Co-Location Decision Tree for Enhancing Decision-Making of Pavement Maintenance and Rehabilitation

Guoqing Zhou

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

In
Civil and Environmental Engineering

Linbing Wang, Committee Chair

Montasir Abbas, Committee Member

Gerardo W. Flintsch, Committee Member

Antoine G. Hobeika, Committee Member

Antonio A. Trani, Committee Member

January 17, 2011

Blacksburg, Virginia

Keywords: Pavement Management, Maintenance and Rehabilitation, Decision Tree, Spatial Data Mining, Co-Location, GIS

Co-Location Decision Tree for Enhancing Decision-Making of Pavement Maintenance and Rehabilitation Strategy

Guoqing Zhou

ABSTRACT

A pavement management system (PMS) is a valuable tool and one of the critical elements of the highway transportation infrastructure. Since a vast amount of pavement data is frequently and continuously being collected, updated, and exchanged due to rapidly deteriorating road conditions, increased traffic loads, and shrinking funds, resulting in the rapid accumulation of a large pavement database, knowledge-based expert systems (KBESs) have therefore been developed to solve various transportation problems. This dissertation presents the development of theory and algorithm for a new decision tree induction method, called *co-location-based decision tree (CL-DT.)* This method will enhance the decision-making abilities of pavement maintenance personnel and their rehabilitation strategies. This idea stems from shortcomings in traditional decision tree induction algorithms, when applied in the pavement treatment strategies. The proposed algorithm utilizes the co-location (co-occurrence) characteristics of spatial attribute data in the pavement database. With the proposed algorithm, one distinct event occurrence can associate with two or multiple attribute values that occur simultaneously in spatial and temporal domains.

This research dissertation describes the details of the proposed CL-DT algorithms and steps of realizing the proposed algorithm. First, the dissertation research describes the detailed co-location mining algorithm, including spatial attribute data selection in pavement databases, the determination of candidate co-locations, the determination of table instances of candidate co-locations, pruning the non-prevalent co-locations, and induction of co-location rules. In this step, a hybrid constraint, *i.e.*, spatial geometric distance constraint condition and a distinct event-type constraint condition, is developed. The spatial geometric distance constraint condition is a neighborhood relationship-based spatial joins of table instances for many prevalent co-locations with one prevalent co-location; and the distance event-type constraint condition is a Euclidean distance between a set of attributes and its corresponding clusters center of attributes. The dissertation research also developed the spatial feature pruning method using the multi-resolution pruning criterion. The cross-correlation criterion of spatial features is used to remove the non-prevalent co-locations from the candidate prevalent co-location set under a given threshold. The dissertation research focused on the development of the co-location decision tree (CL-DT) algorithm, which includes the non-spatial attribute data selection in the pavement management database, co-location algorithm modeling, node merging criteria, and co-location decision tree induction. In this step, co-location mining rules are used to guide the decision tree generation and induce decision rules.

For each step, this dissertation gives detailed flowcharts, such as flowchart of co-location decision tree induction, co-location/co-occurrence decision tree algorithm, algorithm of co-location/co-occurrence decision tree (CL-DT), and outline of steps of SFS (Sequential Feature Selection) algorithm. Finally, this research used a pavement database covering four counties, which are provided by NCDOT (North Carolina Department of Transportation), to verify and test the proposed method. The comparison analyses of different rehabilitation treatments proposed by NCDOT, by the traditional DT induction algorithm and by the proposed new method are conducted. Findings and conclusions include: (1) traditional DT technology can make a consistent decision for road maintenance and rehabilitation strategy under the same road conditions, *i.e.*, less interference from human factors; (2) the traditional DT technology can increase the speed of decision-making because the technology automatically generates a decision-tree and rules if the expert knowledge is given, which saves time and expenses for PMS; (3) integration of the DT and GIS can provide the PMS with the capabilities of graphically displaying treatment decisions, visualizing the attribute and non-attribute data, and linking data and information to the geographical coordinates. However, the traditional DT induction methods are not as quite intelligent as one's expectations. Thus, post-processing and refinement is necessary. Moreover, traditional DT induction methods for pavement M&R strategies only used the non-spatial attribute data. It has been demonstrated from this dissertation research that the spatial data is very useful for the improvement of decision-making processes for pavement treatment strategies. In addition, the decision trees are based on the knowledge acquired from pavement management engineers for strategy selection. Thus, different decision-trees can be built if the requirement changes.

This dissertation research has demonstrated the advantages of the proposed method on the basis of the experimental results and several comparison analyses including the induced decision tree parameters, the misclassified percentage, the computational time taken, support, confidence and capture for rule induction, and the quantity and location of each treatment strategy. It has been concluded that (1) the proposed CL-DT algorithm can make better decisions for pavement treatment strategies when compared to the traditional DT method; (2) the proposed CL-DT method misclassified from 61.2% to 9.7%, which implies that the training data can contribute to decision tree induction; (3) the proposed CL-DT algorithm saves the 20% computational time taken in tree growing, tree drawing, and rule generation; (4) the percentage of support, confidence and capture of the FDP treatment strategy for the proposal CL-DT algorithm increases from 71.6%, 55.6%, and 66.2% to 83.2%, 84.4% and 77.7%, respectively; (5) the quantity of each treatment strategy discovered by CL-DT is very close to those proposed by the ITRE(Institute for Transportation Research and Education (ITRE); and (6) the location of each treatment strategy proposed by CL-DT is also very close to those proposed by the ITRE.

ACKNOWLEDGEMENT

I sincerely wish to express my gratitude to my advisor Dr. Linbing Wang for his support, guidance, patience, and encouragement throughout the course of this research. Also, I am greatly indebted to him for his critical review of the manuscript of my dissertation.

I would like to thank my committee members, Dr. Montasir Abbas, Dr. Gerardo W. Flintsch, Dr. Antoine G. Hobeika, and Dr. Antonio A. Trani for serving on my committee, and providing me with excellent course instruction during my graduate study at Virginia Tech.

I would like to take this opportunity to express my deepest thanks to all of those who assisted me throughout my academic career and made the completion of this dissertation possible. Their names that I can mention here include, Dr. Oktay Baysal, Prof. Gary Crossman, and Dr. Mileta Tomovic at Old Dominion University.

It is my greatest pleasure to dedicate this small achievement to my wife and two sons. Throughout my education I totally have relied on their love and support.

Finally, I would like to extend my deepest thanks to many friends for their support and understanding throughout the Virginia Tech life.

Table of Contents

List of Tables

List of Figures

1. INTRODUCTION

1.1 Background

The U.S. Department of Transportation (U.S. DOT) initiated the Commercial Remote Sensing and Spatial Information Technology Application to the transportation program in 1999 in collaboration with the National Aeronautics and Space Administration (NASA), in accordance with Section 5113 of the Transportation Equity Act for the 21st Century (DOT-NASA, 2003; 2002). The collaborative program with NASA is administered by the U.S. Department of Transportation (U.S. DOT) Research and Special Programs Administration (RSPA). The program was intended to focus on unique and cost-effective application, of remote sensing and spatial information technologies for delivering smarter, more efficient and responsive transportation services with enhanced safety and security (DOT-NASA, 2003; 2002). The five (originally four) application areas within the program have been (DOT-NASA, 2002) (Figure 1.1):

1) Environmental assessment, integration, and streamlining for faster decision making at reduced costs;
2) Transportation infrastructure management for improving maintenance service efficiency;
3) Traffic surveillance, monitoring and management for monitoring and managing traffic and freight flow;
4) Safety hazards and disaster assessment for unplanned disasters and security of critical transportation lifeline systems; and
5) Highway and runway pavement construction, quality control and maintenance.

The above four/five major priority areas of the collaborative program were deployed through national consortia, each of which consists of teams from leading institutions, industries and service providers. The major administrations from U.S. Department of Transportation are

- Bureau of Transportation Statistics,
- Federal Aviation Administration,
- Federal Highway Administration,
- Federal Motor Carrier Safety Administration,

- Federal Railroad Administration,
- Federal Transit Administration,
- Maritime Administration,
- National Highway Traffic Safety Administration, and
- Research and Special Programs Administration.

The research centers of NASA consist of

- Ames Research Center,
- Dryden Flight Research Center,
- Glenn Research Center at Lewis Field,
- Jet Propulsion Laboratory,
- Johnson Space Center,
- Kennedy Space Center,
- Langley Research Center,
- George C. Marshall Space Flight Center,
- Goddard Space Flight Center, and
- John C. Stennis Space Center.

Remote Sensing and Geospatial Information Technologies
Application to Multimodal Transportation
U.S. DOT Research Pays Off in Achieving Smarter
Transportation Services, Safety and Security

Figure 1.1: Collaborative program between US DOT and NASA on remote sensing and geospatial information technologies application to transportation (Courtesy of DOT, 2003)

The leasing institutions, industries and service providers in the national consortia for the four major priority areas are (DOT-NASA, 2003):

1) *Area of the Environmental assessment, integration and streamlining*. Leading by Mississippi State University (www.ncrste.msstate.edu), consisting of University of Alabama in Huntsville, University of Mississippi, Auburn University, U.S.RA, NASA Marshall Space Flight Center, Digital Globe, Intermap Technologies Corportion, Earth Data Technologies, LLC, ITRES Corporation, Virginia DOT, EarthData, ICF Consulting, Washington State DOT, and Veridian Systems Division. The focuses of this consortium are:

- Developing new solutions for transportation relocation and corridor planning,
- Developing algorithms for using raster and vector geospatial data in corridor planning,
- Relocating the CSX railroad in the Mississippi coastal corridor,
- Assessing urban growth in coastal corridors,
- LIDAR applications for terrain mapping and hydrologic analysis,
- LIDAR application for alignment optimization,
- Hyper spectral image data for wetland vegetation mapping and analysis,
- Geospatial data fusion application to transportation environmental assessment,
- Analysis of transportation, development, and population growth impacts on urban watersheds,
- LIDAR measurements of air pollutants and air quality modeling,
- Assessing urban growth and transportation impacts on human and natural environments,
- Developing computational mapping resources and geospatial data libraries for environmental assessment and transportation corridor planning, and
- Assessing user needs for geospatial and remote sensing technologies in transportation.

2) *Area of the safety hazards and disaster assessment*. Leading University of New Mexico (www.trans-dash.org), consisting of University of Utah, Oak Ridge National Laboratory, George Washington University, York University, Image Cat Inc., Digital Globe, and AERIS Inc. The focuses of this consortium are:

- Integrating remote sensing technology for planning evacuations in emergencies,
- Detecting damaged bridges for emergency response in southern California,

- Planning community evacuations for large populations,
- Tools for managing highway bridges,
- Transportation hazards consequence tool,
- Accessing and delivering geospatial data and toolkits for transportation applications,
- Protecting the critical infrastructure using Rational Mapper—a tool for processing high-resolution images,
- Assessing pipeline and airport safety using automated processing of LIDAR data,
- Hyper spectral analysis of urban surface materials,
- Lane-based evacuation routing tools to reduce evacuation times,
- Detailed evacuation simulations for identifying communities that could be trapped in a bottleneck,
- Mapping areas of potential damage to highways and pipelines due to land subsidence,
- New remote sensing technologies for planning and maintaining pipeline corridors,
- Managing rural roads in Indian reservations,
- Calculating mileages for highway performance monitoring for FHWA,
- Identifying glide path safety obstructions at the Santa Barbara Municipal Airport,
- Weather-related road hazards assessment and monitoring system for real-time weather monitoring and rural road condition assessment, and
- High-resolution satellite data updates E-911 road information.

3) *Area of the transportation infrastructure management for improving maintenance service efficiency*. Leading University of California at Santa Barbara (www.ncgia.ucsb.edu/ncrst), consisting of University of Wisconsin-Madison, Iowa State University, University of Florida, Digital Geographic Research Corp., Geographic Paradigm Computing Inc., Florida DOT, University of Massachusetts, Orbital Imaging Corporation, and Tetra Tech, Inc. The focuses of this consortium are:

- Responding to security threats, hazards and disasters,
- Evacuating a small neighborhood: infrastructure adequacy,
- Meeting the challenge of inventory assessment,
- Urban hyper spectral sensing and road mapping,

- LIDAR applications for highway design and construction,
- LIDAR for engineering design,
- BridgeView – a tool for bridge inventory and assessment,
- Security sitting of off-port inspection facilities,
- Tools for managing highway bridges for the National Bridge Inventory, and
- Aviation infrastructure planning and development support.

4) *Area of traffic surveillance, monitoring and management*. Leading by the Ohio State University (www.ncrst.org), consisting of George Mason University, University of Arizona, GeoData Systems Inc., TerraMetrics Inc., Veridian Grafton Technologies, Technology Service Corp., and Bridgewater State College. The focuses of this consortium are:

- Improving a real-time bus information system with image-based backdrops,
- Applications for traffic operations,
- Cheaper and more accurate traffic measures using satellite and airborne imagery,
- Determining highway level of service using airborne imagery,
- Improving freight flow management,
- High resolution georeferencing from airborne images for traffic flow,
- "Bird's-eye" views of transportation networks for mitigating urban congestion,
- Exploring LIDAR applications for traffic flow,
- Pioneering traffic data collection from UAVs,
- Automated vehicle tracking from airborne video,
- UAV applications for multi-modal operations, and
- Airborne data acquisition system (ADAS) for traffic surveillance.

1.2 Relevant Efforts Under the Collaborative Program

The collaborative program accomplishments have created a new model for R&D application by combining resources from U.S. DOT with NASA research capabilities in partnerships with universities and technology service providers (Usher and Truax, 2001). A detailed survey and analysis for the applications of land satellite remote sensing in transportation infrastructure and systems engineering has been made by Zhou and Wei (2009a; 2008b and 2008c). The applications of commercial remote sensing and geospatial information technologies can briefly categorized by the following area:

(1) Geospatial technology applied in transportation infrastructure, such as pavement construction and maintenance, and management,
(2) Geospatial technology applied in transportation planning,
(3) Geospatial technology applied in transportation safety analysis and monitoring,
(4) Geospatial technology applied in transportation operation and analysis, and
(5) Geospatial technology applied in transportation environmental analysis.

The overview of relevant efforts in the above four fields has been made by Zhou et al. (2009a; and 2008b). This Chapter will highlight the relevant efforts on the geospatial technology applied in transportation infrastructure.

Pavement construction is one of the most important aspects of transportation infrastructure. Pavement construction quality monitoring and evaluation for early scheduling of repair and maintenance are important in many areas of pavement engineering, especially in a pavement management system (Gilly et al., 1987). Remote sensing technologies using electromagnetic waves from various parts of the energy spectrum can acutely reflect the physical and chemical properties of pavement material changes, and thus can be used for monitoring and evaluating pavement construction (Usher and Truax, 2001). This program has been carried out for a couple of years in cooperation with NASA and a consortium of university research centers. Many successful examples sponsored from this program can be found in the recent published papers and symposiums, for instance, Reginald (2004); Karimi et al. (1999); Shauna et al. (2001); Zhou and Wang (2010a; 2010b; 2008a; 2008b); .

Many current researchers applied hyper spectral images for pavement mapping (Herold et al., 2005; 2004) or high-resolution satellite imagery (e.g., IKONOS 1.0 m resolution) (Noronha et al., 2002) for pavement quality management. This is because high-resolution multispectral/hyper spectral satellite images can clearly observe/monitor the road conditions, such as loss of oily components, hydrocarbon absorption, pavement condition deterioration, exposing rocky components of the pavement, structural damages like cracking. The spectral signals for these pavement conditions can theoretically be reflected in high-resolution hyper spectral images (Herold et al., 2004; Cloutis, 1989). Lalitha (1989) and use an available enhancement technique to a Landsat TM (Thematic Mapper) urban scene to ascertain which technique is effective in improving the contrast of the road features in the image. Kelley (2002) used remote sensing technology for obtaining real-time pavement specific weather information, which was used as assistance for pavement. Moriyoshi (1999) described the infrared sensing analysis of asphaltic mixture and asphaltic pavement and presented the various application of infrared sensing analysis for civil engineering. Ayalew et al. (2003; 1998) identified spatial and spectral requirements for successful large-scale road feature extraction, and further examined the benefits of using hyper spectral imaging over traditional methods of roadway maintenance and rehabilitation for pavement management applications. Spagnolini and Rampa (1999) used monostatic ground penetrating radar (GPR) for pavement profiling, such as layer thickness. Guo et al. (2007) developed an algorithm for suburban road segmentation in high-resolution aerial images. Many researchers, such as Beaumont (1985) have demonstrated that information acquired from the interpretation of satellite imagery can play a significant role in the planning, management, and implementation of highway maintenance or rehabilitation. For example, Yoo, et al. (2005) used space-borne imagery of 1.0 meter high resolution with KOMPSAT-EOC to help road construction or repair planning. Lin et al. (2004) measured the concrete highway rough surface parameters by an X-Band scatterometer. To verify the test results, a laser profiler and a radar system were used to provide a direct measurement result. Irick and Hudson (1964) presented their research project, which contains principles and rules that can be used to design selected pavement sections and relate their behavior to similarly designed sections on the AASHO Road Test. In addition, they developed the guidelines to provide the basis for merging data of individual studies with data collected in the overall program, and provide means for translating road test findings to local conditions. Starks et al. (2002) used satellite image referencing algorithms to characterize asphaltic concrete mixtures. They demonstrated, from satellite imagery analysis that the corresponding mixture, the elasticity E depends on the frequency f in the range from 0.1 to 10^5 Hz. They measured the dependence of E on moderate frequencies f for different temperatures. Eckardt and White (1997) used Landsat thematic mapper to assist analysis of coarse gravel overlying a silty substrate. The silty material, known as a stone pavement, is prone to erosion.

Additionally, other research, such as Keaton and Brokish (2003), used IKONOS multispectral images to evaluate the evolving roads. Morain (2002) presented the application of image intelligence from space-based and aerial sensors for the critical infrastructure protection. He described "*America's transportation systems are predicated on economic, social, and political stability. After the epiphany of September 11, and subsequent national alerts, however, all sectors of transportation, not just in the USA, but around the world have become keenly aware of the vulnerabilities inherent in such systems; and of the cascading consequences that can arise from attacks at critical nodes in any one or more of the transportation sectors*." Liu et al. (2006) presented an algorithm for pavement cracking detection based on multi-scale space, since conventional human-visual and manual field pavement crack detection methods are very costly, time-consuming, dangerous, labor-intensive and subjective. A robust and high-efficient parallel pavement crack detection algorithm based on multi-scale space was presented.

1.3 Overview of Pavement Management System

A pavement management system (PMS) is a valuable tool and one of the critical elements of the highway transportation infrastructure (Tsai et al., 2004; Kulkarni et al., 2003). The earliest PMS concept can be traced back to the 1960s. With rapid increase of advanced information technology, many investigators have successfully integrated the Geographic Information System (GIS) into PMS for storing, retrieving, analyzing, and reporting information needed to support pavement-related decision making. Such an integration system is thus called G-PMS (Lee et al., 1996). The main characteristic of a GIS system is that it links data/information to its geographical location (e.g., latitude/longitude or state plane coordinates) instead of the milepost or reference-point system traditionally used in transportation. Moreover, the GIS can describe and analyze the topological relationship of the real world using the topological data structure and model (Goulias, 2002; Lee et al., 1996). GIS technology is also capable of rapidly retrieving data from a database and can automatically generate customized maps to meet specific needs such as identifying maintenance locations. Therefore, a G-PMS can be enhanced with features and functionality by using a geographic information system (GIS) to perform pavement management operations, create maps of pavement condition, provide cost analysis for the recommended maintenance strategies, and long-term pavement budget programming.

With the increasing amount of pavement data collected, updated and exchanged due to deteriorating road conditions, increasing traffic loading, and shrinking funds, many knowledge-based expert systems (KBESs) have been developed to solve various transportation problems (e.g., Abkowitz et al., 1990; Nassar, 2007; Spring and Hummer 1995; Zhang et al., 2001). A comprehensive survey of KBESs in transportation is summarized and discussed by Cohn and Harris (1992). However, only a few scholars have investigated applying data mining and knowledge discovery (DMKD) to PMSs. For example, Attoh-Okine (2002; 1997) presented application of Rough Set Theory (RST) to enhance the decision support of the pavement rehabilitation and maintenance. Prechaverakul and Hadipriono (1995) applied knowledge-based expert system and fuzzy logic for minor rehabilitation projects in Ohio. Wang et al. (2003) discussed the decision-making problem of pavement maintenance and rehabilitation. Leu et al. (2001) investigated the applicability of data mining in the prediction of tunnel support stability using an artificial neural networks (ANN) algorithm. Sarasua and Jia (1995) explored an integration of Geographic Information System Technology with knowledge discovery and expert system for pavement management. Ferreira et al. (2001) explored the application of probabilistic segment-linked pavement management optimization model. Chan et al. (1994) applied the genetic algorithm for road maintenance planning. Amado et al. (2002) applied knowledge discovery for pavement condition evaluation. Soibelman et al. (2000) discussed the data preparation process for construction knowledge generation through knowledge discovery in databases, as well as construction knowledge generation and dissemination.

