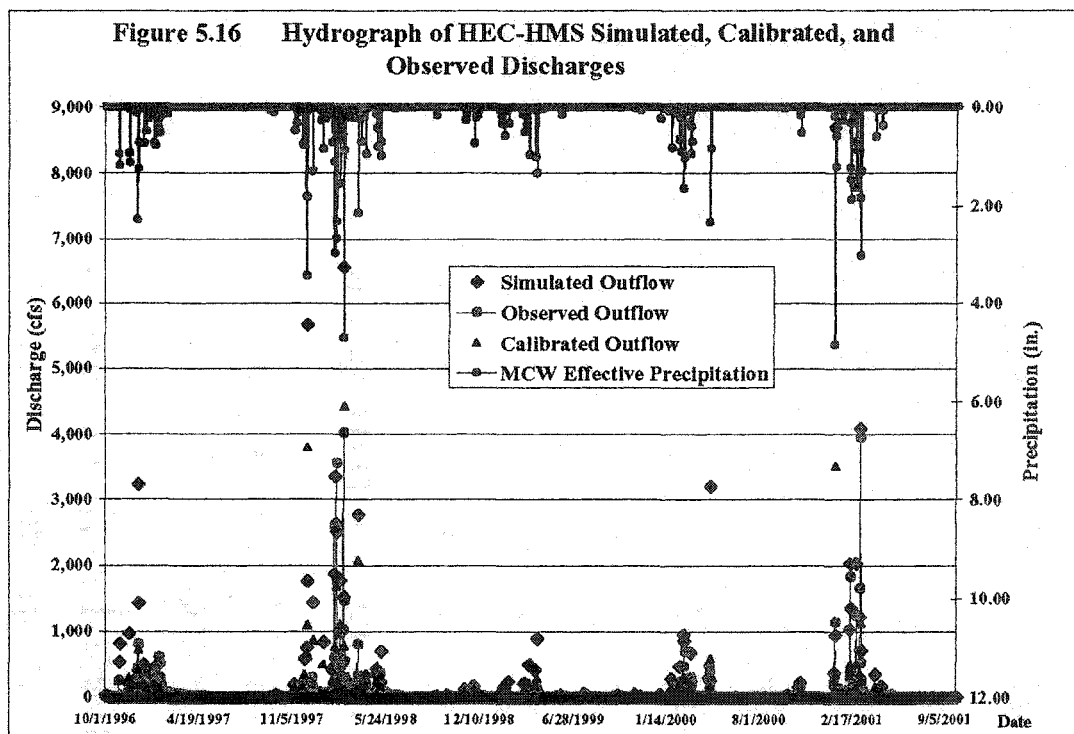
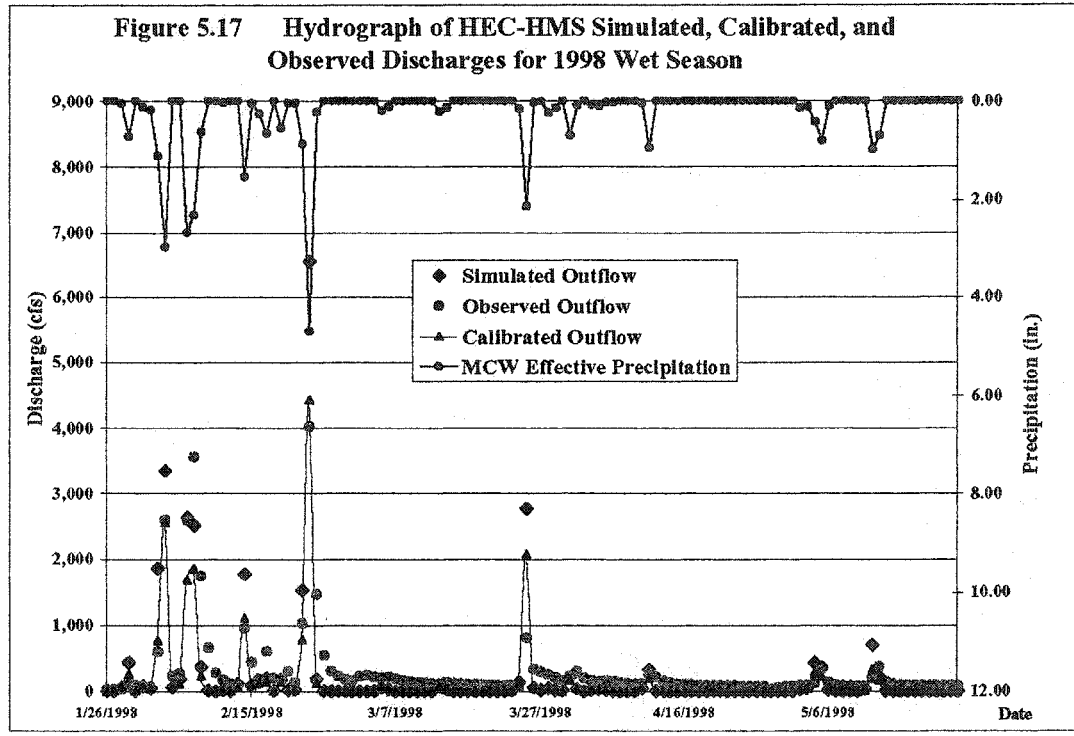
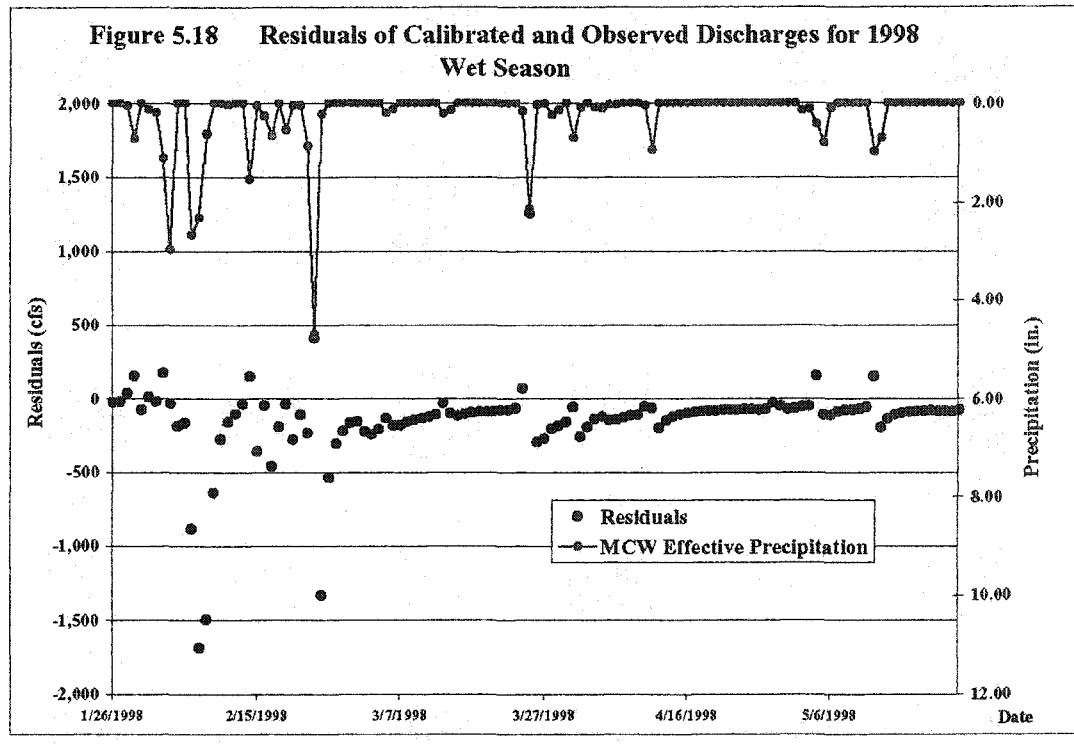