1.4 Motivation

Zhou et al., (2010a; 2008a; 2008b) has initially investigated the application of the decision tree induction method in the decision-making of pavement maintenance and rehabilitation strategies. It is found that:

(1) The use of data mining and knowledge discovery methods for road maintenance and rehabilitation can largely increase the speed of decision-making, save time and money, and shorten the project period;

(2) The use of data mining and knowledge discovery for pavement management can make a consistent decision for road-network treatment strategies, thus avoid any human factors for decision-making of treatment.

(3) A decision tree is used to organize the obtained knowledge from experts in a logical order. Thus, decision trees can determine the technically feasible rehabilitation strategies for each road segment in a reasonable manner.

However, many shortcomings of applying traditional decision tree induction method have been discovered (see Chapter 2.4). Thus, *the motivation of this dissertation is to develop an innovative method for decision tree induction in order to overcome the shortcomings discovered, and to further enhance decision-making of the maintenance and rehabilitation strategies*.

On the other hand, the decision trees are based on the knowledge acquired from pavement management engineer for rehabilitation strategy selection. Thus, different decision trees can be built if the requirement changes. For example, the decision trees were based on severity levels of individual distresses in this research. If the pavement layer thickness and material type are taken as knowledge, or work history, pavement type, and ride data are taken as knowledge for generating decision-trees, these decision-trees are different. This means the decision rules generated by different knowledge are different. Thus, *motivation of this dissertation is to investigate and develop an "optimal" decision tree (decision rules) to largely enhance the decision-making of pavement treatment strategies.*

1.5 Enhances of Decision-Making in PMS

The highway transportation system is vital to the mobility of goods and people in USA and through worldwide. Pavements are an important component of the highway transportation infrastructure, accounting for the single largest share of the overall investment in highway infrastructure. Because of the large network of highways in each state, a tremendous amount of money is spent each year on the construction of new pavements and the maintenance and rehabilitation (M&R) of existing pavement. To maximize the benefits and minimize the overall costs associated with the process, a systematic and scientific approach is needed to manage the pavements. Many investigators have developed different system for effectively managing pavement and making reasonable decision in combination with GIS, data mining and knowledge discovery, artificial intelligences, etc. Thus, *successful implementation of this dissertation research will significantly enhance the decision-support of pavement management, maximize the benefits and minimize the overall costs of pavement management, since the proposed research attempts to find knowledge hidden in the pavement database.*

On the other hand, a pavement management system is a planning tool that is able to model pavement and surface deterioration due to the effects of traffic and pavement ageing, and contains a series of decision units used to determine how and when to repair the roads surface. Pavement management decisions need to integrate diverse spatially referenced data for decision-making. The data include geospatial data (e.g., XY coordinates, pavement width, number of lanes, width of lane, central line, etc), economic data (such as initial cost, total cost), pavement condition (such as skid resistance measurements, cracking, rutting, traffic counts, bridge conditions, etc.) and other data (such as construction history, sign inventories, and construction and maintenance records, etc.). Understanding the relationships between pavement condition data, street locations, and networks is very important for pavement management decision making. Thus, *successful implementation of this dissertation research will largely increase our understanding to the relationship between these pavement condition data and the street location and networks, since the proposed methods attempts to combine the spatial data (e.g., XY coordinates, etc.) and non-spatial data (e.g., distress data) for inducing co-location decision tree.*

1.6 Organization of the Dissertation

The organization of this dissertation is:

(1) Data mining techniques applied in pavement management is overviewed in Chapter 2.

(2) Spatial decision tree induction methods and their advantages and disadvantages are described in Chapter 3.

(3) Development of theory and algorithm for co-location decision tree induction is presented in Chapter 4.

(4) Data and experiment using traditional decision tree induction are presented in Chapter 5.

(5) Experiment design and experimental result analyses using the developed co-location decision tree induction are described in Chapter 6.

(6) Conclusion and future work are described in Chapter 7 and Chapter 8, respectively.

A list of published papers related to this dissertation in the duration of Ph.D. period is described in the end of dissertation.

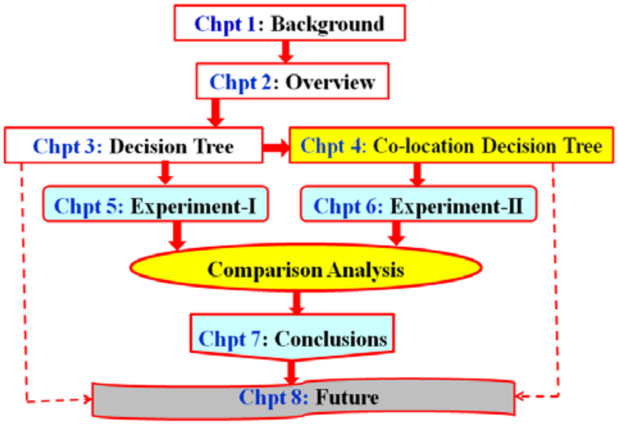


Figure 1.2 Relationship of the chapters

References for Chapter 1

AASHTO (2001): Pavement Management Guide. AASHTO, Washington, D.C., 2001. 64 Paper No. 02-3100, *Transportation Research Record,* vol. 1816.

Abkowitz, Mark; Walsh, Stephen; Hauser, Edwin; Minor, Larry (1990): Adaptation of geographic information systems to highway management, *Journal of Transportation Engineering*, vol. 116, no. 3, May-Jun, 1990, pp. 310-327.

Al-Turk, E.; Uddin, W. (1999). Infrastructure inventory and condition assessment using airborne laser terrain mapping and digital photography, *Transportation Research Record*, n 1690, 1999, pp. 121-125.

Ayalew, Balehager; Gomez, Richard; Roper, William; Carrasco, Oscar (2003). Pavement management using hyper spectral imagery *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 5097, 2003, pp. 207-214.

Ayalew, Balehager; Gomez, Richard; Roper, William; Carrasco, Shirotori, Takeo (1998). Route planning system using high-resolution commercial remote sensing data, *Quarterly Report of RTRI (Railway Technical Research Institute) (Japan)*, vol. 39, no. 4, December 1998, pp. 203-207.

Beaumont, T. E. (1985). Application Of Satellite Imagery For Highway Maintenance And Rehabilitation In Niger. *International Journal of Remote Sensing*, vol. 6, no. 7, Jul, 1985, pp. 1263-1267.

Cloutis Edward (1989). Spectral Reflectance Properties of Hydrocarbons: Remote-Sensing Implications, *Science*, 14 July 1989, vol. 245. no. 4914, pp. 165 – 168.

Cohn, L. F., and Harris, R. A. (1992). Knowledge-based expert system in transportation. *NCHRP Synthesis* 183, *TRB*, National Research Council, Washington, D.C.

DOT-NASA (2003): Remote Sensing and Geospatial Information Technologies Application to Multimodal Transportation 2003, *www.ncgia.ucsb.edu/ncrst/.../SynthRep2003/6pager-2003.pdf*

DOT-NASA (2002): Commercial Remote Sensing and Spatial Information Technologies Application to Transportation, a partnership for advancing transportation practice, a collaborative research Program, *Progress Report, January 2002. U.S. Department of Transportation, National Aeronautics and Space Administration. http //www.ncgia.ucsb. edu/ ncrst/synthesis/Brochure200201/brochure2001.pdf*

Eckardt, F.; White, K. (1997). Human induced disruption of stone pavement surfaces in the Central Namib Desert, Namibia: observations from landsat thematic mapper. *International Journal of Remote Sensing*, vol. 18, no. 16, Nov. 10, 1997, pp. 3305-3310.

Ferguson, C. Roger (1985). Transportation Application of Remote Sensing Information. *Technical Papers of the American Society of Photogrammetry, Annual Meeting*, vol. 2, 1985, pp. 642-650.

Ferreira, A., A. Antunes, and L. Picado-Santos (2001). Probabilistic Segment-Linked Pavement Management Optimization Model. *Journal of Transportation Engineering*, vol. 128, no.

Gilly, B.A., A. Touran, and T. Asai (1987). Quality Control Circles in Construction," *ASCE Journal of Construction Engineering and Management*, vol. 113, no. 3, 1987, pp. 432.

Guo, D.; Weeks, A.; Klee, H. (2007). Robust approach for suburban road segmentation in high-resolution aerial images, *International Journal of Remote Sensing*, vol. 28, no. 2, January, 2007, pp. 307-318.

Herold, M.; Roberts, D. (2005). Spectral characteristics of asphalt road aging and deterioration: implications for remote-sensing applications, *Applied Optics*, vol. 44, no. 20, 10 July 2005, pp. 4327-4334.

Herold, M., Roberts, D., Gardner, M. and P. Dennison (2004). Spectrometry for urban area remote sensing: Development and analysis of a spectral library from 350 to 2400 nm, *Remote Sensing of Environment*, vol. 91, no. 3-4, Jun 30, pp. 304-319.

Irick, P.E.; Hudson, W.R. (1964). Guidelines for satellite studies of pavement performance, *National Research Council -- Highway Research Board -- NCHRP Report*, 1964, 2A, 182p.

Karimi, H.A., X. Dai, S. Khorram, A. J. Khattak, and J.E. Hummer (1999). Techniques for Automated Extraction of Roadway Inventory features from High-Resolution Satellite Imagery. *Geocarto International*, vol. 14, no. 2, June, pp. 5-16.

Keaton, T.; and J. Brokish, (2003). Evolving roads in IKONOS multispectral imagery, *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, pp. III-1001-4 vol.2, 2003

Kelley, Joe R. (1993). Pavement weather sensing for the transportation industry, *Sensors (Peterborough, NH)*, vol. 10, no. 2, Feb, 1993, pp. 14-20.

Lalitha, L. (1989). Technique for road detection from high resolution satellite images, *Digest - International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 4, 1989, pp. 2246-2249.

Leu, Sou-Sen; Chee-Nan Chen; Shiu-Lin Chang, (2001). Data mining for tunnel support stability: neural network approach, *Automation in Construction*, vol. 10, no. 4, May 2001, pp. 429-41.

Lin, Jiangtao; Liu, Ce Richard; Li, Jing; Chen, Xuemin (2004). Measurement of concrete highway rough surface parameters by an X-Band scatterometer. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, June, 2004, pp. 1188-1196.

Liu, Xiang-Long; Li, Qing-Quan (2006). An algorithm to pavement cracking detection based on multi-scale space, *The International Society for Optical Engineering*, vol. 6419, *Geoinformatics 2006: Remotely Sensed Data and Information*, 2006, pp. 64190X.

Morain, Stanley A. (2002) Critical Infrastructure Protection Using Image Intelligence from Space-based and Aerial Sensors: ASME International Mechanical Engineering Congress and Exposition, Nov 17-22 2002, New Orleans, LA, *Transportation: Making Tracks for Tomorrows Transportation*, 2002, pp. 159-168.

Moriyoshi, Akihiro (1989). Application of infrared sensing analysis for civil engineering, *Doboku Gakkai Rombun-Hokokushu/Proceedings of the Japan Society of Civil Engineers*, n 409, pp. 177-180, (Japanese)

Noronha, V., M. Herold, D. Roberts, and M. Gardner (2002). Spectrometry and Hyperspectral Remote Sensing For Road Centerline Extraction And Evaluation of Pavement Condition, Proceedings of the Pecora Conference, *Proceedings of the Pecora Conference*, Denver, CO, November 2002.

Oscar, E. (2003). Pavement management using hyper spectral imagery, *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 5097, 2003, pp. 207-214.

Spagnolini, U. and Rampa, V., (1999). Multitarget detection/tracking for monostatic ground penetrating radar: Application to pavement profiling. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 1, part. 2, pp. 383-394.

Spring, G. S., and Hummer, J. (1995). Identification of hazardous highway locations using knowledge-based GIS: A case study. *Transportation Research Board,* No. 1497, Washington, D.C., pp. 83-90.

Shauna L. Hallmark, Kamesh Mantravadi, David Veneziano, Reginald R. Souleyrette, (2001). Evaluating remotely sensed images for use in inventorying roadway infrastructure features, Center for Transportation Research and Education, The Iowa State University, Final Report May, 20 p.

Starks, Scott A.; Nazarian, Soheil; Kreinovich, Vladik; Adidhela, Joseph (2002). Use of satellite image referencing algorithms to characterize asphaltic concrete mixtures, *IEEE International Conference on Plasma Science*, vol. 1, 2002, May 12-17 2002, Honolulu, HI, pp. 536-540.

Usher, J., and D. Truax (2001). Exploration Of Remote Sensing Applicability Within Transportation, NASA, Stennis Space Center, June 2001, 157p.

Veneziano, David; Hallmark, Shauna L.; Souleyrette, Reginald R.; Mantravadi, Kamesh, (2002). Evaluating remotely sensed images for use in inventorying roadway features, *Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering*, 2002, pp. 378-385.

Yoo, Hee-Young; Lee, Kiwon; Kwon, Byung-Doo (2005). Transportation analysis with high-resolution satellite imagery by wavelet analysis scheme, *IEEE International Geoscience and Remote Sensing Symposium*, 2005, pp. 1184-1187.

Zhang, Zhanmin; Smith, Stephen G.; Hudson, W. Ronald (2001). Geographic information system implementation plan for pavement management information system: Texas department of transportation. *Transportation Research Record*, n 1769, 2001, pp. 46-50.

Zhou, G., and L. Wang (2010a). GIS and Data Mining to enhance pavement rehabilitation decision-making, *Journal of Transportation Engineering,* vol. 136, no. 4, February, 2010, pp. 332-341.

Zhou, G., L. Wang (2010b). Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation, *submitted to Journal of Transportation Engineering,* March, 2010.

Zhou, G. and Jingyu Wei (2009a). Survey and analysis of land satellite remote sensing applied in highway transportations infrastructure and system engineering, *American Society for Photogrammetry and Remote Sensing (ASPRS) Annual Meeting*, Baltimore, MD, March 9 – 13, 2009.

Zhou, G., and L. Wang(2009b). Analysis of Flexible Pavement Distresses on IRI Model, *Pavements and Materials: Modeling, Testing, and Performance,* pp. 150-160, [ed] Zhanping You, Ala R. Abbas, and Linbing Wang, *American Society of Civil Engineering (ASCE), Geo Institute* (978-0-7844-1008-0), 2009, Reston, VA, ISBN: 498704679

Zhou, G., and L. Wang (2008a). Integrating GIS and Data Mining to Enhance the Pavement Management Decision-Making, *The 8th International Conference of Logistics and Transportation,* Chengdu, China July 31 –August 2, 2008.

Zhou, G., and L. Wang (2008b). 3D In-Vehicle Navigation Using Photorealistic Urban Model For Intelligent Transportation System, *87th TRB Annual Meeting*, Washington DC, January 13-17, 2008**.**

Zhou, G., and D. Wei (2008c), Traffic Spatial Measures and Interpretation of Road Network Using Aerial Remotely Sensed Data. *2008 IEEE International Geoscience and Remote Sensing Symposium* (*IGARSS 2008*), Boston, MA, USA, July 7-11, 2008

Zhou, G., and D. Wei (2008d). Survey and Analysis of Satellite Remote Sensing Applied in Transportations Infrastructure and System Engineering, *2008 IEEE International Geoscience and Remote Sensing Symposium* (*IGARSS 2008*), Boston, MA, USA, July 7-11, 2008.

2. LITERATURE REVIEW

2.1 Pavement Management

Pavement is one of the critical elements of the highway transportation infrastructure worldwide (Zhang et al. 2001). Although billions of dollars are spent annually on maintaining and rehabilitating pavements in the United States, deteriorating pavement conditions, increasing traffic loads, and limited funds present a complex challenge for pavement maintenance and rehabilitation activities (Zhang et al. 2001; Zhang et al. 1994). During the last several decades, Pavement Management Systems (PMS) have been developed to cope with these challenges in almost every state in order to identify maintenance needs, help to allocate funds, support cost-effective decision making, reduce the cost of pavement maintenance and rehabilitating activities under the constraint of limited funds.

The term “Pavement Management System” was first introduced to include the management of all aspects of pavement-related activities, including planning, design, maintenance, and rehabilitation of highway pavements (Zhang et al. 2001). Since introduction of the concept of pavement management system in 1960s, pavement management largely progressed in optimization of decision-making in the 1970s, and a significant accomplishment was implemented in the 1990s (Haas et al. 1994; AASHTO 2001). It is demonstrated that pavement management system (PMS) is a valuable tool, and can save money and maximize benefits for the highway system. It has become increasingly popular among local highway agencies, since many county and city government agencies have realized the benefits of having a decision-support system that helps them find cost-effective strategies for keeping their pavements in good condition (Fitch et al., 2001)), and maintain a highway system at an acceptable level of service that continues to support economic growth with a small amount of resources. (Tsai et al., 2004; Kulkarni et al., 2003; 1984)

With a significant advance in implementation of PMS in the 1990s, FHWA (Federal Highway Administration) mandated the adoption of PMS by all state departments of transportation (DOTs) in 1993 (Tsai and Lai, 2002; 2001). The purposes of a PMS are to identify maintenance needs, generate the pavement-rehabilitation plans, help to allocate funds, and maintain good pavement conditions under the constraint of limited funds, assist highway engineers and upper management in making consistent and cost-effective decisions related to maintenance and rehabilitation of pavements (Medina et al., 1999; Osman et al., 1994; Zhang et al., 1994). An effective PMS will therefore substantially save money and time, and maximize benefits for the highway management system.

FHWA mandated the adoption of PMS by all state departments of transportation (DOTs) in 1993. Several computer programs are available to help local communities develop their management systems. All roadway management problems are, to different degrees, geographical because they involve the spatial relations between objects and events. Road networks extend over a wide area and are affected by various land elements, such as rivers, mountains, buildings, and so forth. Since the required data always have a spatial component, the rational way to store and relate this information is through a spatial consistent-referencing system such as a geographic information system (GIS) (Hudson and Hass, 1995). Studies of linear reference systems (LRSs) provide ways to facilitate the information integration through a common spatial reference along the transportation network (Adams et al. 1998; Opiela 1997; Scarponcini 2001). Several studies (Hudson and Hass, 1995, pp. 102–103, 8–10) also mention the use of geographic information systems (GISs) to facilitate the transportation spatial information integration. Guo (2001, pp. 12–13) has further introduced the temporal reference for storing different pavement treatment methods applied to different pavement Chapters at different times along a common LRS. This allows DOTs to conduct the temporal analysis to study the interaction of different features, such as traffic accidents and pavement quality for providing better and more cost-effective management of maintaining pavement performance.

2.2 GIS Applied in Pavement Management

2.2.1 Two Basic Data Types in Transportation GIS (T-GIS)

GIS is defined as “a system of computer hardware, software and procedures designed to support the capture, management, manipulation, analysis, modeling, and display of spatially referenced data for solving complex planning and management problems.” (Chang, 2006; www.

disdevelopment.net/technology/gps/techgp0045c.htm) The main GIS characteristic is the potential of spatially linking information into its geographical location to (www.colorado.edu/ geography/gcraft/notes/intro/intro.html):

- Manage geographically-referenced information by integrating a database and mapping software,
- Provide the tools to analyze spatial relationships between events or phenomena,
- Allow us to view, understand, question, interpret, and visualize data in many ways that reveal relationships, patterns, and trends in the form of maps, globes, reports, and charts, and
- Be integrated into any enterprise information system framework.

GIS deals with the two basic types of data, vector data types and raster data types, both of which refers the data to a geographical coordinate system (e.g., latitude/longitude or state plane coordinates) instead of the milepost or reference-point system traditionally used in transportation. We call this, *geospatial data*.

A) Vector Data Types

Vector data is composed of discrete coordinates that can be used as points or connected to create lines and polygons (see Figure 2.1):

- *Point:* zero-dimensional objects, which have a position but no spatial extension.
- *Line:* one-dimensional objects with length as the only measurable spatial extension. Line Objects are built up of connected line segments.
- *Polygon:* two-dimensional or two and half-dimensional objects with area and perimeter as measurable spatial extension, which are composed of facet patches.
- *Body:* three-dimensional objects with volume and surface area as measurable spatial extensions, which are bordered by facets.

For the above four types of objects, *Point* is the basic geometric element. For example, a *Point* can present a point object, and also can be the start or end point of an *Edge*. The *Edge* is a line segment, which is an ordered connection between two points: begin point and end point. The *Facet* is the intermediate geometrical element. It is completely described by the ordered edges that define the border of the facet. *Entity* is the highest level geometrical element, and it can carry shape information, body object and DTM object. Each facet is related to an image patch through a corresponding link.