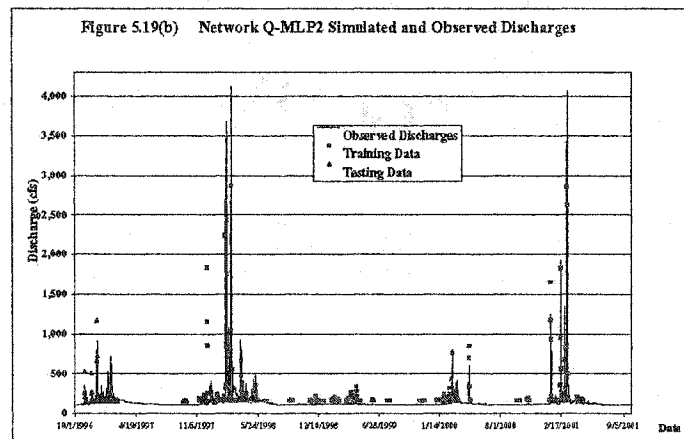
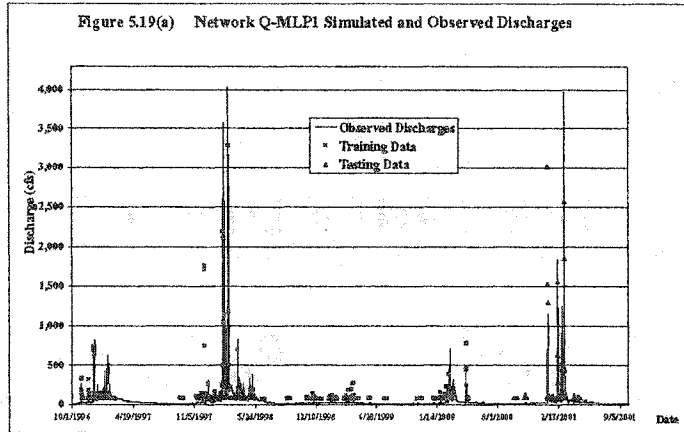
Figure 5.14
LADPW Stream Gauging Station F-130R