Corresponding to the above four types of vector data; they have examples in transportation have such as:

- Point: The simplest geospatial element, point, in transportation might be used to represent such as accident sites, traffic posts, traffic signs, or branches of road, or road interaction in small scale.
- Line: A line geospatial element, line, in transportation might be used to represent roads, transit routes, and so forth.
- Area: A polygon geospatial element, area, in transportation might be used for boundary data, such as traffic analysis zones, engineer districts, city limits, and so forth.
- Entity: A body geospatial element, entity, in transportation might be used for describing 3D transportation characteristics, car itself, accident event 3D reconstruction, and so forth.

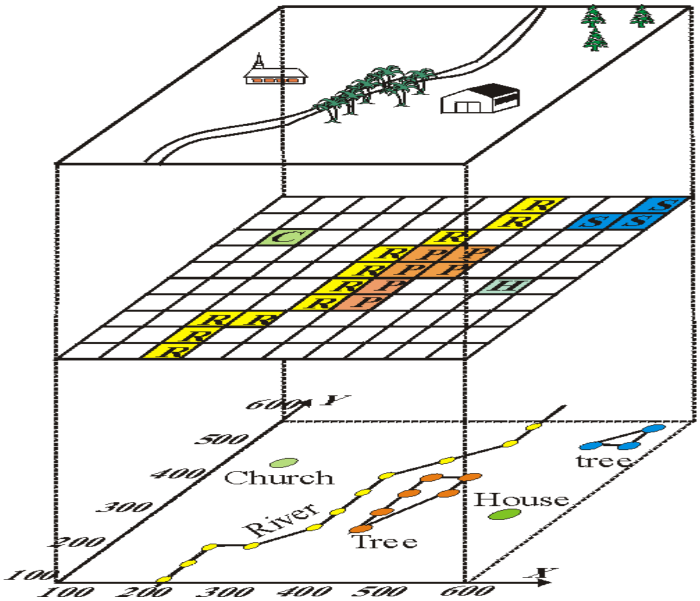


Figure 2.1: Vector data structures are based on elemental points whose locations are known to arbitrary precision, in contrast to the raster or cellular data structures.

B) Raster Data Types

Raster data represent features as a matrix of cells within rows and columns in continuous space. These cells are formed by pixels of a specific dimension size, and can be described as either "cell-based" or "image-based" data (see Figure 2.2):

- *Cell-based Data:* The cell size used for a raster layer depends on the requirement of the spatial analysis, map scale and the minimum mapping unit of the other GIS data. Using too large a cell size will cause some information to be lost. Using too small a cell size will significantly increase the storage space and processing time required, without adding precision to the map.
- *Image-based Data*: Image data ranges from satellite images and aerial photographs, to scanned maps.
- *Grid Data*: The grid provides the simplest way of dealing with the data. Grids speed the calculation time required for the computer to determine the location of the data points within the polygon. For example, elevation data are stored in this layer.

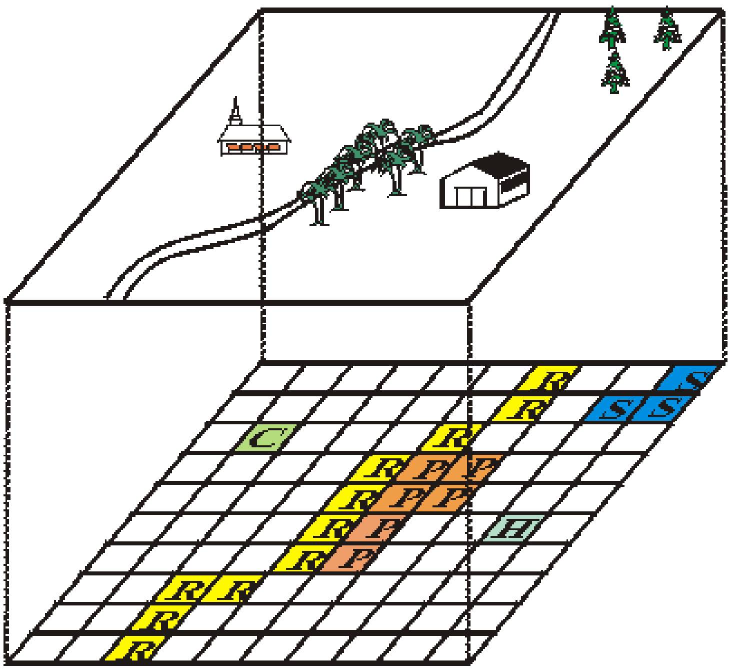


Figure 2.2: Raster data represent features as a matrix of cells within rows and columns in continuous space. As compared with Figure 2.1, the same real-world can be represented by two data types.

2.2.2 Integration of GIS into Pavement Management System (PMS)

With rapid increase of advanced information technology, many investigators have successfully integrated the GIS (Geographic Information System) into PMS for storing, retrieving, analyzing, and reporting information needed to support pavement-related decision making. Such an integration system is thus called G-PMS (Lee et al., 1996). In fact, a GIS system provides capabilities with which all aspects of the PMS process can be built and be enhanced, including data collection, data storage, data analysis, data interpretation, data visualization (spatial, and nonspatial data), system assessment, determination of strategies, project identification and development, and project implementation (Tsai et al., 2000; Zhang et al., 2001; Abkowitz et al., 1990). The main characteristic of a GIS system is that it links data/information to its geographical location, i.e., geographical coordinate system (e.g., latitude/longitude or state plane coordinates) instead of the milepost or reference-point system traditionally used in transportation, which is fundamental when integrating separate databases (Medina et al., 1999). Moreover, GIS can describe and analyze topological relationship of real world using topological data structure and model, which relates the geographical elements and attributes by mathematical rules, concerned with contiguity, order, and relative position (Goulias, 2002; Goulias et al. 2000; Lee et al., 1996). GIS is also capable of rapidly retrieving data from database and automatically generating customized maps to meet specific needs such as identifying maintenance locations. The attribute data manipulated in GIS is basically the same as those used in any traditional pavement management database, e.g., width of roads, number of lanes, condition of pavement, and history of construction and maintenance. Thus, the attribute data in the pavement management system can be stored in the GIS database by location and attribute (Harter 1998). So a G-PMS can be enhanced with features and functionality by using a geographic information system (GIS) to perform pavement management operations, create maps of pavement condition, provide cost analysis for the recommended maintenance strategies, and long-term pavement budget programming. Zhou et al., (2010a; 2010b; 2009; 2008a) have initially investigated the application of decision tree method in the pavement maintenance and rehabilitation. It has been demonstrated that the decision tree induction method using only attribute data cannot completely make correct decisions for rehabilitation and maintenance strategies, thus it is suggested that the spatial data should be used as well in combination with non-spatial data.

2.3 Data Mining and Knowledge Discovery Applied in Pavement Management

Several data mining techniques have been developed over the last decade in artificial intelligence community. Generally, the data mining techniques can be categorized in four categories, depending on their functionality: classification, clustering, numeric prediction, and association rules (Michalski, 1983; Tan et al., 2001).

- *Classification:* Generates predictive models for analyzing an existing database to determine categorical divisions or patterns in the database. It is focused on identifying the characteristics of the group or class to which each record belongs in the database.
- *Clustering:* Is to group or class the items that seem to fall naturally together in the database when there is no pre-identified class or group.
- *Numeric Prediction:* A classification learning technique, whose outcome is a numeric value (numeric quantity) rather than a category (discrete class). Thus, the numeric values are used for prediction.
- *Association Rule:* Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules shows attribute value conditions that occur frequently together in a given dataset. Association rules provide information of this type in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature.

The main difference between the above techniques depends on the method that (algorithms and methods used) is used to extract knowledge and how the mined knowledge and discovery rules are expressed. Many inductive learning algorithms, which mainly come from the machine learning, have been presented, such as AQ11 (Michalski, et al., 1975), AQ15 (Michalski et al., 1986 and Hong et al., 1995) and AQ19 (Kaufman et al., 1999), AE1 and AE9 (Hong, et al., 1995), CLS (Concept Learning System) (Hunt et al., 1966), ID3, C4.5 and C5.0 (Quinlan, 1979; 1987; 1993), and CN2 (Clark, 1989), and so on. CLS (Hunt, 1966) used a heuristic look ahead method to construct decision trees. Quinlan extended CLS method by using information content in the heuristic function, called ID3. ID3 method adopts a strategy, called “divide and conquer”, and selects classification attributes recursively on the basis of information entropy. Quinland (1993) further developed his method, called C4.5, which only dealt with strings of constant length. C4.5 not only can create a decision tree, but induce equivalent production rules as well, and deals with multi-class problems with continuous attributes. C5.0 is an upgraded version of C4.5. This method requires the training data, which are usually constructed by several tuples, each of which has several attributes. Obviously, if the records in the database are taken as tuples and fields as attributes, C5.0 algorithm is easily realized in a pavement database. Thus, the knowledge discovered by C5.0 algorithm is a group of classification rules and a default class, and with each rule, there is a confidence value (between 0 and 1). The main advantage of the C5.0 method is that it can generate a decision tree and associate rule.

The application of DMKD technology in pavement management system would be able to discover any interesting patterns in the database by e.g., creating a decision tree or decision rules. Some knowledge is stored in a "shallow place of the database", which can be obtained by traditional database query operation, such as maximum and minimum width of the roads; some knowledge is hidden a "deep place", (e.g., pavement condition pattern), which cannot be obtained by simple operations, but can be mined by intelligent technology, such as data mining.

With a vast amount of collected, and frequently and continuously updating and exchanging pavement database because of the rapid accumulation of large pavement database, meanwhile, with the technical advances in machine learning research, database, and visualization technologies (Lee et al., 1996; Simkowitz, 1990), many knowledge-based expert systems (KBESs) have been developed to solve various transportation problems (Jia, 2000; Richie and Prosser 1990; Prechaverakul and Hadipriono 1995; Nassar, 2007; Sarasua and Jia 1995; Spring and Hummer 1995). A comprehensive survey of KBESs applied in transportation is summarized and discussed by Cohn and Harris (1992). However, only a few scholars have conducted investigations of applying data mining and knowledge discovery for PMSs in order to discover any hidden rules of patterns stored within the databases. Attoh-Okine (2002; 1997; 1993) and Chang et al. (2006) presented application of Rough Set Theory (RST) to enhance the decision support of the pavement rehabilitation and maintenance. Prechaverakul and Hadipriono (1995) applied a knowledge-based expert system and fuzzy logic for minor rehabilitation projects in

Ohio. Wang et al. (2003) discussed the decision-making problem of pavement maintenance and rehabilitation. Leu et al. (2001) investigated the applicability of data mining in the prediction of tunnel support stability using an artificial neural networks (ANN) algorithm. Sarasua et al. (1995) explored an integration of GIS-T with knowledge discovery and expert system for pavement management. Ferreira et al. (2001) explored the application of probabilistic segment-linked pavement management optimization model. Chan et al. (1994) applied the genetic algorithm for road maintenance planning. Amado et al. (2002) applied knowledge discovery for pavement condition evaluation. Soibelman et al. (2000) discussed the data preparation process for construction knowledge generation through knowledge discovery in databases, as well as construction knowledge generation and dissemination.

2.4 Statement of Problems Pertaining to Data Mining Applied in Pavement Management

Although many advantages, when applying the data mining and knowledge discovery (DMKD) in pavement management system, have been found (see Chapter 1.3), many shortcomings are also discovered on the basis of the initial experimental results (Zhou et al., 2010a; 2008a; 2008b). These shortcomings can be briefly described as follows:

(1) *Post-processing*: the DMKD method is not quite as smart as people's imagine, since it is based on severity levels of individual distresses. Consequently, the induced decision rules for pavement treatment rehabilitation and maintenance are not completely correct. So, post-processing for verification is needed.

(2) *Many leaves and nodes, and decision rules*: The current algorithms of decision tree induction, such as C4.5, produce many tree nodes and leaves, resulting in redundant individual decision rules. The organization of individual rules into a logically ordered decision rules is time-consuming, sometime, incorrect.

(3) *Attribute selection*: The current algorithms of decision tree induction, such as C4.5, produce a decision tree through selecting attribute data. This implies that the algorithm does not consider relationship among the attribute data, such as co-location, co-occurrence, and cross-correlation.

(4) *Spatial data:* The data set of pavement database includes geospatial data in addition to the attribute data. As known, these geospatial elements basically have three

characteristics: attributes, geographical location, and topological relationship. The non-spatial (attribute) data is basically the same as those used in any traditional database, e.g., condition of pavement, and history of construction and maintenance. Spatial data that links the geospatial elements to its geodetic position gives a map-based coordinate system, such as State Plane Coordinate System, to unify all data sets in the same reference. The topological data structure or topology relationship describes the spatial relationships between adjacent features, and uses x, y coordinates to identify the location of a particular point, line, or polygon. Using such data structures enforces planar relationships, and allows GIS specialists to discover relationships between data layers, to reduce artifacts from digitization, and to reduce the file size required for storing the topological data. Unfortunately, the two major characteristics of spatial data in current decision tree induction method have not been considered.

2.5 Objectives of This Dissertation Research

With the shortcomings above, the primary objectives of this dissertation research are as follows:

1) Develop an advanced decision tree induction method to enhance the decision-making of rehabilitation and maintenance strategies,
2) Exploit the combination between co-location mining algorithm and decision tree induction method, and pioneer its application in rehabilitation and maintenance treatment strategies,
3) Integrate the pavement spatial data and non-spatial data into decision-making of rehabilitation and maintenance to minimize the cost and inconsistent decision of rehabilitation and maintenance strategies made by human, and
4) Integrate data mining technology into GIS system to graphically display treatment decisions, visualize the attribute and non-attribute data and link the data and information to the geographic coordinates in order to enhance the pavement management decision.

References for Chapter 2

AASHTO (2001): Pavement Management Guide. AASHTO, Washington, D.C., 2001. 64 Paper No. 02-3100 *Transportation Research Record,* vol. 1816.

AASHTO (1999): AASHTO guidelines for pavement management system. (1990). *American Association of State Highway and Transportation Officials,* Washing ton, D.C.

Abkowitz, Mark; Walsh, Stephen; Hauser, Edwin; Minor, Larry (1990): Adaptation of geographic information systems to highway management, *Journal of Transportation Engineering*, vol. 116, no. 3, May-Jun, 1990, pp. 310-327.

Adams, T., A. P. Vonderohe, and J. A. Butler (1998). Multimodal, Multidimensional Location Referencing System Modeling Issues. *Presented at NCHRP 20-27(3) Workshop on Functional Specifications for Multimodal, Multidimensional Transportation Location Referencing Systems*, December, 3–5, 1998.

Amado, Vanessa; Bernhardt, Kristen L. Sanford (2002): Expanding the use of pavement condition data through knowledge discovery in databases, *Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering*, 2002, pp. 394-401.

Attoh-Okine, Nii O. (2002): Combining use of Rough set and artificial neural networks in Doweled-Pavement-performance modeling – a hybrid approach, *Journal of Transportation Engineering*, vol. 128, no. 3, 2007, pp. 270-275.

Attoh-Okine, Nii O. (1997): Rough set application to data mining principles in pavement management database, *Journal of Computing in Civil Engineering*, vol. 11, no. 4, Oct, 1997, pp. 231-237.

Attoh-Okine, B. N., and Martinelli, D. (1994): Belief function framework for handling uncertainties in pavement management system design decision making, *Transportation Research Board (TRB)*, no. 455, Washington, D.C.

Attoh-Okine, B. (1993). Potential application of Bayesian influence diagram in pavement management. *Proc.. 2nd International Symposium (IEEE) on Uncertainty Modeling and Anal. (ISUMA),* College Park, Md.

Chan, W. T., T. F. Fwa, and C. Y. Tan (1994). Road Maintenance Planning Using Genetic Algorithms. I: Formulation. *Journal of Transportation Engineering*, vol. 120, no. 5, 1994, pp. 693–709.

Chang, Kang-tsung (2006). Introduction to Geographic Information Systems, 3rd edition, Boston; McGraw Hill.

Chang, Jia-Ruey, R., Hung, Ching-Tsung, Tzeng, Gwo-Hshiung, and Hsiao Wen-Hshiung (2006). Pavement maintenance and rehabilitation decisions derivd by Rough Set Theory, *Joint international Conference on Computing and Decision Making in Civil and Building Engineering*, June 14-16, 2006, Montréal, Canada, pp. 3396-3405.

Chung, Hung Chi; Girardello, Roberta; Soeller, Tony; Shinozuka, Masanobu (2003): Automated management of pavement inspection system (AMPIS), *Proceedings of The International Society for Optical Engineering*, vol. 5057, 2003, pp. 634-644.

Clark, P. and Niblett, T. (1989). The CN2 Induction Algorithm, *Machine Learning* 3, pp. 261-283, 1989.

Cohn, L. F., and Harris, R. A. (1992). Knowledge-based expert system in transportation. *NCHRP Synthesis* 183, TRB, National Research Council, Washington, D.C.

Ferreira, A., A. Antunes, and L. Picado-Santos (2001). Probabilistic Segment-Linked Pavement Management Optimization Model. *Journal of Transportation Engineering*, vol. 128, no. 6, 2001, pp. 568–577.

Fitch, G. Michael; Anderson, John E. (2001). Use of digital multispectral videography to capture environmental data sets for Virginia Department of Transportation. *Transportation Research Record*, no. 1756, 2001, pp. 87-93.

Goulias, D.G. (2002). Management systems and spatial data analysis in transportation and highway engineering, *Management Information Systems*, 2002: Incorporating GIS and Remote Sensing, 2002, pp. 321-327.

Goulias, D.G.; Goulias, K.G. (2000). GIS in pavement and transport management, *Management Information Systems*, 2000, pp. 165-175.

Guo, B. A. (2001). Feature-Based Linear Data Model Supported by Temporal Dynamic Segmentation. Ph.D. thesis. University of Kansas, Lawrence, 2001.

Haas, R., W. R. Hudson, and J. Zaniewski (1994). *Modern Pavement Management.* Krieger Publishing Co., Malabar, Fla., 1994.

Harter, Gerald L. (1998). Integrated geographic information system solution for estimating transportation infrastructure needs: a Florida example, *Transportation Research Record*, no. 1617, September 1998, pp. 50-55.

Hong, J., Mozetic, I., Michalski, R.S. (1995). AQ15: Incremental Learning of Attribute-Based Descriptions from Examples, the Method and User's Guide, *Reports of the Intelligent Systems Group*, University of Illinois at Urbana-Champaign, ISG 86-5, May, 1986.

Hunt, E. B. Marin, J. and Stone, P. T. (1966). Experiments in Induction, *Academic Press*, New York, N.Y., 1966.

Hudson, W. R., and R. C. G. Haas (1995). Future Directions and Need for Innovation in Pavement Management. *Conference Proceedings 1,* vol. 3, 1995, pp. 121–130.

Hudson, W. R., and S. W. Hudson (1995). Pavement Management Systems Lead the Way for Infrastructure Management Systems. *Conference Proceedings 1,* vol. 2, 1995, pp. 99–112.

Jia, X. (1996). A client/server-based intelligent GIS shell for transportation, PhD dissertation, Georgia Institute of Technology, Atlanta, GA.

Kaufman, K.A. and Michalski, R.S. (1999). Learning in an Inconsistent World: Rule Selection in AQ19, *Reports of the Machine Learning and Inference Laboratory*, MLI 99-2, George Mason University, Fairfax, VA, 1999.

Kulkarni, R. B., and R. W. Miller (2003). Pavement Management Systems: Past, Present, and Future. *Journal of the Transportation Research Board*, no. 1853, National Research Council, Washington, D.C., 2003, pp. 65–71.

Kulkarni, R. (1984). Dynamic decision model for a pavement management system. *Transportation Research Board (TRB)*, no. 997, National Research Council, Washington, D.C.

Lee, H.N.; Jitprasithsiri, S.; Lee, H.; Sorcic, R.G. (1996). Development of geographic information system-based pavement management system for Salt Lake City, *Transportation Research Record*, no. 1524, September 1996, pp. 16-24.

Leu, Sou-sen, Chee-Nan Chen, Shiu-Lin Chang (2001). Data Mining for tunnel support stability neural network approach, *Automation in construction*, vol. 10, no. 4, May 2001, pp. 429-441.

Medina, Alejandra; Flintsch, Gerardo W.; Zaniewski, John P. (1999): Geographic information systems-based pavement management system. A case study, *Transportation Research Record*, vol. 2, no. 1652, 1999, pp. 151-157.

Michalski, R.S, (1983). A Theory and Methodology of Machine Learning, in Michalski, R.S, Carbonell, J.G. and Mitchell, T.M. (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Tioga Publishing Company, 1983, pp. 83-134.

Michalski, R.S. and Larson, J. (1975). AQVAL/1 (AQ7) User's Guide and Program Description, *Report No. 731, Department of Computer Science*, University of Illinois, Urbana, June 1975.

Michalski, R.S., Mozetic, I., Hong, J., and Lavrac, N. (1986). The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains, *Proceedings of AAAI-86*, Philadelphia, PA, 1986, pp. 1041-1045.

Nassar K. (2007). Application of data-mining to state transportation agencies' projects databases*, ITcon,* vol. 12, pp. 139-149.

NCHRP (1991): NCHRP Research Results Digest No 180*: Implementation of Geographic Information Systems (GIS) in State DOTs.* TRB, National Research Council, Washington, D.C., 1991.

Opiela, K. S. (1997). A Generic Data Model for Linear Referencing Systems. *NCHRP Research Results Digest,* no. 218, *TRB, National Research Council*, Washington, D.C., 1997.

Osman, Omar; Hayashi, Yoshitsugu (1994). Geographic information systems as platform for highway pavement management systems, *Transportation Research Record*, no. 1442, Oct, 1994, pp. 19-30.

Paredes, Miguel; Fernando, Emmanuel; Scullion, T. (1990). Pavement management applications of GIS. A case study, *Transportation Research Record*, no. 1261, 1990, pp. 20-26.

Prechaverakul, S., and Hadipriono, F. C. (1995). Using a knowledge based expert system and fuzzy logic for minor rehabilitation projects in Ohio. *Transportation Research Board,* No. 1497, Washington, D.C., 19–26.

Primer: *GASB 34.* FHWA, U.S. Department of Transportation, 2000.

Quinlan, J. R. (1979): Discovering rules from large collections of examples: a case study. In D. Michie, editor, *Expert Systems in the Microelectronic Age,* Edinburgh University Press, Edinburgh, 1979.

Quinlan, J. R. (1987). Induction of decision trees. *Machine Learning*. 1987; vol. 1, pp. 81–106.

Quinlan, J.R. (1993). C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, CA.

Rasdorf, W., E. Shuller, R. Poole, O. Abudayyeh, and F. Robson (2000). Information Management at State Highway Departments: Issues and Needs. *Journal of Transportation Engineering,* vol. 126, no. 2, 2000, pp. 134–142.

Road Surface Management Manual*:* Pavement Condition Evaluation System (PACES*). Georgia Department of Transportation*, Atlanta, 1990.

Ritchie, Stephen G. and Prosser, Neil A. (1990). Development of a prototype real-time expert system for managing non-recurring freeway congestion, *VTT Symposium (Valtion Teknillinen Tutkimuskeskus)*, vol. 1, no. 116, 1990, pp. 129-153.

Sarasua, Wayne A.; Jia, Xudong (1995). Framework for integrating GIS-T with KBES: a pavement management system example, *Transportation Research Record*, no. 1497, July 1995, pp. 153-163.

Scarponcini, P. (2001). Linear Reference System for Life-Cycle Integration. *Journal of Computing in Civil Engineering,* vol. 15, no. 1, 2001, pp. 81–88.

Simkowitz, Howard J. (1990). Using geographic information system technology to enhance the pavement management process, *Transportation Research Record*, no. 1261, 1990, pp. 10-19.

Soibelman, Lucio; Kim, Hyunjoo (2000). Generating construction knowledge with knowledge discovery in databases, *Computing in Civil and Building Engineering*, vol. 2, 2000, pp. 906-913.

Spring, G. S., and Hummer, J. (1995). Identification of hazardous highway locations using knowledge-based GIS: A case study. *Transportation Research Board,* no. 1497, Washington, D.C., pp. 83-90.

Tan, Pang-Ning, Michael Steinbach and Vipin Kumar (2006). Introduction to Data Mining, Pearson Addison Wesley, ISBN 0-321-32136-7.

Tsai, Yichang; Gao, Bo; Lai, James S. (2004). Multiyear pavement-rehabilitation planning enabled by geographic information system: Network analyses linked to projects, *Transportation Research Record*, no. 1889, pp. 21-30.

Tsai, Yichang; Lai, James S. (2002). Framework and strategy for implementing an information technology-based pavement management system, *Transportation Research Record*, no. 1816, 2002, pp. 56-64.

Tsai, Y., and J. Lai (2001). Utilization of Information Technology to Enhance Asphalt Pavement Evaluation. *The International Journal of Pavement Engineering*, vol. 2, no. 1, 2001, pp. 17–32.

Tsai, Y., L. J. Lai, and Y. Wu (2000). Using Geographic Information Systems for Supporting Network-Level Pavement Maintenance Management. *Proc., 2nd International Conference on Decision Making in Urban and Civil Engineering,* Lyons, France, vol. 1, 2000, pp. 461– 471.

Tsai, Y., J. Lai, and L. Sun (1998). Developing a Prototype Knowledge-Based System in Identifying the Causes of Asphalt Pavement Distress. *Proc., 1st Conference on New Information Technologies for Decision Making in Civil Engineering,* Montreal, Quebec, Canada, vol. 1, 1998, pp. 515–526.

Vonderohe, A. P., L. Travis, R. L. Smith, and V. Tsai (1993). NCHRP Report 359: Adaptation of Geographic Information Systems for Transportation. *TRB, National Research Council*, Washington, D.C., 1993.

Wang, F., Z. Zhang, and R. B. Machemehl (2003). Decision-Making Problem for Managing Pavement Maintenance and Rehabilitation Projects. *Journal of the Transportation Research Board*, no. 1853, National Research Council, Washington, D.C., 2003, pp. 21–28.

Zhang, Zhanmin; Smith, Stephen G.; Hudson, W. Ronald (2001). Geographic information system implementation plan for pavement management information system: Texas department of transportation. *Transportation Research Record*, no. 1769, 2001, pp. 46-50.

Zhang, Zhanmin; Dossey, Terry; Weissmann, Jose; Hudson, W. Ronald (1994). GIS integrated pavement and infrastructure management in urban areas, *Transportation Research Record*, no. 1429, May 1994, pp. 84-89.

Zhou, G., and L. Wang (2010a). GIS and Data Mining to enhance pavement rehabilitation decision-making, *Journal of Transportation Engineering,* vol. 136, no. 4, February, 2010, pp. 332-341.

Zhou, G., L. Wang (2010b). Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation, *submitted to Transportation Research Part C, revised,* November 2010.

Zhou, G., L. Wang, (2009). Analysis of Flexible Pavement Distresses on IRI Model, *Pavements and Materials: Modelling, Testing, and Performance,* pp. 150-160, [ed] Zhanping You, Ala R. Abbas, and Linbing Wang, *American Society of Civil Engineering (ASCE), Geo Institute* (978-0-7844-1008-0), 2009, Reston, VA, ISBN: 498704679

Zhou, G., and L. Wang (2008). Integrating GIS and Data Mining to Enhance the Pavement Management Decision-Making, *The 8th International Conference of Logistics and Transportation,* Chengdu, China, July 31 –August 2, 2008.

3. SPATIAL DECISION TREE MODELING

3.1 Data Mining

3.1.1 Data Mining and Its Architecture

Data mining is the process of automatically extracting hidden useful information from large data repositories in order to find novel and useful patterns that might remain unknown. The data mining has become a powerful technology and tools for (Tan et al., 2006):

- Finding predictive information and pattern, future trends and behavior that experts may miss,
- Allowing decision-maker to make proactive, knowledge-driven decisions,
- Making prospective analyses and interpretability that are beyond the provision by retrospective tools, such as decision support systems, and
- Answering those questions that traditionally were too laborious and time-consuming to resolve.

Figure 3.1 illustrates a basic architecture of data mining including data collection, selection, transformation, mining and interpretation, and knowledge discovery. The starting point is a data warehouse, where the different types of attribute data and/or spatial data are collected and archived, and further managed in a variety of relational database systems. The data mining technology is integrated with the data warehouse to analyze these data using the data mining algorithms. The discovered knowledge in the last step is rendered to improve the whole process. Reporting, visualization, and other analysis tools can then be applied to plan the future actions and confirm the impact of those plans.

As seen from Figure 3.1, the process of data mining technology consists of a series of transformation steps, from data pre-processing to post-processing of data mining results. As mentioned in Tan et al. (2006), the input data can be in a variety of formats (e.g., flat files, spreadsheets, or relational tables), or be stored in a centralized data repository or be distributed across multiple sites connected by internet. The pre-processing includes fusing data from multiple sources, cleaning data to remove noise and duplicate observations, and selecting records and features that are relevant to the data mining task so that the raw input data can be transformed into an appropriate format for subsequent data mining analysis. The post-processing step is required in order to eliminate spurious data mining results so that only valid and useful results are incorporated into the decision support system. The post-processing algorithm includes such as statistical measures, hypothesis testing methods, etc. (Tan et al., 2006). Such a "closing the loop" form can ensure the final decision will be optimal as possible, since the mined information from database can be recycled and refined recursively.

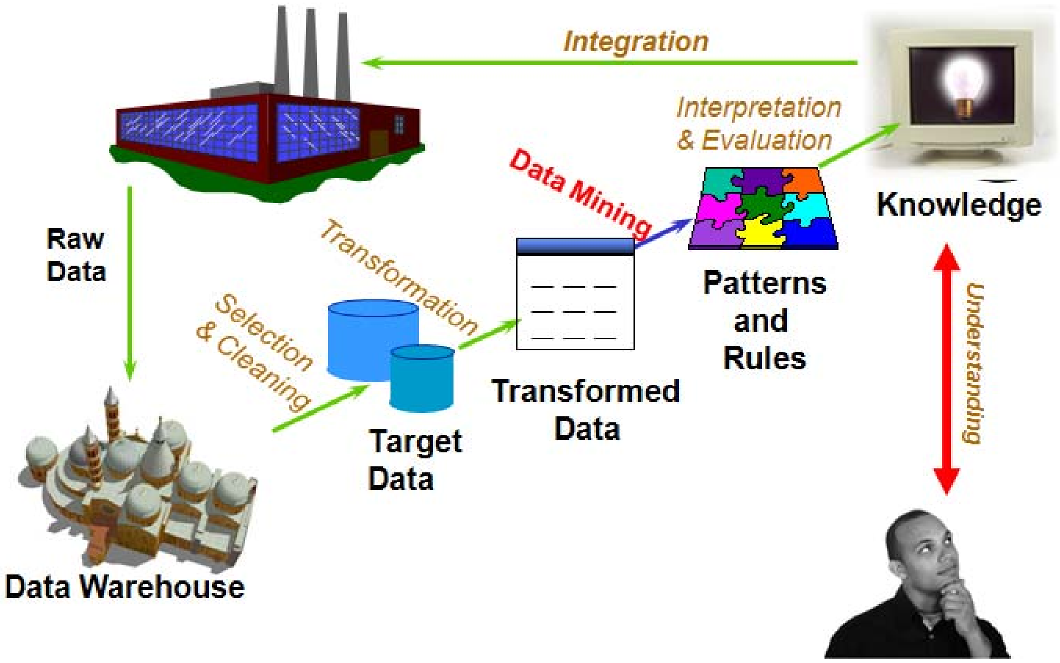


Figure 3.1: The basic architecture of data mining

3.1.2 Data Mining and Geographic Knowledge Discovery

Knowledge Discovery (KD) is a process including data warehousing, target data selection, cleaning, preprocessing, transformation and reduction, data mining, model selection (or combination), evaluation and interpretation, and use of the extracted knowledge (Fayyad 1996). Data mining is an integral part of knowledge discovery in databases (KDD) (Tan et al. 2006, page 3). Data mining aims to develop algorithms for extracting new useful patterns from database in which experts may miss, while Knowledge Discovery aims to enable an information system to transform information to knowledge through hypothesis testing and theory formation (Tan et al. 2006).

3.1.3 Data Mining with Other Disciplines

A number of other disciplines have played key supporting roles to development of data mining. The germinative idea of data mining was from such as sampling, estimation, and hypothesis testing from statistics, and search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning (Tan et al., 2006). With the advanced technologies in other disciplines, such as optimization, evolutionary computing, information theory, signal processing, visualization, spatial database, genetic algorithm, and information retrieval, the data mining obtained a sustainable development. In particular, techniques from high performance (parallel) computing are often important in addressing the massive size of some data sets, such as database system for efficient storage, indexing, and query processing. Distributed techniques can also help address the issue of size and are essential when the data cannot be gathered in one location. The most commonly used techniques in data mining are:

- Database technology,
- Information science,
- Statistics,
- Machine learning,
- Visualization, and
- Others such as
    - Artificial neural networks,
    - Decision trees,
    - Genetic algorithms,
    - Classification, and
    - Rule induction.

3.1.4 Data Mining Tasks

Mennis and Guo (2009) have summarized the common tasks in the spatial data mining. These tasks include (1) spatial classification and prediction, (2) spatial association rule mining, (3) spatial clustering regionalization and point pattern analysis, and (4) geo-visualization. Generally speaking, spatial data mining tasks are divided into two major categories (Tan et al., 2006):

- *Predictive tasks.* Like data mining, the major task of spatial data mining is to predict the values and behaviors of attributes on the basis of mined knowledge and pattern.
- *Descriptive tasks.* This task majorly describes the mined knowledge and spatial patterns, such as correlations, trends, neighbor, clusters, trajectories, co-location, co-occurrence, and anomalies.

3.2 Spatial Data Mining in Pavement Database

A pavement database is a type of spatial database, since the spatial data, such as XY coordinates, etc, are recorded for describing the pavement location. Thus, study of pavement management on the basis of pavement database should apply the spatial data mining technology. In fact, spatial data mining is a natural extension of data mining techniques applied to a spatial database. Spatial data mining is also used to extract the useful information and pattern in geography which is unknown and missed by exporters for offering great potential benefits for applied GIS-based decision-making. Thus, spatial data mining has the same objectives and goals as the data mining does, and even more. Many researchers in information technology (IT), digital mapping, remote sensing, geoinformatics, spatial science, and spatial databases have made tremendous efforts. These efforts include the development of theory, algorithm, methodology, and practice for the extraction of useful information and knowledge from geographically referenced spatial data and drive inductive approaches to geographical analysis and modeling (e.g., Andrienko and Andrienko, 1999; Chawla et al., 2000; Gahegan, 2003; Guo et al., 2008; Guo et al., 2006; Han et al., 1997; Keim et al., 2004; Knorr and Ng, 1996; Kulldorff, 1997; Mennis and Liu, 2005; Miller and Han, 2009; 2001; Openshaw et al., 1987; Shekhar et al., 2004; Yan et al, 2009; Yao and Thill, 2007; Zhang and Pazner, 2004; Huang et al., 2006; May and Savinor, 2002; Zhou et al., 2010b).

3.3 Comparison between Spatial Data Mining and Data Mining in Pavement Database

The common points between spatial data mining and data mining are they can share common method, algorithm, theory and practice. The differences of the two branches can be briefly summarized as follows (Zhou et al., 2010a; 2010b):

A) Spatial Data in Pavement Database

Data describing an object in pavement database consist of spatial data and non-spatial data. So-called spatial data generally consists of two basic properties, geometric and topological properties. The geometric properties can be spatial location (e.g., geodetic coordinates), area, perimeter, volume, etc. Meanwhile, the topological properties can be adjacency, inclusion, left/right hand side, clockwise/counter-clockwise, etc. In a traditional database, describing an object usually only uses non-spatial data, i.e., no spatial data. The non-spatial data can be stored and managed using a relational database where one attribute of an object has no spatial relationship (Aref and Samet, 1991). In pavement database, the object and event are described by spatial data simultaneously.

In addition, geographic attributes used for describing an object often exhibit the properties of spatial dependency and spatial heterogeneity (Yuan, 1997; Gahegan, et al., at http://www.ucgis.org/ priorities/ research/ research _white/ 2000%20Papers/emerging/gkd.pdf). The former implies that the attributes at some locations in space are related with others, the latter implies that most geographic processes are unstable, so that global parameters do not represent well the process occurring at a particular location (e.g., Glymour et al., 1997; Han et al., 1993; Hornsby and Egenhofer, 2000; Lu et al., 1996; Ng et al., 2002).

These distinct features present challenges and bring opportunities for mining useful information and spatial pattern from non-spatial and/or spatial properties of pavement treatment strategies. Thus, decision tree induction and decision rules induction for pavement management should consider both spatial data and non-spatial data simultaneously. Thus, if ignoring the properties of spatial dependency and spatial heterogeneity, the accuracy of pavement treatment strategies derived from data mining techniques will be affected.

B) Spatial Database

A pavement database is a type of spatial database. The primary methods for spatial data mining focus on the spatial database, which stores spatial objects represented by spatial data, non-spatial data, and spatial relationships (Han et al., 1993; Agrawal et al., 1993). In addition to extraction of hidden knowledge, spatial pattern, and information, spatial data mining, or knowledge discovery is also the extraction of implicit spatial relations that are not explicitly stored in spatial databases (Koperski and Han, 1995). Also the most studies of spatial data mining focus on the relational and transactional databases. The methods strived to combine the already mature techniques in such as machine learning, databases and statistics (Han et al., 1993; Ng and Han, 2002).

The fundamental idea of spatial data mining is on the basis of spatial data of pavement database, which has some characteristics and bring more challenging than the tradition data mining. Existing traditional data mining methods may not have been sufficient to deal effectively with geospatial data, since it can change in spatial and temporal domain. Thus this research considers the characteristics of spatial data's co-location and co-occurrence.

3.4 Decision Trees and Decision Rules

3.4.1 Decision Tree Induction

Decision tree (DT) induction is one of the most popular and powerful data mining techniques, and have thus widely applied in various pattern classifications (Chandra and Varghese, 2009; Witten and Frank, 2000). A decision tree can be understood as a type of classifier, which classifies the data set using a tree structure representation of the given decision problem (Osei-Bryson, 2007), and is usually composed of three basic elements (Tan et al., 2006) (see Figure 3.2):

(1) **A root node,** which is also called *decision node*. It has no incoming edges and zero or more outgoing edges,

(2) **Internal nodes**, which is also called *edge*, each of which has exactly one incoming edge and two or more outgoing edges, and

(3) **Leaf,** which is also called *terminal node* or *answer node*, each of which has exactly one incoming edge and no outgoing edges.

Over the past few decades, a lot of efforts have been made on how to construct an "optimal" DT. Dietterich (1990) discussed improvement to decision tree design methods, and provided a good background to these and more classical decision tree development methods. Lim, et al. (1998) compared several decision trees, such as statistical and neural network methods on a variety of datasets. Both of these works showed that a wide range of speed and accuracies can be obtained from the different decision tree algorithms commonly used, and that the effectiveness of different algorithms varies greatly with the dataset. One of the most commonly 'benchmark' methods of inducting decision tree structure is ID3 (Interactive Dichotomizer 3) (Quinlan, 1986) and C4.5 (Quinlan, 1993), which deals with datasets in which variables are continuous or integer, or where there is missing data, and CART (Classification And Regression Trees) algorithm (Breiman et al., 1984). These algorithms are typically called Top-Down Induction on Decision Trees (TDIDT), with which the knowledge obtained in the learning process is represented in a tree where each internal node contains a question about one particular attribute (corresponding decision variable) and each leaf is labeled with one of the possible classes (associated with a value of the target variable) (Osei-Bryson, 2007). The typical algorithm also includes; SLIQ (Mehta, et al. 1996), PUBLIC (Rastogi and Shim, 1998), SPRINT (Shafer et al., 1996), RAINFOREST (Gehrke et al., 2000), BOAT (Gehrke et al., 1999), MMDT (Chen et al., 2003), and TASC (Chen et al., 2006). In addition, Friedman et al. (1996) discussed the problems of constructing decision trees, and showed that the problem of constructing a decision becomes harder as one deals with larger and larger data sets, and with more and more variables. Fulton et al. (1996) analyzed the problems of generating decision trees capable of dealing with large, complex data sets, and showed that it is simpler to construct decision trees that can deal with a small subset of the original data set. Alsabti et al. (1998) discussed the problems of scaling decision trees up to large datasets, with the loss of accuracy that often occurs as a result. Mehta et al. (1996) emphasized the importance of classification in mining of large datasets, and also discussed the wide range of uses that classification can be put to in economic, medical and scientific situations. Garofalakis et al. (2000) discussed methods for constructing decision trees with user-defined constraints such as size limits or accuracy. These limits are often important for users to be able to understand or use the data sets adequately, or to avoid over-fitting the decision tree to the data that is available. Ankerst et al. (1999) used an interactive approach, with the user updating the decision tree through the use of a visualisation of the training data. This method resulted in a more intuitive decision tree and one that the user was capable of implementing according to their existing knowledge about the system in question. On the other hand, evolutionary computation for decision tree induction has been increasingly interested to many researchers. Li and Belford (2002) demonstrated that slight changes in the training set could require dramatic changes in the tree topology, i.e., the instability inherent in decision tree classifications. Llorà and Garrell (2001), and Papagelis and Kalles (2001) showed that evolutionary methods, when used to develop classification decision trees, allowed both important and unimportant attributes and relationships to be developed and for unimportant factors to be recognized. Cantú-Paz and Kamath (2000) meanwhile discussed an evolutionary method specifically used to develop classification trees, while Turney (1995) used a definition of fitness for decision tree evolution that included not only error rates but also other costs, such as size. Endou and Zhao (2002) examined a decision tree implementation method that relied on evolution of the training data set used. The training data set was evolved to give the best coverage of the domain knowledge. Siegel (1994) discussed the implementation of competitively evolving decision trees as a method of enhancing evolutionary methods.