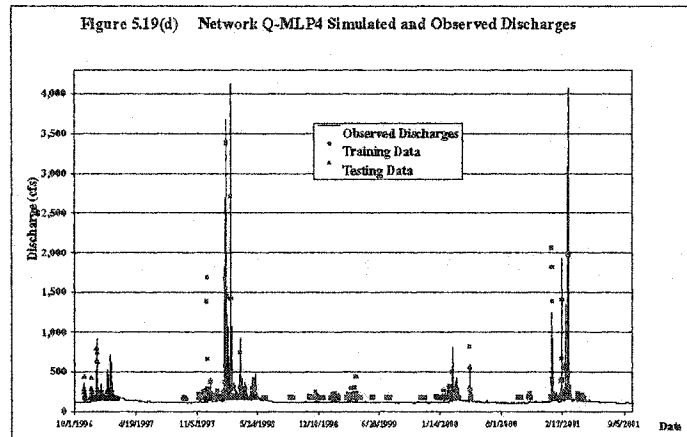
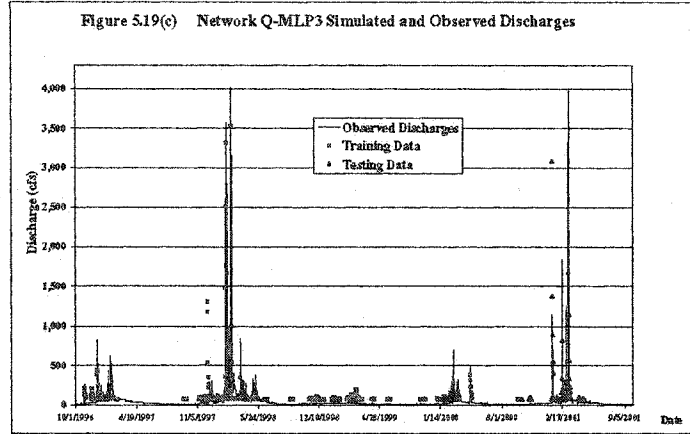