Among these methods, one of the most common 'benchmark' methods, and also probably the most popular one is C4.5 algorithm developed by Quinlan (1986, 1993), which is based on the ID3 (Interactive Dichotomizer 3) method. Thus, this research will emphasize the analysis of the algorithm's advantages and disadvantages in order to presents our new method in Chapter 4.

3.4.2 Decision Tree Modeling

In principle, there are exponentially many decision trees that can be constructed from a given set of attributes (Tan et al., 2006), but investigators in fact only endeavor to find a most appropriate decision tree through making a series of locally optimum decisions about which attribute to use for partitioning the data during growing a decision tree, since the optimal tree is computationally infeasible because of the exponential size of the search space (Olaru and Wehenkel, 2003). This most appropriate decision tree is believed to be reasonably highest accurate, albeit suboptimal, and a reasonable amount of time. No matter which algorithm employed, the basic process of a decision tree usually consists of two major phases: the *growth phase* and the *pruning phase* (Aptė and Weiss, 1997).

3.4.2.1 Growth Phase

The basic process of growth phase is: a decision tree is generated in a top-down by successive divisions of the training set where each division represents a question about an attribute value.

The initial state of a decision tree is the root node that is assigned all of the attributes from the training set. If all attributes belong to the same class, then no further decisions need to be made to partition the attributes, and the solution is complete. If attributes at this node belong to two or more classes, then a split attribute operation will be made by a test. The process is recursively repeated for each of the new intermediate nodes until a completely discriminating tree is obtained (Aptė and Weiss, 1997). With the generated decision tree, each leaf node is assigned a class label. The non-terminal nodes, which include the root and other internal nodes, contain attribute test conditions to separate records that have different characteristics.

The above algorithm, i.e., starting from the root to the leaves, is called *generic decision tree algorithm*, which can be is briefly characterized by the following three properties (Elouedi et al., 2001):

(1) *Attribute selection measure.* How to choose an attribute is a critical issue because a most appropriate choosing will result in partitioning the training set in an *optimized* manner. When a decision node relative to this attribute is created after a test. This node becomes the root of the decision tree.

(2) *Partitioning strategy.* How to partition the training set with a given criterion or multiple criteria is very important. It consists in decomposing the training set into many subsets. In order to "optimally" partition the attributes, many criteria have been presented before; meanwhile many new criteria have still been being proposed.

(3) *Stopping criteria.* What criteria will be satisfied for stopping so that a training subset is declared as a leaf? This means that stopping criteria determines whether or not a training subset will be further divided. Some investigators applied the different steps recursively on the training subsets for verifying the stopping criteria.

One of the above most important properties is the attribute selection measurement, which measures how to select the attribute which characterizes the root of the decision tree and those of the different sub-decision trees. Quinlan (1993) has defined a measure called *information gain*, and further developed a well-known popular decision tree modeling algorithm, called *C4.5*. The details of attribute selection measures will be described in Chapter 3.4.3. Briefly, the basic idea of this attribute selection measure is to compute the information gain of each attribute in order to find how well each attribute alone classifies the training examples, and then one presenting the highest value will be chosen. In fact, this attribute generates a partition where the record classes are as homogeneous as possible within each subset created by the attribute.

In order to explain this basic process, Figure 3.2 illustrates the data set and corresponding tuple-table is listed in Table 3.1. The figure also shows the three axis parallel lines, one at height=4.2, the second line at height=6.3, and third line at volume=34. The three lines seems completely partition the training data set into three different sub-areas.

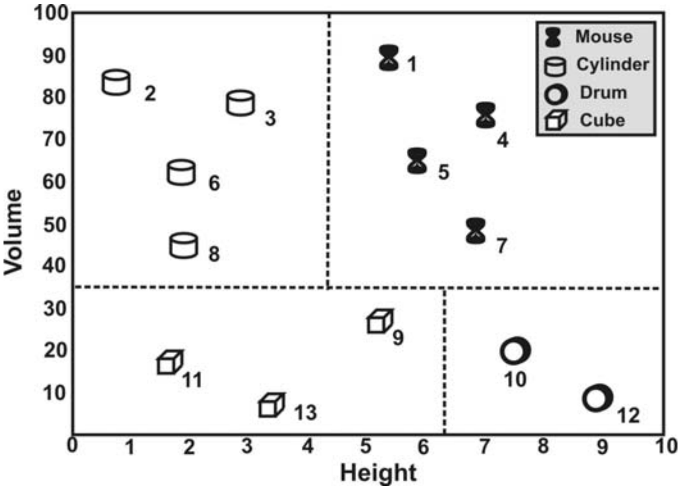


Figure 3.2 Example of data set for decision tree generation

Table 3.1: Example of data set for decision tree generation

ID	Height	Volume	Class
1	5.2	91.3	⧗
2	0.85	84.4	⛁
3	2.96	78.3	⛁
4	6.99	75.5	⧗
5	5.92	65.0	⧗
6	1.87	62.2	⛁
7	6.83	48.5	⧗
8	1.79	45.6	⛁
9	5.33	26.4	⧉

10	7.55	19.2	○
11	1.87	18.4	❒
12	8.99	8.9	○
13	3.41	8.2	❒

Figure 3.2 illustrates the process of a decision tree growth phase for training set listed in Table 3.1 and Figure 3.2. At first step, all attributes are assigned to the top level of the tree, i.e., root node, at which the classification process begins with a condition test for all examples at *volume* >34. Examples that satisfy this test conditions with TRUE are passed down to the left internal node, with FALSE are passed down to the right internal node. This means that the right edge from root node receives examples are not yet purely from one class so further testing is required at this intermediate node. The second test at this level for the left node (TRUE) is for *height* >4.2; and for the right node (FALSE) is for *height* >6.2. For the left node, examples that satisfy the test condition (*height* >4.2) are all in one class (MOUSE), and those that do not are all in another class (Cylinder). For the right node, examples that satisfy that test condition (*height* >6.2) are all in one class (DRUM), and those that do not are all in another class (CUBE). At this stage, both edges from this node lead to leaf nodes, i.e., no more tests are needed, thus the decision tree solution is complete. Note that this example illustrates a binary tree, where each intermediate node can split into at most two sub-trees. In fact, a decision tree may be non-binary tree, where each intermediate node may split into more than two sub-trees.

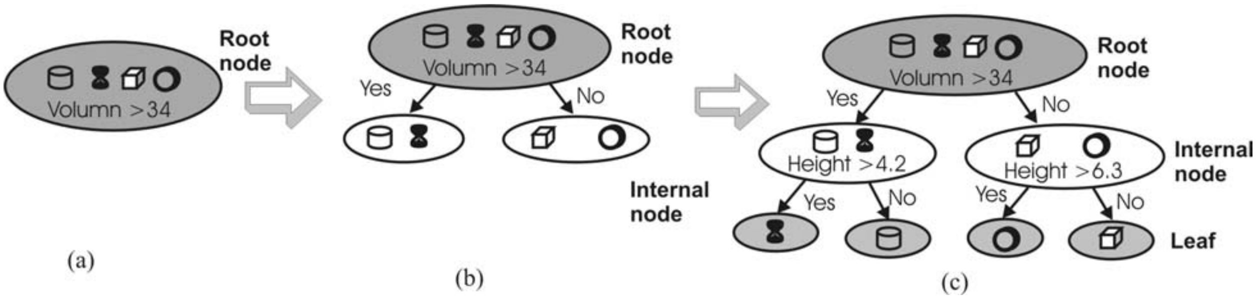


Figure 3.3 Process of tree growth phase

3.4.2.2 Pruning Phase

Due to noise and outliers in the training data, the generated decision tree at the above stage is potentially an over-fitted solution. The over fitting can heavily influence the classification accuracy of new datasets. Thus, a second phase, called *pruning,* is required to eliminate sub-trees in order to minimize the real misclassification error produced in growth phase (Aptė and Weiss, 1997). The actions of the *pruning phase* are often referred to as *post-pruning* in contrast to the *pre-pruning* that occurs during the *growth phase*. In order to create a small and interpretable decision tree, numerous post-pruning methods have been proposed (e.g. Almuallim, 1996; Bohanec and Bratko, 1994; Fournier and Cremilleux, 2002; Li et al., 2001; Mingers, 1989; 1987; Niblet and Bratko, 1986; Quinlan, 1986; 1987; 1993; Mansour, 1997; Mitchell, 1997; Elouedi et al., 2000; Säuberlich, 2000; Witten and Frank, 2000). These methods can be grouped by (Osei-Bryson, 2007)

- *Error-based method*. Some post-pruning approaches attempted to identify a sub-tree that gives the smallest error on the validation dataset, such as Reduced Error Pruning method proposed by Quinlan (1987); while others use an error estimation that is derived from training dataset only, such as Minimum Error Pruning method developed by Niblet and Bratko (1986);
- *Top-down or down-top method*. Some researchers propose a top-down approach, such as Pessimistic Error Method (Quinlan, 1987); while some researchers take a bottom-up approach, such as Error-Based Pruning method (Quinlan, 1993).
- *Optimal or sub-optimal method*. Some methods are sub-optimal heuristics (e.g. Mingers, 1987); some methods proposed to produce optimal solutions (e.g. Almuallim, 1996; Bohanec and Bratko, 1994).
- *Criterion method*. Some methods used signal criterion (e.g., Quinlan, 1987); some methods used a multi-criteria approach for evaluating the "best" DT in a set of generated DTs (e.g., Osei-Bryson, 2007; 2004).

3.4.3 Measures for Selecting the Best Split

Many measures have been developed to determine the best way to split the attributes based on the degree of impurity of the child nodes during growth phase of a decision tree. Most of these measures are defined in terms of the class distribution of the records before and after splitting. The smaller the degree of impurity, the more skewed the class distribution (Tan et al., 2006). The commonly used standard splitting measures are Entropy (Quinlan, 1986), Gain Ratio (Quinlan, 1993) and Gini Index (Breiman et al., 1984). The first two measures will be used in this research.

3.4.3.1 Entropy

In information theory, entropy is a measure of the uncertainty associated with a random variable. Also, the entropy is a measure of the average information content one is missing when one does not know the value of the random variable (http://en.wikipedia.org/wiki/Entropy (information _theory)). Entropy was first adopted in decision tree generation by Quinlan (1986) in his ID3 algorithm as split measure. The formula is (Tan et al., 2006)

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

Where $p(i|t)$ is the fraction of records belonging to class i at a given node t, and c is the number of classes. The ID3 algorithm utilized the entropy criteria for splitting nodes. The process is: Giving a node t, computing the splitting criterion, *Entropy (t)* = p_i*log(p_i), where p_i is the probability of class i within node t. An attribute and split are selected that minimize entropy. Splitting a node produces two or more direct descendants. Each child has a measure of entropy. The sum of each child's entropy is weighted by its percentage of the parent's cases in computing the final weighted entropy used to decide the best split.

3.4.3.2 Information Gain

For a training set T on attribute A, information gain in information theory and machine learning is defined as (Elouedi et al., 2001)

$$Gain(T, A) = Info(T) - Info_A(T) \qquad (3.1)$$

Where

$$Info(T) = \sum_{i=1}^{n} \frac{freq(C_i, T)}{|T|} \cdot \log_2 \frac{freq(C_i, T)}{|T|}, \qquad (3.2)$$

$$Info_A(T) = \sum_{v \in D(A)}^{n} \frac{|T_v|}{|T|} \cdot Info(T_v) \qquad (3.3)$$

where $\Theta = \{C_1, C_1, \cdots, C_n\}$ are the set of n mutually exclusive and exhaustive classes, $freq(C_i, T)$ denotes the number of objects in the set T that belong to the class C_i, and T_v is the subset of objects for which that attributes A has the values v.

Theoretically, the best attribute is the one that maximizes Gain (T, A). Once the best attribute is allocated to a node, the training set T is split into several subsets. The procedure is then iterated for each subset.

3.4.3.3 Gain Information Ratio

Elouedi et al. (2001) demonstrated that the Gain Information has good results, but it is limited to those attributes with a large number of values over those with a small number of values. To overcome this drawback, Quinlan (1993) has proposed gain ratio criterion, which is mathematically defined by

$$Gain\,ratio(T, A) = \frac{Gain(T, A)}{Split\ Info(T, A)} \qquad (3.4)$$

Where $Split\ Info(T, A) = -\sum_{v \in D(A)} \frac{|T_v|}{|T|} \cdot \log_2 \frac{|T_v|}{|T|}$ measures the information in the attribute due to the partition of the training set T into |D(A)| training subsets. Split Info (T, A) is also the information due to the split of S on the basis of the value of the categorical attribute A. With gain ratio, the attributes with many values will be adjusted.

In C4.5 algorithm (Quinlan, 1993), the attribute value that maximizes the Gain Ratio is chosen for the splitting attribute. The Gain Ratio is computed using attributes having Gain greater than Average Gain. This gain ratio expresses the proportion of information generated by a split that is helpful for developing the classification. The numerator (the information gain) in this ratio is the standard information entropy difference achieved at node t, expressed in Eq. 3.4.

3.4.4 Decision Rule Induction

Decision rules are directly induced by translating a decision tree either in a bottom-up specific-to-general style, or in a top-down general-to-specific style (Aptė and Weiss, 1997). In other words, the decision rules are constructed by forming a conjunct of every test that occurs on a path between the root node and a leaf node of a tree.

Algorithms of inducing decision rules can be grouped into two categories: ordered rule sets, or un-ordered rule sets (Aptė and Weiss, 1997):

(1) Ordered rule sets are induced by ordering all the classifications, and then using a fixed sequence, such as the smallest to the largest class, to combine them together. When this rule is applied to new data set, the new data example is required in exactly the same sequence as they were generated in the training data. Based on the example in Figure 3.3, the induced decision rules are depicted in Figure 3.4.

(2) Un-ordered rule sets are induced without a fixed sequence. Thus, when this rule is applied to new data, the new data example can be independent and more flexible.

For the above basic process of decision rule induction, i.e., a tree generation first, and then translation of the tree into a set of rules, discovered some problems. For example, for certain data spaces, this nature of partitioning may not always be capable of producing appropriate/optimal solutions. On the other hand, if algorithms are employed that directly generate tree, it is possible to create rules. These rules essentially correspond to decision regions that overlap each other in the data space. Thus, some people suggested the techniques that directly generate rules from data are also available, which overcome some of the drawbacks of decision tree modeling.

```
IF (volume > 34)
  THEN If (height > 4.2)
         Then MOUSE
         ELSE CYLINDER
ELSE IF
  IF (height > 6.3)
        THEN DRUM
         ELSE CUBE
End
```

Figure 3.4: Decision rule induction

3.4.5. Evaluation of the Performance of Decision Tree

Once a decision tree and/or decision rule is induced, it can be used for estimating or predicting new data set. Many methods have been developed to evaluate the performance of a decision tree or decision rules. The most well-known criteria are accuracy, speed, and interpretability. In other words, decision tree and decision rules derived using different approaches can be compared in terms of their predictive accuracy on the new data set, on the computational cost, and the level of understanding and insight that is provided by the solution. Accuracy and speed vary from algorithm to algorithm, and in most instances these two issues are coupled, i.e., a high predictive accuracy tends to require increased computational effort (Aptė and Weiss, 1997). This research will use the following criteria to evaluate the performance of decision tree and decision rules.

3.4.5.1 Accuracy of Performance

The performance accuracy of a decision tree is defined as a ratio between the number of correct or incorrect classified instances. The mathematical formula is (Tan et al., 2006):

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \qquad (3.5)$$

The above classification accuracy gives a general assessment of the number of correctly classified examples in total.

3.4.5.2 Two-Fold Cross-Validation

Two-fold cross-validation will be applied in this research to evaluate the performance of decision tree and decision rules. The basic process is: the whole dataset is split into 2 parts, one part of the dataset being dedicated to the training and the other one for the test. The training set is used to learn the algorithm and generate the tree and rules, and the test set is used to estimate the generated decision tree and rules. This procedure is repeated after every part of the dataset is used for both training and testing, respectively. Afterwards, the overall accuracy parameters were calculated as means from the evaluation of the individual cross-validation subset.

3.4.6 Problems of Decision Tree Induction for Spatial Data

Decision tree induction is capable of extracting implicit, previously unknown, and potentially useful information from large databases, and has therefore been successfully and widely used in various domains, including data mining (Quinlan, 1986, 1993), text mining (Yang and Pedersen, 1997), web intelligence (Cho et al., 2002; Zamir & Etzioni, 1998), and many other industrial and business domains for credit evaluations (Piramuthu, 1999), fraud detection (Bonchi et al., 1999), and customer-relationship management (Berry and Linoff, 2000).

The decision tree induction method has several advantages over other data mining methods including, being human-interpretable, well-organized, computationally inexpensive, and capable of dealing with noisy data (Breiman et al., 1993; Brodley and Utgoff, 1995; Duda et al., 2001; Durkin, 1992; Fayyad and Irani, 1992; Li et al., 2001).

However, the decision tree induction method entails the following drawbacks:

(1) Up to until now, decision tree construction algorithms have usually assumed that the class labels were Boolean variables. This means that the algorithms operate under the assumption that the class labels are flat. In other words, decision tree construction take each attribute through one-by-one manner without considering the simultaneous occurrence of multiple attributes. In real-world applications, there are more complex class scenarios, where the classification labels to be predicted are co-occurrence. Unfortunately, existing research has paid little attention to the classification of data with co-occurrence class labels. To the best of our knowledge, no method has been developed to construct DTs directly from data with co-occurrence class labels. This research work intends to remedy this research gap.

(2) Almost all of the decision tree generation methods did not consider the spatial features of geospatial data, such as geographic relationship and topological relationship. In other words, the spatial data contains objects which are characterized by a spatial location and/or extension as well as by several non-spatial attributes. Fig. 3.6 shows an example of spatial objects, which occur at a co-location pattern, i.e., CYLINDER always co-occurs with MOUSE. In a real-world, some instances are often located in close to geographically to another instance, such as gasoline station and road. Thus, identification of such a classification pattern, associated with spatial relationship and topological relationship, needs to be studied.

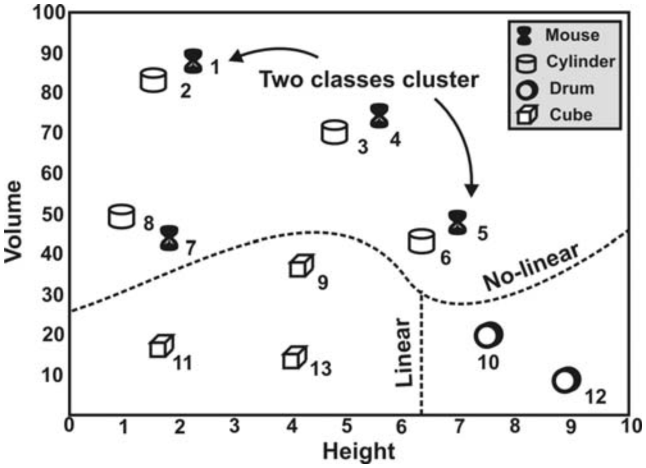


Figure 3.5: Instance co-location pattern and non-linear classification

(3) Mugambi et al. (2004) divided the decision trees into three main types on the basis of how they partition the feature space:

- *Uni-variate or axis-parallel decision tree*. This type of decision trees carries out tests on a single variable at each non-leaf node, and split the attributes using axis-parallel hyperplanes in the feature space (see Figure 3.2). C4.5 algorithm (Quinlan, 1993) belongs to the axis-parallel class of decision trees. This type of tree is called *Linear Decision Tree.*
- *Multivariate linear or oblique*. This type of decision trees carries out tests and split the attributes using an oblique orientation to the axis of the feature space geometrically.
- *Non-linear multivariate decision trees*. This type of decision trees carries out tests using non-linear partitioning of the feature space (see Figure 3.5), such as polynomial-fuzzy decision tree (Mugambi et al., 2004).

A linear decision tree is known to perform well in small and linear feature spaces but very poorly in large and non-linear ones. Theoretically, exploring information pattern using decision tree is based on a large database. In fact, in our pavement management database, the database is not large enough as expected in principle. This means that the pavement data mining only uses linear decision better than non-linear decision tree. However, the spatial features in the pavement database are not a linear mode in the real-world. Thus, this fact requires us to develop a robust linear decision tree method to handle small data with linear spatial feature.

References for Chapter 3

Agrawal, R., T. Imielinski, and A. Swami (1993). Mining Association Rules Between Sets of Items in Large Databases. *Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data*, Washington, D.C., May 1993, pp. 207-216.

Almuallim, H. (1996). An algorithm for the optimal pruning of decision trees. *Artificial Intelligence* 1996; vol. 83, no. 2, pp. 347–62.

Alsabti, K., Ranka, S., and Singh, V. (1998). CLOUDS: a decision tree classifier for large datasets. *In Proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98)*, August 27–31, New York City, New York, USA: AAAI Press, pp. 2–8.

Ankerst, M., Elsen, C., Ester, M., Kriegel, H.-P. (1999). Visual classification: an interactive approach to decision tree construction. *In Proceedings of international conference on knowledge discovery and data mining (KDD'99)*, San Diego, CA.

Andrienko, G., & Andrienko, N. (1999). Data mining with C4.5 and interactive cartographic visualization. In N. W. G. T. Paton (Ed.), User interfaces to data intensive systems. Los Alamitos, CA: IEEE Computer Society, pp. 162–165.

Aref, W. G. and H. Samet (1991). Optimization Strategies for Spatial Query Processing. In Proc. 17th Int. Conf. VLDB, Barcelona, Spain, September 1991, pp. 81-90,

Apté, Chidanand, Sholom Weiss (1997). Data mining with decision trees and decision rules, *Future Generation Computer Systems*, vol. 13, no. 2-3, November 1997, pp. 197-210

Bohanec, M. and Bratko, I. (1994). Trading accuracy for simplicity in decision trees. *Machine Learning* 1994, vol.15, pp. 223–50.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. California, USA: Wadsworth.

Chawla, S., Shekhar, S., Wu, W., & Ozesmi, U. (2000). Extending data mining for spatial applications: A case study in predicting nest locations. *In ACM SIGMOD workshop on research issues in Data Mining and Knowledge Discovery (DMKD 2000)*, Dallas, TX.

Cantú-Paz, E., & Kamath, C. (2000). Using evolutionary algorithms to induce oblique decision trees. In D. Whitley, D. E. Goldberg, E. Cantu´-Paz, L. Spector, I. Parmee, & H.-G. Beyer (Eds.), GECCO-2000: *Proceedings of the genetic and evolutionary computation conference,* San Francisco, CA: Morgan Kaufmann, pp. 1053-1060.

Chen, Y.L., Chen, Hsiao-Wei, and Hu, Kwei Tang (2009). Constructing a decision tree from data with hierarchical class labels, *Expert Systems with Applications*, vol. 36, no. 3, Part 1, April 2009, pp. 4838-4847.

Chen, Y. L., Hsu, W. H., & Lee, Y. H. (2006). TASC: Two-attribute-set clustering through decision tree construction. *European Journal of Operational Research*, vol. 174, no. 2, pp. 930–944.

Chen, Y. L., Tang, K., Shen, R. J., & Hu, Y. H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, vol. 40, no. 2, pp. 339–354.

Chen, Y. L., Hsu, C. L., & Chou, S. C. (2003). Constructing a multivalued and multi-labeled decision tree. *Expert Systems with Applications*, vol. 25, no. 2, pp. 199–209.

Dietterich, T. G. (1990). Machine learning. *Annual Review of Computer Science*, 4, 1990.

Dianhong, D. J. Liangxiao, An improved attribute selection measure for decision tree induction, *Proceedings of Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 4, 2007, pp. 654–658.

Elouedi, Zied, Khaled Mellouli, Philippe Smets (2001). Belief decision trees: theoretical foundations, *International Journal of Approximate Reasoning*, vol. 28, no. 2-3, November 2001, pp. 91-124.

Elouedi, Z., Mellouli, K., Smets, P., 2000. A pre-pruning method in belief decision trees. *In: Artificial Intelligence: Methodology, Systems, and Applications, 9th International Conference*, 80–90, Varna, Bulgaria.

Endou, T., & Zhao, Q. F., 2002. Generation of comprehensible decision trees through evolution of training data. *Proceedings of IEEE congress on evolutionary computation (CEC'2002)*, pp. 1221–1225.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, Fall 1996, pp. 37-54.

Friedman, J. H., Kohavi, R., & Yun, Y. (1996). Lazy decision trees. *Proceedings of the 13th national conference on artificial intelligence and eighth innovative applications of* *artificial intelligence conference*, Vol. 1. AAAI Press/The MIT Press. AAAI 96, IAAI 96, August 4–8, 1996, pp. 717–724.

Fulton, T., Kasif, S., Salzberg, S., & Waltz, D. (1996). Local induction of decision trees: towards interactive data mining. *Proceedings of Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR, August 1996, pp. 14–19.

Fournier, D. and Cremilleux, B. (2002). A quality index for decision tree pruning. *Knowledge-Based Systems* 2002, vol. 15, pp. 37–43.

Gahegan, M., Wachowicz, M., Harrower, M. and Rhyne, T. M. (2001). The integration of geographic visualization with knowledge discovery in databases and geocomputation. *Cartography and Geographic Information Systems*, special issue on the ICA research agenda.

Garofalakis, M., Hyun, D., Rastogi, R., & Shim, K. (2000). Efficient algorithms for constructing decision trees with constraints. *Proceedings of the 6th ACM SIGKDD international conference on knowledge discovery and data mining*, Boston, Massachusetts, August 2000, pp. 335–339.

Gehrke, J., Ganti, V., Ramakrishnan, R., & Loh, W-Y. (1999). BOAT—Optimistic decision tree construction. *Proceedings of the 1999 ACM SIGMOD international conference on management of data*, pp. 169–180.

Gehrke, J., Ramakrishnan, R., & Ganti, V. (2000). RainForest—A framework for fast decision tree construction of large datasets. *Data Mining and Knowledge Discovery*, vol. 4, no. 2–3, pp. 127–162.

Glymour, C., Madigan, D., Pregibon, D., and Smyth P. (1997). Statistical themes and lessons for data mining, *Journal of Data Mining and Knowledge Discovery*, vol. 1, pp. 11-28.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, vol. 22, no. 7, pp. 801–823.

Han, J., Koperski, K., & Stefanovic, N. (1997). Geominer: A system prototype for spatial data mining. *In ACM SIGMOD International conference on management of data*, Tucson, Arizona, USA, pp. 553–556.

Han, J., Y. Cai, and N. Cercone. Data-driven Discovery of Quantitative Rules in Relational Databases. *IEEE Trans. Knowledge and Data Eng.,* vol. 5, 1993, pp.29-40

Huang, Y., Pei, J., & Xiong, H. (2006). Mining co-location patterns with rare events from spatial data sets. *Geoinformatica*, vol. 10, no. 3, pp. 239–260.

Hornsby, K. and Egenhofer, M. J., (2000). Identity-based change: A foundation for spatiotemporal knowledge representation. *International Journal of Geographical Information Science*, vol. 14, pp. 207-224.

Jun, B.H. Jun, C.S. Kim, H. Song, A new criterion in selection and discretization of attributes for the generation of decision trees*, IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 12, 1997, pp. 1371–1375.

Keim, D. A., Panse, C., Sips, M., & North, S. C. (2004). Visual data mining in large geospatial point sets. *IEEE Computer Graphics and Applications*, vol. 24, no. 5, pp.36–44.

Kervahut, Tanguy and Jean-Yves Potvin (1996). An interactive-graphic environment for automatic generation of decision trees, *Decision Support Systems,* vol. 18, no. 2, October 1996, pp. 117-134.

Knorr, E. M., & Ng, R. T. (1996). Finding aggregate proximity relationships and commonalities in spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 884–897.

Koperski, K. and Han, J. (1995). Discovery of Spatial Association Rules in Geographic Information Databases, *Proc. 4th International Symposium on Large Spatial Databases*, SSD95, Maine, pp. 47-66.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics—Theory and Methods*, vol. 26, pp. 1481–1496.

Lee, H. and Ong, H. (1996). Visualization support for data mining. *IEEE Expert*, vol. 11, no. 5, pp. 69-75.

Li, X-B, Sweigart, J, Teng, J. Donohue, J., Thombs, L. (2001). A dynamic programming based pruning method for decision trees. *INFORMS Journal on Computing* 2001, vol. 13, no. 4, pp. 332–44.

Li, R.-H., & Belford, G. G. (2002). Instability of decision tree classification algorithms. *In Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining*, Edmonton, Alberta, Canada, pp. 570–575.

Lim, T.-S., Loh, W.-Y., & Shih, Y.-S. (1998). An empirical comparison of decision trees and other classification methods. *Technical report, no. 979. Department of Statistics, University of Wisconsin.*

Llorà, X., & Garrell, J. M. (2001). Evolution of decision trees. *In Proceeding of the 4th Catalan conference on artificial intelligence (CCIA'2001)*. ACIA Press.

Lu, H., Setiono, R., and Liu, H. (1996). Effective data mining using neural networks. *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 957-961.

Miller, H. J., (2000). Geographic representation in spatial analysis. *Journal of Geographical Systems*, vol. 2, pp. 55-60.

Mansour, Y., 1997. Pessimistic decision tree pruning based on tree size. In: van Someren, M., Widmer, G. (Eds.), *Proceedings of the 9th European Conference on Machine Learning (ECML-97)*. Springer Press, Berlin, Heidelberg.

May, M.; Savinov, A. (2002). An integrated platform for spatial data mining and interactive visual analysis, *Management Information Systems*, vol. 6, 2002, *Data Mining III,* pp. 51-60.

Mehta, M., Agrawal, R., & Rissanen, J. (1996). SLIQ: A fast scalable classifier for data mining. *In Proceedings of the 5th international conference on extending database technology*, Avignon, France, March 25–29, 1996, pp. 18–32.

Mennis, Jeremy; Guo, Diansheng (2009). Spatial data mining and geographic knowledge discovery-An introduction, *Computers, Environment and Urban Systems*, vol. 33, no. 6, November 2009, pp. 403-408.

Mennis, J., & Liu, J. W. (2005). Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS*, vol. 9, no. 1, pp.5–17.

Miller, H., & Han, J. (2009). Geographic data mining and knowledge discovery: An overview. In H. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery*. CRC Press, Taylor and Francis Group, pp. 1–26.

Miller, H. J., & Han, J. (2001). Geographic data mining and knowledge discovery: An overview. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery*. London and New York: Taylor & Francis, pp. 3–32.

Mingers, J. (1987). Expert systems—rule induction with statistical data. *Journal of the Operational Research Society*, vol. 38, pp. 39–47.

Mingers, J., 1989. An empirical comparison of pruning methods for decision tree induction. *Machine Learn*, vol. 3, no. 4, pp. 319–342.

Mugambi, E. M., Andrew Hunter, Giles Oatley, Lee Kennedy (2004). Polynomial-fuzzy decision tree structures for classifying medical data, *Knowledge-Based Systems*, vol. 17, no. 2-4, May 2004, pp. 81-87.

Ng, Raymond T.; Han, Jiawei (2002). CLARANS: A method for clustering objects for spatial data mining, *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 5, , September/October 2002, pp. 1003-1016.

Niblet, T. and Bratko, I. (1986). Learning decision rules in noisy domains. *Proceedings of Expert System*s 1986; vol. 86, pp. 25–34.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Science*, vol. 1, no. 4, pp. 335–358.

Osei-Bryson, Kweku-Muata (2007). Post-pruning in decision tree induction using multiple performance measures, *Computers & Operations Research*, vol. 34, no. 11, November 2007, pp. 3331-3345.

Osei-Bryson, Kweku-Muata (2004). Evaluation of decision trees: a multi-criteria approach, *Computers & Operations Research*, vol. 31, no. 11, September 2004, pp. 1933-1945.

Papagelis, A., and Kalles, D. (2001). Breeding decision trees using evolutionary techniques. *In Proceedings of the eighteenth international conference on machine learning (ICML 2001)*, pp. 393–400. Williams College, Williamstown, MA, USA: Morgan Kaufmann.

Quinlan, J.R., (1986). Induction of decision trees. *Mach. Learn.* 1, pp. 81–106.

Quinlan, J.R., (1987). Simplifying decision trees. *Int. J. Man-Machine Stud.* 27, pp. 221–234.

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

Rastogi, R., & Shim, K. (1998). PUBLIC: A decision tree classifier that integrates building and pruning. *Proceedings of 24th international conference on very large data bases*, pp. 404–415.

Shekhar, S., Zhang, P., Huang, Y., and Vatsavai, R. (2004). Trends in spatial data mining. In H. Kargupta, A. Joshi, K. Sivakumar, & Y. Yesha (Eds.), Data mining: Next generation challenges and future directions. AAAI/MIT Press, pp. 357–381.

Säuberlich, F., 2000. KDD und Data Mining als Hilfsmittel zur Entscheidungsunterst¨utzung [KDD and Data Mining as aid for decision support]. Dissertation Thesis. Peter Lang, Frankfurt a. Main, Germany.

Shafer, J. C., Agrawal, R., & Mehta, M. (1996). SPRINT: A scalable parallel classifier for data mining. *Proceedings of 22nd international conference on very large data bases*, pp. 544–555.

Siegel, E. V. (1994). Competitively evolving decision trees against fixed training cases for natural language processing. In K. Kinnear (Ed.), Advances in genetic programming. Cambridge, MA: MIT Press.

Tan, Pang-Ning, Michael Steinbach and Vipin Kumar (2006). Introduction to Data Mining, Pearson Addison Wesley, ISBN 0-321-32136-7.

Turney, P. D. (1995). Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, vol. 2, pp. 369–409.

Witten, I.H., Frank, E., 2000. Data mining—practical machine learning tools and techniques with Java implementation. Morgan Kaufmann, San Mateo, CA.

Wu, F., Zhang, J., and Honavar, V. (2005). Learning classifiers using hierarchically structured class taxonomies. *Proceedings of symposium on abstraction reformulation, and approximation,* pp. 313–320.

Yan, J., and Thill, J.C. (2009). Visual data mining in spatial interaction analysis with self-organizing maps. *Environment and Planning B*, 36, pp. 466–486.

Yao, X., and Thill, J.C. (2007). Neurofuzzy modeling of context–contingent proximity relations. *Geographical Analysis*, vol. 39, no. 2, pp. 169–194.

Yuan, M. (1997). Use of knowledge acquisition to build wildfire representation in geographic information systems. *International Journal of Geographical Information Systems,* vol. 11, pp.723-745.

Zhang, X., and Pazner, M. (2004). The icon image map technique for multivariate geospatial data visualization: Approach and software system. *Cartography and Geographic Information Science*, vol. 31, no. 1, pp. 29–41.

Zhou, G., and L. Wang (2010a). GIS and Data Mining to enhance pavement rehabilitation decision-making, *Journal of Transportation Engineering,* vol. 136, no. 4, February, 2010, pp. 332-341.

Zhou, G., L. Wang (2010b). Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation, *submitted to Transportation Research Part C,* November 2010.

Zhou, G., and L. Wang (2008). Integrating GIS and Data Mining to Enhance the Pavement Management Decision-Making, *The 8th International Conference of Logistics and Transportation,* Chengdu, China July 31 –August 2, 2008.

4. SPATIAL CO-OCCURRENCE DECISION TREE MODELLING

4.1 Introduction

Chapter 3 introduced the fundamental theory of decision tree generation including tree construction, algorithm modeling, and attributes splitting criteria, pruning, and accuracy evaluation of decision tree performance. This Chapter will present an innovative method, called *co-location spatial decision tree induction*, which is to incorporate co-location (also called co-occurrence) mining into the decision tree. This Chapter will describe the details of the co-location decision tree construction, algorithm, modeling, decision rule, node splitting criterion, node merging criterion, and leaf stopping criteria, and then will give an example for illustrating the calculation process of the proposed co-location spatial decision tree induction algorithm.

4.2 Co-Location Mining Algorithm

Huang et al. (2004) presented the first general framework of mining spatial co-location patterns. Afterwards, Huang and her research team made further research and exploration on how mining co-location rules can be applied in spatial data analysis, spatial data pattern classification and spatial geographic knowledge discovery. For example, Huang et al. (2005; 2006) adjusted the measure to treat the case with rare events, and Huang (2008) used density ratios of different features to describe the neighborhood constraint together with a clustering approach. Xiao et al. (2008) presented a density-based algorithm for mining a spatial co-location pattern, and Xiong et al. (2004) presented a buffer-based model to describe the neighborhood constraint for dealing with extended spatial object such as lines and polygons.

On the other hand, different researchers have made efforts to improve the efficiency of mining process of co-location. Yoo et al. (2004) proposed partial-join algorithm. Yoo et al. (2005, 2006) and Wang et al. (2008) proposed a join-less algorithm and N-most prevalent collocated event in 2009 (Yoo et al., 2009). Complex spatial co-location patterns are presented by Munro et al. (2003) and Verhein et al (2007). Sheng et al. (2008) introduced the definition of influence function based on Gaussian kernel to describe the neighborhood constraint, in which the algorithm assumed a distribution of features on the global space. Hsiao et al. (2006) applied the spatial data mining of co-location pattern for support agriculture decision-making. Zhang et al.

(2004) enhanced the algorithm proposed in Hunag et al. (2004) to mine special types of co-location relationships in addition to cliques, namely; the *spatial star*, and *generic patterns*. Celik et al. (2006a) proposed the problem of mining mixed-drove spatial-temporal co-occurrence patterns (MDCOPs) which extends the co-location pattern mining to the scope of both time and space. Afterward, Celik et al. (2006b) further considered some constraints based on the result of MDCOP and the most top-k ranking issues, and Celik et al. (2007a; 2007b) partitioned a global space into small zones and applied the co-location mining algorithms on every zone for accumulated computation. Eick et al. (2008) also proposed to find regional co-location patterns based on clustering. Qian et al. (2005) presented spatial co-location patterns with dynamic neighborhood constrain, and further spatial-temporal co-occurrence over zones (Qian et al., 2009).

4.2.1 Some Definitions of Co-Location Mining Algorithm

The basic concept of spatial co-location or called spatial co-occurrence implies the presence of two or more spatial objects at the same location or at significantly close distances from each other. Co-location patterns can indicate interesting associations among spatial data objects with respect to their non-spatial attributes. In these methods, the neighborhood constraint is described by a distance threshold which is the maximal distance allowed for two events to be neighbors. Mathematically, the co-location can be modeled by (Huang et al., 2004; Yoo et al., 2006):

Given

a) The training data is a set $S = \{s_1, s_2, \cdots, s_K\}$. Each sample $s_i = \{x_1, x_2, \cdots, x_N\}$ is a vector, representing example-id, spatial feature type, and location $\vee$, where location $\in$ spatial framework. The training data is augmented with a vector C = c_1,c_2,... where c_1,c_2,... represent the class to which each sample belongs.

b) A neighbor relation $\Re$ over examples in S.

We have

a) A co-location, C is defined as a subset of Boolean spatial features, $C \subseteq S$, whose instances form a clique under a neighbor relationship $\Re$. Usually, the neighbor relationship $\Re$ is a Euclidean distance metric. For example, if the two spatial objects satisfy the neighbor relationship, i.e., distance $(s_i, s_j) \leq d$, they are called *neighbor*. If an instance is co-location with another instance, the objects of all features forms a clique relationship in the co-location.

b) Accompanying with co-location mining process, a co-location rule can be formed, and expressed as $c_1 \Rightarrow c_2(p, cp)$, where $c_1 \subseteq S$, $c_2 \subseteq T$, and $c_1 \cap c_2 = \Omega$. p is a number representing the prevalence measure, and *cp* is a number measuring conditional probability (Huang et al., 2004).

With the modeling given above, it can be noted that an important part in the co-location is proximity neighborhood, which is expressed using neighbor relation, $\Re$. This relationship is based on the semantics of the application domains for forming a clique (Huang et al., 2006). For this reason, many researchers have presented different methods and algorithms to mode the neighbor relationship, $\Re$, such as:

- Spatial relationships (e.g. connected, adjacent in GIS (Xiong et al. (2004)),
- Metric relationships (e.g. Euclidean distance (Yoo et al., 2006)),
- Combined relationship (e.g. shortest-path distance in a graph such as a road-map), and
- Constrained relationship (e.g., Sheng et al. (2008), Qian et al. (2005))

It is also noted that the $\Re$-proximity neighborhood concept is different from the neighborhood concept in topology, since some sets of a $\Re$-proximity neighborhood may not qualify to be $\Re$-proximity neighborhoods (Huang et al., 2006).