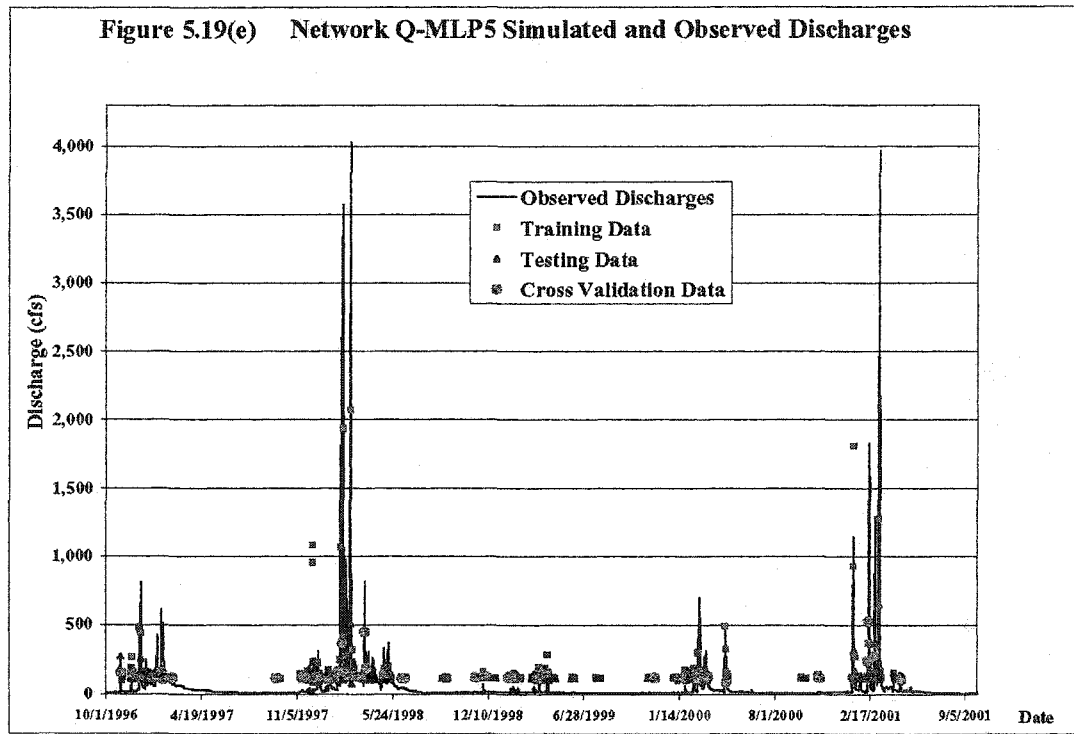


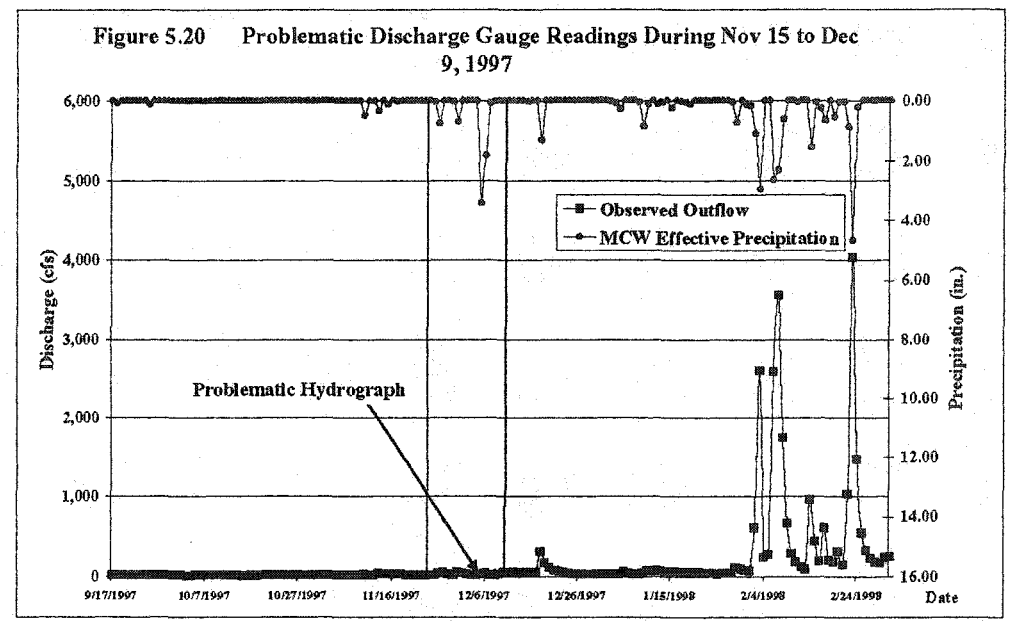


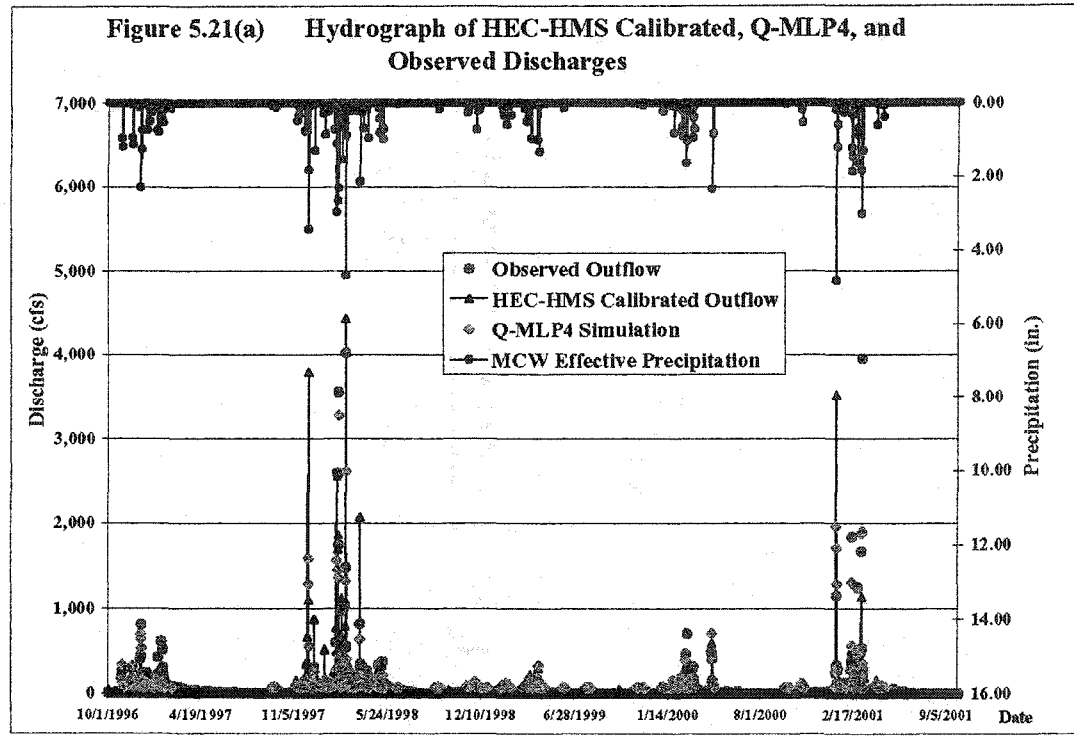


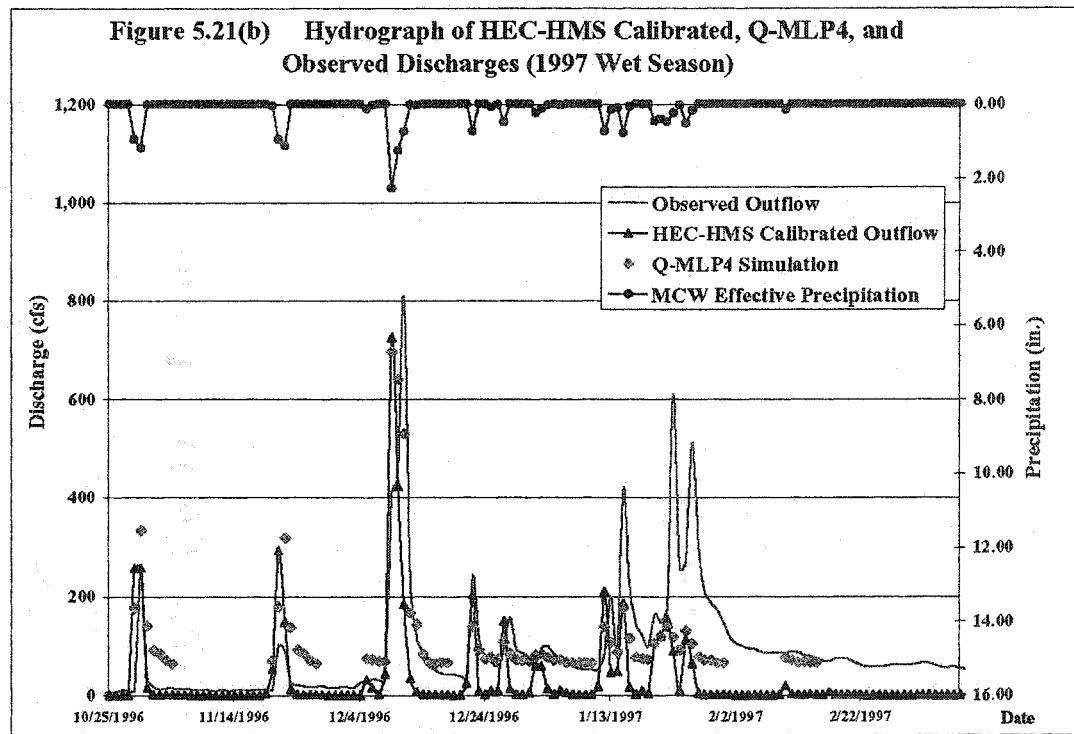


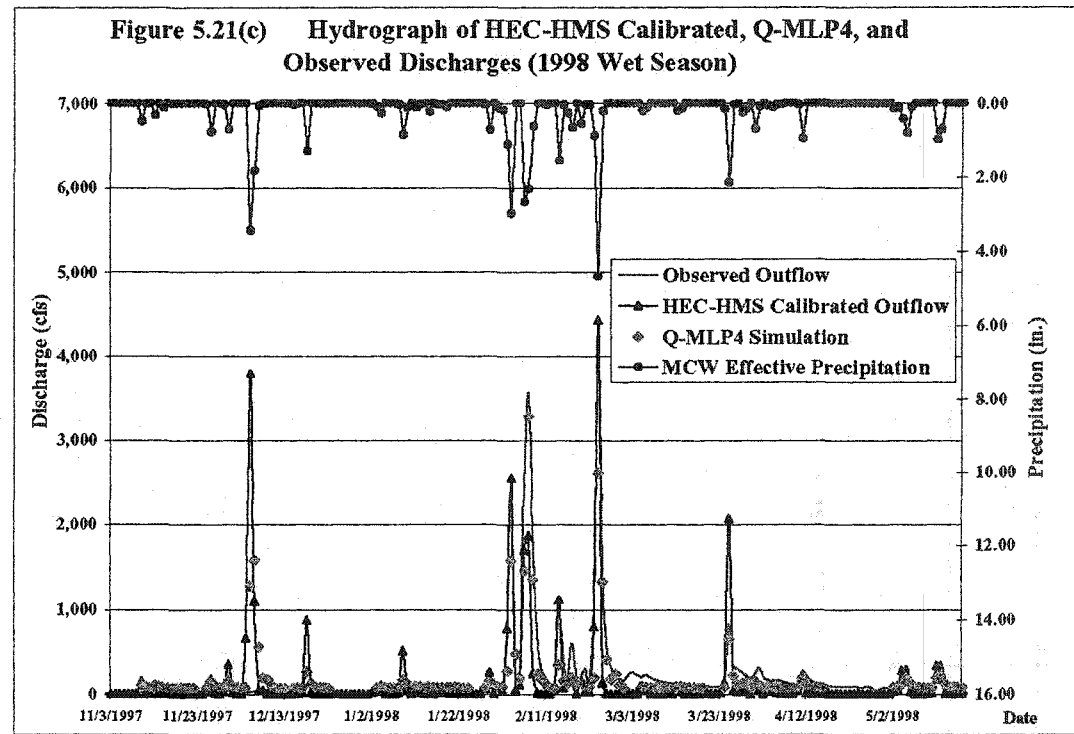












6 CONCLUSIONS

Using the unprecedented technological combination of GIS and ANN algorithms, this study was able to elicit new information from new data and old, thereby not only answering its immediate questions, but also securing topics of future research endeavors. Something that was already known, but could be now used to increased benefit was that of all the parameters, land use information played a critical role in monitoring stormwater runoff quantity and quality.