In order to describe the co-location algorithm, we first give several definitions (Huang et al., 2004).

A) **Participation ratio**

The participation ratio $pr(c,s_i)$ for feature type s_i in a size-k co-location $c=\{s_1,s_2,\cdots,s_K\}$ is the fraction of instances of feature s_i $\Re$-reachable to some row instance of co-location $c-\{f_i\}$.

B) **Participation index**

The participation index $pi(c)$ of a co-location $c=\{s_1,s_2,\cdots,s_K\}$ is $min_{i=1}^{k}\{pr(c_1,f_i)\}$. The participation index is used as the measure of prevalence of a co-location. The participation ratio can be computed as:

$$pi(c)=\frac{\pi_{s_i}(|table\_ins\tan ce(c)|)}{|\,table\_ins\tan ce(f_i)\,|} \qquad (4.1)$$

Where π is the relational projection operation with duplication elimination.

C) **Conditional Probability**

The conditional probability $cp(c_1\Rightarrow c_2)$ of a co-location rule $c_1\Rightarrow c_2$ is the fraction of row instances of c_1 $\Re$-reachable to some row instance of c_2. It is computed as:

$$cp=\frac{|\pi_{c1}(table\_instance(\{c_1\cup c_2\}))|}{|\,table\_instance(\{c_1\})\,|} \qquad (4.2)$$

Where π is the relational projection operation with duplication elimination.

4.2.2 Steps of Co-location Mining Algorithm

Different types of co-location mining algorithms have been proposed in the past several years, for instance, Huang et al. (2004; 2005; 2006), Xiao et al. (2008), Xiong et al. (2004), Yoo et al. (2005, 2006), Verhein et al (2007). Sheng et al. (2008), Celik et al. (2006a; 2006b; 2007a; 2007b), Qian et al. (2005; 2009). All of these proposed algorithms for mining co-location rules iteratively perform five basic tasks, namely (1) initialization, (2) determination of candidate co-locations, (3) determination of table instances of candidate co-locations, (4) pruning, and (5) generation of co-location rules. These tasks are carried out inside a loop iterating over the size of the co-locations.

A. Initialization

The task of initialization is to assign starting values to various data-structures. Obviously, the value of the participation index is 1 for all co-locations of size 1, i.e., there is no need for either the computation of a prevalence measure or prevalence-based filtering, since all co-locations are prevalent.

B. Determination of Candidate Co-locations

Determination of candidate co-location is usually realized using an approximately computation with rough threshold, so that a number of features with potential co-location can be found as much as possible. Huang et al. (2004) applied *apriori_gen* proposed by Agarwal (1994) to generate size k+1candidate co-locations from size *k* prevalent co-locations. This research will use only one geometric condition, spatial neighbor to generate candidate co-location.

C. Determination of Table Instances of Candidate Co-locations

The determination of table instances of candidate co-locations can be realized through join query from k+1 candidate co-location. The query takes the k+1 candidate co-location set, C_{k+1} and k prevalent co-locations in table instances as arguments and works.

In addition, during the join computation of generating table instances, Huang et al. (2004; 2006) presented three spatial neighbor relationship constraint conditions, geometric approach $(\text{i.e.}, (p.ins\tan ce_k, q.ins\tan ce_k) \in \Re)$, a combinatorial distinct event-type constraint $(\text{i.e.}, p.ins\tan ce_1 = q.ins\tan ce_1, \cdots, p.ins\tan ce_{k-1} = q.ins\tan ce_{k-1})$, and hybrid constraint, which combine the spatial neighbor relation constrain and combinatorial distinct event-type constraint. This research will adopt the hybrid constraint, but a slight modification will be made as follow:

- ***Geometric Constraint Condition:*** The geometric constraint condition will be neighborhood relationship-based spatial joins of table instances of prevalent co-locations of size k with table instance sets of prevalent co-locations of size 1. The spatial join operations consist of filter step and refinement. For these algorithms, Huang et al. (2004; 2006) has given a detailed description.
- ***Event-type Constraint Condition:*** The distinct event-type constraint is:

Let $V = \{v_1, v_2, \cdots, v_c\}$ is a set of corresponding clusters center of feature a_1, a_2, … a_c, the distinct event-type constrain is defined as:

$$\Gamma = \sum_{i=1}^{S} \sum_{k=1}^{c} \left(\| f_i - v_k \| \right)^2 \qquad (4.3)$$

Where $\| x_i - v_k \|$ represents the Euclidean distance between f_i and v_k; Γ is a squared error clustering criterion. $v_k, \forall k = 1,2,\cdots,c$ can be calculated by:

$$v_k = \sum_{i=1}^{N} f_i / N, \quad \mathrm{N} = 6; \forall k = 1,2 \qquad (4.4)$$

So, if the Γ is greater than a given threshold, Γ_θ, i-th instance is assumed the distinct event.

D. Pruning

The purpose of pruning is to remove the non-prevalent co-locations from the candidate prevalent co-location set using the given threshold θ on the prevalence measure. Huang et al. (2004) proposed two basic pruning methods, called *prevalence-based pruning method***,** and *multi-resolution pruning***.** In this research, we will develop the spatial features pruning method. The multi-resolution pruning used criterion of the coarse participation index based on the coarse table instance to eliminate the co-location. If its coarse participation indexes fall below the threshold, the co-location will be eliminated. This research will use the autocorrelation criterion of spatial features to eliminate the co-location features. The basic idea is:

For a training data set $S = \{s_1, s_2, \cdots, s_K\}$*, if the instance* s_i*, and* s_j *are co-location, where* $s_i \in S$*,* $s_j \in S$ *and* $S_i = \{x_1, x_2, \cdots, x_N\}$*, autocorrelation of the features vectors,* x_i*, and* x_j*, where* $x_i \in S_i$ *and* $x_j \in S_j$*, is calculated by:*

$$\rho_{ij} = \frac{\sum_{i,j=1}^{N} (S_i - \overline{S}_i)(S_j - \overline{S}_j)}{\sqrt{\sum_{i,j=1}^{N} (S_i - \overline{S}_i)^2} \sqrt{\sum_{i,j=1}^{N} (S_j - \overline{S}_j)^2}} \qquad (4.5)$$

If the two feature vectors are strong cross-correlation when greater than a given threshold $T_{\rho_{ij}}$ in the training data, the co-location will be eliminated from the candidate co-location. Under this condition, a new neighbor relationship $\Re^{p}$ will have to be re-computed on the basis of the original relationship $\Re$ so that any two instances from each of the two partitions are $\Re$ neighbors. In this research, this computation is implemented under a local zone, i.e., not a global extend.

E. Generating Co-location Rules

Accompanying with the generation of co-location set, all the co-location rules with the user defined conditional probability threshold from the prevalent co-locations and their table instances can be generated (Huang et al., 2004). The conditional probability of a co-location rule $cp(c_1 \Rightarrow c_2)$ in the event centric model is the probability of c_1 reachable to a $\Re$-proximity neighborhood containing all the features in c_2.

An overview of the co-location mining algorithm is depicted in Figure 4.1.

```
Find-Co-location Instance ()      /* function

Input:
    (a) Spatial data set
    (b) Criteria, including Minimum prevalence threshold and other thresholds.

Output:
    A set of co-locations rules

Variables Setup:
    $k$ :  co-location size ¢
    $C_k$ :  set of candidate size-$k$ co-locations
    $T_k$ :  set of table instance of co-location in $C_k$
    $P_k$ :  set of prevalent size
    $R_k$ :  set of co-location rules of size
    $T\_C_k$ : set of coarse-level table instances of size-k co-locations in $C_k$

Steps:
     Step 1:  Co-location size $k$ =1;
     Step 2:  IF (fmul=TRUE) THEN
                  $T\_C_1$ =generate _table_instance($C_1$,multi_event);
     Step 3: While(not empty $P_k$ and $k < K$ ) do {
              generate candidate co_location;
              IF (fmul=TRUE) THEN
                  $C_{k+1}$ = candidate size-$k$ co-locations
              $T_{k+1}$ = table instance of co-location in $C_k$
              $P_{k+1}$ : = select prevalent colocation
              $R_{k+1}$ : = generate co-location rule
              $k = k + 1$;
             }
    Step 4:  return union
```

Figure 4.1: Overview of the co-location mining algorithm (modified from Huang et al., 2006)

4.3 Co-Location Decision Tree (CL-DT) Algorithm

The basic idea of the presented co-location decision tree (CL-DT) algorithm is depicted in Figure 4.2. The co-location mining is used to induce the co-location rules. These induced co-location rules are used to guide the decision tree generation. The co-location mining algorithm has been described in Chapter 4.2. This Chapter will focus on how a co-location decision tree is induced.

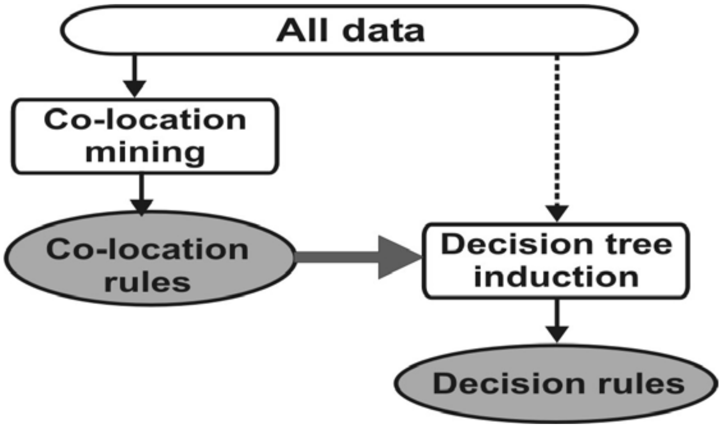


Figure 4.2: Flowchart of co-location decision tree induction

4.3.1 CL-DT Algorithm Modeling

Let each sample $s_i = \{x_1, x_2, \cdots, x_N\}^T$ in data set $S = \{s_1, s_2, \cdots, s_d\}^T$ be a vector, representing example-id, spatial feature type, and location Π, where d is the number of features, T is transpose, and the spatial location $\in$ spatial framework. The training data is augmented with a vector $C = \{c_1, c_2, \cdots, c_K\}$, where $\{c_1, c_2, \cdots, c_K\}$ represent the class to which each sample belongs. In order to assign an example to one of the classes, $C = \{c_1, c_2, \cdots, c_K\}$ ($K \geq 2$), each internal node, m_i, carries out a decision or discriminant function, denoted by $g_{m_i}(x)$ for this purpose.

The functional of $g_{m_i}(x)$ varies due to various decision tree algorithms, such as univariate decision trees, linear multivariate decision tree, and nonlinear multivariate decision tree (Altincay, 2007). This Chapter first discuss the generation of *univariate* co-location decision trees, and the linear multivariate co-location will be discussed in Chapter 4.4

As usual, the CL-DT also utilizes a divide-and-conquer strategy to partition the instance space into decision regions by generating internal or test nodes. During the generation of *the univariate* decision trees, each internal node uses only one attribute to define a *decision* or a *model.* The mathematical model can be expressed by:

$$g_{m_i}(x) = s_i + b_{m_i} \tag{4.6}$$

where b_{m_i} is a constant. The selection of the best attribute s_i, where $s_i \in S$, and corresponding b_{m_i} for the instance subset reaching at the node b_{m_i} are the main tasks in the generation of the decision function.

As shown in Figure 4.3, the proposed algorithm for generating the CL-DT consists of a binary tree structure. At beginning, the root node "accepts" all of examples, $S = \{s_1, s_2, \cdots, s_d\}^T$, the best feature is selected from input data set, and then splitting criterion is used for determining whether the root node will be split using binary decision, ***Yes*** and ***No***, with which the two intermediate nodes, noted by m_i, and m_j (i =1 and j = 2 in Figure 4.3). For each of intermediate nodes, m_i, and m_j, splitting criterion will be used to determine whether the node (e.g., m_i or m_j) should be further split. If **No**, this node is considered as a leaf node, then one of class labels is assigned to this leaf node. If **YES**, this node will be split by selecting one "best" feature. Once this "best" attribute is selected, the co-location criterion will be used to determine whether the sample with the "best" feature is co-occurred with the sample with the previously selected features (see Figure 4.3). If **YES**, this node will be "merged" into the same classification as the co-location's, and one new "best" attribute will be selected again, re-determine whether the selected "best" feature co-occurs with the last best attribute; If **NO**, the node will further be split into sub-set by repeating the above work. This selection process is repeated until a non-co-occurrence feature is found.

The above process continues recursively until all vectors are classified correctly. Finally, termination criterion is satisfied; all leaf nodes are reached, and the class labels are assigned to each of the leaf node.

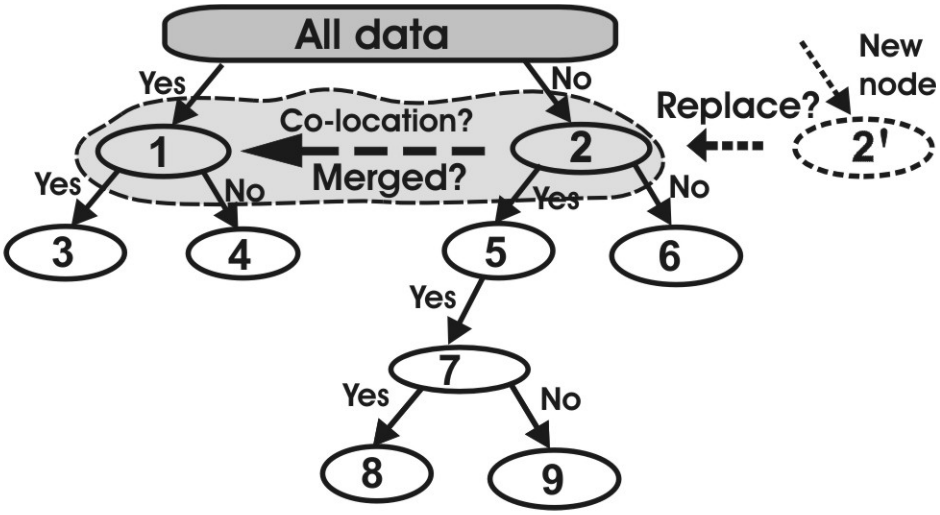


Figure 4.3: Co-location/Co-occurrence decision tree

The outline of the algorithm is depicted in Figure 4.4. The input to this algorithm consists of the training records $S = \{s_1, s_2, \cdots, s_d\}^T$ and the attribute set $s_i = \{x_1, x_2, \cdots, x_N\}^T$. The algorithm works by recursively selecting the best attribute to split the data and expanding the intermediate nodes of the tree, and checking whether or not the attributes co-occur until the stopping criterion is met.

```
Input:
    Training dataset D,
    Splitting criterion,
    Co-location threshold and criterion
    Terminal node threshold

Output:
    A LC-DT decision tree with multiple condition attributes.

Process:
    Step 1. Co-location mining
    Step 2. Co-location rules
    Step 3. Build an initial tree
    Step 4. Starting with a single node, root. The root node includes all the rules and
              attributes.
    Step 5. For each non-leaf node, e.g., $m_i$
              • Perform label assignment test to determine if there are any labels that can
                be assigned.
              • Take all the unused attributes in node $m_i$, and choose an attribute
                according to splitting criterion to further split $m_{i.}$
                  o If the selected attribute satisfy the splitting criterion, partition the node
                    into subsets.
                  o If terminal condition is satisfied, stop splitting and assign $m_i$ as a leaf
                    node.
    Step 6. For each of two non-leaf nodes in the same layer, e.g., $m_{i,}$ and $m_j$
              • Apply co-occurrence algorithm, and test if the two nodes satisfy the co-
                occurrence criterion. If yes, merging two neighbor nodes; If no, please go
                head Step 5.
    Step 7. Apply the algorithm recursively to each of the not-yet-stopped nodes, and
            update the *bottom nodes* in the tree built in step2.
    Step 8. Generate decision rule by collecting decisions driven in individual nodes.
    Step 9. The decision rules generated in Step 6 are used as initialization of co-location
            mining rule, and apply the algorithm of co-location mining rule to generate
            new associate rules.
    Step 10. Re-organize the input data set, and repeat Step 2 through Step 7, until the
            classified results by the co-location mining rule and decision tree (rules) is
            consistent.
```

Figure 4.4: Outline of algorithm of co-location/co-occurrence decision tree (CL-DT)

4.3.2 Attribute Selection

A pavement management database in fact contains many attributes, which are used to describe different pavement characteristics for various applications. This means that some of attributes in the pavement management database do not in fact contribute to pavement rehabilitation-decision, i.e., these attributes may be irrelevant to pavement decision-making of maintenance and rehabilitation. Applications of these irrelevant attributes may causes negative influences to the pavement decision support, or cause the decision tree to be over-fitted. Thus, to reduce the post-processing for obtaining an accurate and interpretable decision tree, these irrelevant attributes must be eliminated. Schetinin and Schult (2005) proposed a called *Sequential Feature Selection (SFS)* algorithms based on a greedy heuristic to eliminate the irrelevant attributes. The basic idea of this method is a bottom up search method, starting with one attribute, and then iteratively adding the new attributes until a specified stopping criterion is met. The basic steps of the sequential feature selection are described in Figure 4.5.

```
Find_Best_Attribute ()     /* function
    Step 1.  Initiation with Set $i = 1$, $F_b = F_i = F_1$          /* $W_b$ stands for the best feature
    Step 2.  Find the best attribute $F_b$
                 • Run the weighted linear tests $F_1, F_2, \dots, F_T$ with the single attribute
                 • Select the test attribute $F_k$, $k \in T$
                 • Find the best test $F_k$, $k \in T$ , **if** the test $F_k$ is better than $F_b$, **then** $F_b = F_k$
    Step 3.  **if** the stopping criterion is met, **then** stop and return $F_b$.
             **otherwise**, $i := i+1$, and **go to** Step 2.
```

Figure 4.5: The outline of steps of SFS algorithm

4.3.3 Co-Location Mining Rule

With the above co-location pattern mining operation, the co-location rules are traditionally generated with the user defined conditional probability threshold from the prevalent co-locations and their table instances. The conditional probability of a co-location has been given in Chapter 4.2.1, i.e.

$$cp = \frac{\left|\pi_{c1}(table\_instance(\{c_1 \cup c_2\}))\right|}{|\,table\_instance(\{c_1\})\,|} \tag{4.7}$$

Where π is the relational projection operation with duplication elimination.

However, this automatic method encountered problems, since conditional probability computation is time-consuming. Thus, this research manually forms the co-location rules by organizing individual decision-making.

4.3.4 Node Merging Criteria

As mentioned above, in the pavement management database, some attributes are co-occurrence in geography. For example, a co-occurrence attributes, *{car accident, traffic jam, police}* means when a car gets into an accident, the traffic jam will accompany occurrence, further police will arrive the accident site for cleaning up. So the three attributes co-occur frequently in a nearby region. If the three attributes are sequentially selected to generate the decision tree, the generated tree will be over-fitted. Thus, during the generation of the decision tree, the three nodes should be merged into one, or the other two nodes should be pruned.

One of the most major characteristics for the co-occurrence in spatial database is that the attributes occur in nearby regions in geography for an event. For this reason, this research developed the following algorithm to "prune" the nodes.

For a spatial data set *S*, let $F = \{f_1, f_2, \cdots, f_k\}^T$ be a set of *spatial attributes*. Let $I = \{i_1, i_2, \cdots, i_n\}^T$ be a set of *n* instances in *S*, where each instance is a vector instance-id, location, spatial features. The spatial attribute f_i , $f_i \subset F$ of instance i is denoted by $i.\,f$. We assume that the spatial attributes of an instance are from *F* and the location is within the spatial framework of the spatial database. Furthermore, we assume that there exists a neighbor relationship $\Re$ in *S*. In addition, let, V = {v$_1$, v$_2$, . . . , v$_c$} is a set of corresponding clusters center in the data set S, where C is the number of clusters of spatial features, i.e., *C* ⊆ *F*. To capture the concept of "nearby," the criterion of co-occurrence is defined as

$$\Pi_m = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik} \left(\| x_i - v_k \| \right)^2$$

where $\| x_i - v_k \|$ represents the Euclidean distance between x_i and v_k ; Π_m is a squared error clustering criterion; $U = \{u_{ik}\}, i = 1,2,\cdots,C; k = 1,2,\cdots,C$ c is a matrix, and satisfy the following conditions:

$$u_{ik} \in [0,1], \quad \forall_i = 1,2\cdots,N\,, \quad \forall_j = 1,2\cdots,C \qquad (4.8)$$

$$\sum_{k=1}^{C} u_{ik} = 1\,, \quad \forall_i = 1,2\cdots,N\,, \quad \forall_j = 1,2\cdots,C \qquad (4.9)$$

So, if the Π_m is less than a given threshold, the two nodes are considered as co-occurrence, and thus should be merged.