It is known that the performance of any modeling analysis is ultimately constrained by the availability and quality of its input data. Until now, spatial data relating to stormwater systems have been particularly prone to uncertainty and inaccuracy. Thus, one of the major accomplishments of this research was to establish a comprehensive spatial database for the study area. The process involves discovering reliable data sources, making intensive personal contacts, acquiring affordable or public data legitimately under license agreement, understanding constraints of data usage, and transforming data into desirable geospatial formats. The stormwater related data gathered in this study will be timeless and useful for future research activities, no matter how the technology has advanced.

The land use classification was studied first. The Landsat thematic mapper imagery was used as a relative inexpensive data source to investigate more costly land use information. Then GIS was used to extract the spectral signature of each SCAG land use category in the Santa Monica Bay area. Based on the results of several scattered

plots of land use patterns on average pixel values, it was decided that pixel values of various spectral bands would help differentiate the land use patterns.

Five artificial neural network classifiers including one with an input fuzzifier were built to distinguish existing SCAG land use polygons. The network inputs included pixel values of seven landsat spectral bands, digital elevations, average slopes, and coordinate information. Generally speaking, networks with more inputs generated better classification accuracies. The addition of a fuzzifier to the network also enhanced the network performance. Sensitivity studies were performed to evaluate the efficiency of each parameter to classify each land use pattern.

Another three supervised neural network classifiers and three unsupervised networks were constructed to study the land use pixels of the Ballona Wetland and its vicinity areas. Again supervised networks performed better with more inputs. The results also indicated that pixel classification is more accurate than polygon-level classification. It should also be noted that pixel-level classifications required a lot more computing resources. With proper configuration, the unsupervised networks successfully clustered pixels to somewhat imitate certain major land features.

Finally a conventional deterministic hydrological model was developed to simulate the daily discharge hydrograph of the Malibu Creek Watershed. In the mean time, artificial neural network approach was also applied as a comparison to the deterministic model. The results revealed that though with a much less data quantity requirement, neural network simulation could still marginally outperform the deterministic model in predicting certain specific output parameters. It was also

discussed that how a potential problematic precipitation data set could influence the final model predictions.

7 FUTURE WORK

Technology undergoes constant improvement, thus engendering the same in those fields of study that thrive on the immediacy and precision that it offers. Fields such as stormwater have experienced an exhilarating rate of growth—of budgets, of projects, of interest—and of possibilities.

Specifically, dramatic improvements in sensor technology, computational speed, and processing algorithms have stimulated interest and serious professional research in the area of data mining for stormwater systems. GIS technology is also experiencing a boom in applicability. Evolved initially as a vector based system working with points, lines, and polygons, modern GIS technology incorporates powerful raster computing capabilities into its more conventional use.

The research approaches proposed in Chapters Two, Three, and Four should be applied to higher resolution and better quality imagery, such as aerial and IKONOS satellite imagery.

Efforts to integrate GIS with stormwater models should continue and new research should be pursued. Instead of using pixel numbers extracted from satellite imagery to indirectly predict model input parameters, such as imperviousness from land use information, it is suggested that soft computing methodology be used to directly calculate model inputs, or even model outputs. However, the direct approach requires a lot of reliable training data sets from fields in order to yield convincing network learning results.

At the mean time, temporal data modeling in GIS is getting more attention today than ever before. Adding the time element to GIS raises some intriguing questions and present new difficulties. Nevertheless, across the nation, on-going efforts to create a GIS include the dimension of time. Thus, future research activities to incorporate temporal remote sensed data and land use information to monitor stormwater system would be very promising.