4.3.5 Decision Rule Induction from CL-DT

After the co-location decision tree is generated, decision rules will be created by translating a decision tree into semantic expressions. Since a decision tree essentially partitions a data space into distinct disjoint regions via axis parallel surfaces created by its top-down sequence of decisions, decision rules will collect the individual decisions in each node through either top-down or down-up search.

Decision trees present a clear, logical model that can be understood easily by people who are not mathematically inclined.

4.4 Linear Multivariate CL-DT Algorithm

The above discussion is for *univariate* decision tree. In fact, the CL-DT algorithm can easily be extended to *linear multivariate* and/or multi-class trees. For a linear multivariate tree, the decision is based on weighted linear combination of the features can be expressed by (Altincay, 2007)

$$g_m(x) = \sum_{i=1}^{d} w_{mi} x_i + b_m \qquad (4.10)$$

Similar to the univariate decision tree, the linear function at each node generates linear decision hyperplanes in the input space and separates the input space into two or multiple regions. For example, if a data set is partitioned into size-C classes, a maximum of C sub-nodes can be split, and up to C(C-1)/2 linear *multivariate* functions are constructed in each node. Correspondingly, C(C-1)/2 linear hyperplanes are constructed, thus separating each class from one another. It is also noted that an arbitrary hyperplane generated by a linear multivariate node is more powerful compared to the univariate case producing a hyperplane orthogonal to a particular axis (Altincay, 2007). This process continues recursively until all vectors are classified correctly, and a leaf node is reached.

4.5 Example Analysis

This Chapter explains the details of proposed algorithm, and makes a comparison analysis between the proposed method and C4.5 algorithm. Table 4.1 shows a data set of an extend example on the basis of the example adopted in Kervahut and Potvin (1996), where each instance is a member of class c_1, c_2 or c_3, and is described with four discrete attributes, namely a_1 with values f_{11}, f_{12}, f_{13}; a_2 with values f_{21} , f_{22}, f_{23}; a_3 with values f_{31} , $f_{32}, f_{33}, f_{35}, f_{36}$; and a_4 with values f_{41} , f_{42}, f_{43} , f_{45} , f_{46} *(see Table 4.1).*

Table 4.1 Data set of examples for generating a decision tree and co-location decision tree

Example	**Non-spatial attributes**		**Spatial attributes**	**Class results**	
	a_1	a_2	a_3	C4.5 algorithm	Our algorithm
s_1	f_{11}	f_{21}	f_{31}	c_1	c_1
s_2	f_{11}	f_{22}	f_{32}	c_2	c_4
s_3	f_{12}	f_{22}	f_{33}	c_2	c_2
s_4	f_{12}	f_{23}	f_{32}	c_1	c_4
s_5	f_{13}	f_{21}	f_{35}	c_3	c_3
s_6	f_{13}	f_{22}	f_{36}	c_3	c_3

In this example, we have

$$S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$$

$$A = \{a_1, a_2, a_3\}$$

$$c = \{c_1, c_2, c_3\}$$

4.5.1 Decision Tree and Decision Rules Induction using C4.5 Algorithm

The C4.5 algorithm builds decision trees from a data set of training data in the same way as ID3 (Agarwal and Srikant, 1994). At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of instances into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then repeats recursively the work on the smaller sub-lists. For the given example, the details are as follows:

Step 1: Starting with the root as the current node, where the entire set of instance belongs to.

Step 2: Select one attribute, evaluate the entropy of each subset of examples produced by splitting the set of examples at the current node along all possible attribute values. Then, combine these entropy values into a global entropy value. For example, if we evaluate the entropy of attribute a_i, the set of examples S is partitioned into subsets $S_{i,j}$, Each subset $\mathrm{S}_{\mathrm{i,j}}$ contains the instances in S that share the same value $f_{i,j}$ for feature f_i. Then, the entropy values of the subsets $S_{i,j}$ are combined to provide a single global value associated with attribute f_i, namely:

$$Gain(S,f_i) = E(S) - E(S,f_i) \qquad \forall i = 1,2,3,4 \qquad (4.11)$$

Where:

$$E(S,f_i) = -\sum_{f_{ij}\forall a_i}\left(\frac{|S_{ij}|}{|S|}\right)\times E(S_{ij}),\ \text{and}$$

$$E(S) = \sum_{c_k\in C}\left(P_{S|c_k}\log_2(P_{S|c_k})\right)$$

Where:

S = the set of examples at the current node,

|S| = the cardinality of set S,

C = the set of classes, and

$P_{S|c_k}$ = *the* proportion of examples in set S belonging to class c_k.

So, we have

$$\begin{aligned}E(S) &= E(S_{11}) + E(S_{12}) + E(S_{13})\\ &= -0.5\log_2 0.5 - 0.5\log_2 0.5 - 0.0\log_2 0.0 - 0.5\log_2 0.5 - 0.5\log_2 0.5 - 0.0\log_2 0.0\\ &\quad - 0.0\log_2 0.0 - 0.0\log_2 0.0 - 1.0\log_2 1.0\\ &= 0.6934\end{aligned}$$

$$
\begin{aligned}
E(a_1,S) &= \frac{2}{6}E(S_{11})+\frac{2}{6}E(S_{12})+\frac{2}{6}E(S_{13}) \\
&= \frac{2}{6}(-0.5\log_2 0.5-0.5\log_2 0.5-0.0\log_2 0.0)+\frac{2}{6}(-0.5\log_2 0.5-0.5\log_2 0.5 \\
&\quad -0.0\log_2 0.0)+\frac{2}{6}(-0.0\log_2 0.0-0.0\log_2 0.0-1.0\log_2 1.0) \\
&= 0.4621
\end{aligned}
$$

So, $Gain(S,x_i)=0.6934-0.4621=0.2313$

Similarly, we can calculate the information gain for a2,

$$
\begin{aligned}
E(S,a_2) &= \frac{2}{6}E(S_{21})+\frac{3}{6}E(S_{22})+\frac{1}{6}E(S_{23}) \\
&= \frac{2}{6}(-0.5\log_2 0.5-0.0\log_2 0.0-1.0\log_2 1.0)+\frac{3}{6}(-0.0\log_2 0.0-0.6667\log_2 0.6667 \\
&\quad -0.0\log_2 0.0)+\frac{1}{6}(-0.5\log_2 0.5-0.333\log_2 0.333-0.0\log_2 0.0) \\
&= 0.5698
\end{aligned}
$$

So, $Gain(S,x_i)=0.98306-0.3698=0.6133$

$$
\begin{aligned}
E(S,a_3) &= \frac{2}{6}E(S_{21})+\frac{3}{6}E(S_{22})+\frac{1}{6}E(S_{23}) \\
&= 0.5698 \\
E(S,a_4) &= \frac{2}{6}E(S_{21})+\frac{3}{6}E(S_{22})+\frac{1}{6}E(S_{23}) \\
&= 0.5698
\end{aligned}
$$

Based on the above computation of entropy, attribute a_1 is selected and the children of the root are created accordingly.

Step 3: Recursively apply this procedure to the children of the current node. The procedure stops at a given node, when the node is homogeneous, or when all attributes have been used along the path to this node. As shown in Figure 4.6, one child is homogeneous at $a_1 = f_{13}$, and no more processing is needed. The two other children are not homogeneous, and the procedure is recursively applied to each one of them, using the remaining attribute a_2.

Step 4: The stopping criterion is applied to check whether the procedure should be stopped. For this example, the final full decision tree can be created, and illustrated in Figure 4.6.

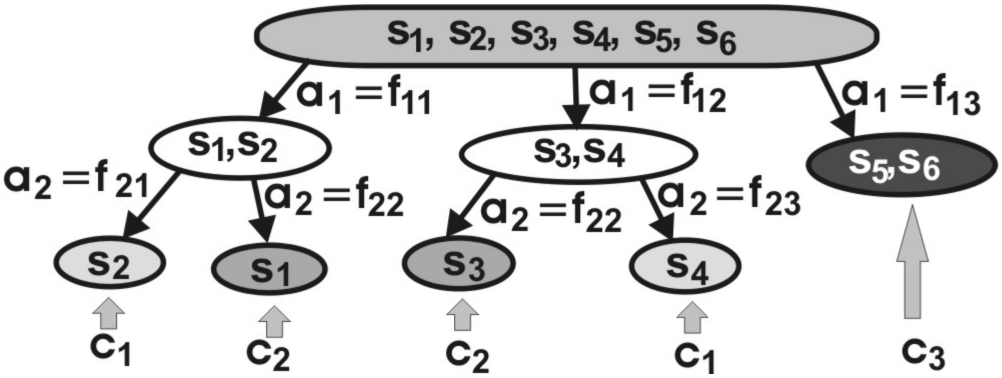


Figure 4.6 Decision tree induced by C4.5 algorithm

Step 5: With the generated decision tree above, this decision tree encodes the following decision rules (see Figure 4.7):

IF ($f_1 = a_l$) **THEN** c_3
IF (($f_1 = a_{12}$ and $f2 = a_{22}$) **OR** ($f_1 = a_{13}$ and $f_2 = a_{22}$)) **THEN** c_2
IF (($f_1 = a_{12}$ and $f2 = a_{23}$) **OR** ($f_1 = a_{13}$ and $f_2 = a_{21}$)) **THEN** c_1

Figure 4.7 decision rules induced by C4.5 algorithm

4.5.2 Our Algorithm

Here, detailed steps for our algorithm would be presented. The proposed algorithm majorly includes two major steps, co-location mining rule induction and decision tree induction. The co-location mining rule induction majorly considers the spatial data and their characteristics, and decision tree induction majorly considers the non-spatial data. Integration of two data sets using two data mining technologies is for being complimentary to the individual technology's shortcoming. The steps of our algorithms are:

4.5.2.1 Co-Location Mining Rule Induction

We first generate a co-location rule to discover which instances are "nearby", i.e., having neighborhood relationship. To this end, we follow up the steps described in Chapter 4.2.2 as follows.

Step 1: Initialization: The purpose of initialization is to set up each variable and assign the memory size for each participation variable.

Step 2: Determination of Candidate Co-locations: The candidate instances with co-location relationship will be determined using the spatial neighborhood criterion with a given threshold, D_θ. In this particular example, the spatial neighborhoods for six instances is computed by:

$$Dist_{i,j} = \sqrt{(f_{3i} - f_{3j})^2} \quad i, j \in 6 \qquad (4.12)$$

With the given data set, the spatial distances of any one pair in this data set can form the following matrix:

$$Dist = \left\{ \begin{matrix} 0 & d_{12} & d_{13} & d_{14} & d_{15} \\ & 0 & d_{23} & d_{24} & d_{25} \\ & & 0 & d_{34} & d_{35} \\ & & & 0 & d_{45} \\ & & & & 0 \end{matrix} \right\}$$

With the given values of instances S_2 and S_4, the matrix can be rewritten as follows:

$$Dist = \left\{ \begin{matrix} 0 & d_{12} & d_{13} & d_{14} & d_{15} \\ & 0 & d_{23} & 0.0 & d_{25} \\ & & 0 & d_{34} & d_{35} \\ & & & 0 & d_{45} \\ & & & & 0 \end{matrix} \right\}$$

With the above computation, instance S_2 and S_4 probably are co-location, since their spatial distance is equal to zero. Thus, S_2 and S_4 are listed as candidate co-locations.

Step 3: Determination of Table Instances of Candidate Co-locations: Based on the above generated potential co-location instances, the determination of table instances of candidate co-locations will be implemented using a combination approach, i.e., spatial neighbor relationship constraint conditions (geometric approach), and distinct event-type constrain. The spatial geometric constrain is expressed below:

$$d_{i,j} \leq D_\theta \quad d_{i,j} \subseteq Dist, \forall_i = 1,2,\cdots,6 \qquad (4.13)$$

where D_θ is given threshold for spatial distance.

With the given example, the distinct event-type constraint is:

$$\Gamma = \sum_{i=1}^{6}\sum_{k=1}^{2}\left(\|f_i - v_k\|\right)^2 \qquad (4.14)$$

where $\|x_i - v_k\|$ represents the Euclidean distance between f_i and v_k; $V = \{v_1, v_2\}$ is a set of corresponding clusters center of feature a_1 and a_2; Γ is a squared error clustering criterion. $v_k, \forall k = 1,2$ can be calculated by:

$$v_k = \sum_{i=1}^{N} f_i / N, \quad \mathrm{N} = 6; \forall k = 1,2 \qquad (4.15)$$

So, if the Γ is greater than a given threshold, Γ_θ, i-th instance is assumed the distinct event.

Step 4: Pruning: As mentioned, this research used cross-correlation of the features vectors, f_i, and f_j, to prune the candidate co-location. Since the features in this example have no correlation, thus the pruning is unnecessary.

Step 5: Generating Co-location Rules: Based on the above co-location mining approach, the co-location rules from the prevalent co-locations and their table instances can be generated. They are depicted in Figure 4.8.

IF $(d_{i,j} \le D_{\theta})$ **THEN** s_i and s_j potential co-location
IF $(d_{i,j} \le D_{\theta})$ and $\Gamma \le T_{\theta})$ **THEN** s_i and s_j co-location
IF (s_i and s_j are co-location) **THEN** c_2
OTHERWISE, c_1

Figure 4.8: Co-location mining rule

4.5.2.2 Co-Location Decision Tree Induction

With the above co-location mining rule, decision tree induction will be carried out on the basis of the induced co-location mining. Thus, during the generation of decision tree at this time, the co-location mining rule will constrain the process of decision tree induction. The steps are as follows:

Step 1: Starting with the root as the current node, where the entire set of instance belongs to.

Step 2: With the similar computation of the entropy of each subset of instances produced by splitting the set of instances at the root node, attribute a_1 is selected.

Step 3: With the selected attribute, a_1, split the instances along the path to this node. As noted, one child is homogeneous at $a_1 = f_{13}$, which implies that no further processing is needed. The two other children are not homogeneous, and the procedure is recursively applied to each one of them, using the remaining attribute a_2.

Step 4: During the recursive procedures to attribute a_2, the process will automatically recall the co-location mining rule, i.e., instances, s_2 and s_4, are co-located, i.e., co-occurred. Thus, the s_2 and s_4 must be the same class.

Step 5: The stopping criterion is applied to check whether the procedure should be stopped. For this example, the final full decision tree can be created, and illustrated in Figure 4.9.

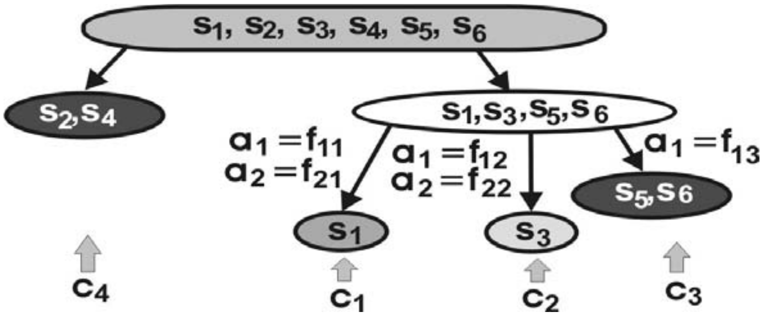


Figure 4.9 Decision tree induced by our algorithm

Step 6: With the generated decision tree above, this decision tree encodes the following decision rules (Figure 4.10):

IF ((s_2, s_4)=co-location) **THEN** c_4
IF ($a_1 = f_{13}$) **THEN** c_3
IF ($a_1 = f_{11}$ and $a_2 = f_{21}$) **THEN** c_1
IF (($a_1 = f_{12}$ and $a_2 = f_{22}$) **THEN** c_2

Figure 4.10 Decision rules induced by our algorithm

4.6 Discussion and Analysis of CL-DT

As observed from above the two examples, the C4.5 algorithm is very sensitive to the entropy formula. If we selected attribute f_2 before f_1, a different tree may be created in the example. Therefore, it can be imagined that many different decision trees may be generated when modifying the entropy formula (Tan et al., 2006). On the other hand, one major weakness of C4.5 algorithm is that a node is created for each value of a given attribute. As mentioned before, a few attributes are co-occurrence of one another, i.e., only a single attribute can get a good global evaluation in some cases, even if its entropy is good only for a few values among all its possible values (Kervahut and Potvin; 1996), where the entropy of an attribute is computed as a linear weighted sum over all values.

The CL-DT uses a co-location mining technology to first classify the co-location attributes. This is in fact equivalently to pruning the nodes whose attributes co-occur with the previous attributes. Consequently, this proposed CL-DT overcomes the weakness of C4.5 algorithm, which creates a node for each value of a given attribute. Obviously, the proposed CL-DT has capability of handling rare event, which may arise naturally in the original data set because of the lower probability of occurrence of certain classes, or the shortage of data for certain classes. Obviously, the CL-DT inherits all the advantages from regular decision trees, such as recursive divide-and-conquer approach, and efficient tree structure for rule extraction. Moreover, the proposed CL-DT allows it to solve classification problems with co-location, and co-occurrence classes, making it more robust in real-world situations.

The quality of a decision tree is based on both its accuracy and complexity. The accuracy is assessed by testing the induced decision tree and/or decision rules to new data set, and then comparing the predicted classes with the real classes. The complexity is related to the shape and size of the tree. Obviously, the proposed CL-DT has capable of creating a simple and high-accurate decision tree because this algorithm has used co-occurrence mining rule as initialization to induce the decision tree and decision rule. However, most classification algorithms sought for the models that attained the highest accuracy, or equivalently, the lowest error rate, but complex tree and rules. For the same accuracy, simple trees are preferred over complex ones.

Traditionally, most of the decision tree induction algorithms have not been capable of producing compact solutions, i.e., free expansion during generation of decision tree, despite pruning adoption. On the other hand, since the decision tree is freely is expanded, the decision rules are freely expanded as well because the decision rules directly capture individual decision of each node. These rules essentially correspond to decision regions that overlap each other in the data space. The proposed CL-DT is capable of create a compact solutions for decision tree and decision rules.

References for Chapter 4

Agarwal, R. and R. Srikant (1994). Fast Algorithms for Mining Association Rules. In *Proc. of the 20th Int'l Conference on Very Large Data Bases*, 1994.

Altincay Hakan (2007). Decision trees using model ensemble-based nodes, *Pattern Recognition*, vol. 40, pp. 3540-3551.

Al-Naymat, Ghazi (2008). Enumeration of maximal clique for mining spatial co-location patterns: *AICCSA 08 - 6th IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 126-133.

Arunasalam, Bavani; Chawla, Sanjay; Sun, Pei; Munro, Robert (2004). Mining complex relationships in the SDSS SkyServer spatial database, *Proceedings in International Computer Software and Applications Conference*, vol. 2, 2004, pp. 142-145.

Celik, M., Shekhar, S., Rogers, J., Shine, J., Yoo, J. (2006a). Mixed-drove spatio-temporal co-occurrence pattern mining: A summary of results. In: *Proceedings of the 6th international conference on data mining*, 2006, pp. 119–128.

Celik, M., Shekhar, S., Rogers, J., Shine, J. (2006b). Sustained Emerging Spatio-Temporal Co-occurrence Pattern Mining: A Summary of Results. In: *Proceedings of the 18th IEEE international conference on tools with artificial intelligence,* pp. 106–115.

Celik, M., Shekhar, S., Rogers, J., Shine, J., Kang, J. (2007a). Mining At Most Top-K Mixed-drove Spatio-temporal Co-occurrence Patterns: A Summary of Results. In: *Proceedings of the 23rd IEEE international conference on data engineering workshop*, pp. 565–574.

Celik, M., Kang, J., Shekhar, S. (2007b). Zonal Co-location Pattern Discovery with Dynamic Parameters. In: *Proceddings of the 7th IEEE international conference on data mining*, 2007, pp. 433–438.

Eick, Christoph F.; Ding, Wei; Stepinski, Tomasz F.; Nicot, Jean-Philippe; Parmar, Rachana (2008). Finding regional co-location patterns for sets of continuous variables in spatial datasets, *Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 2008, pp. 260-269.

He, Jiangfeng; He, Qinming; Qian, Feng; Chen, Qi (2008). Incremental maintenance of discovered spatial colocation patterns, *Proceedings - IEEE International Conference on Data Mining Workshops, ICDM Workshops* 2008, pp. 399-407.

Hsiao, Han-Wen, Meng-Shu Tsai, and Shao-Chiang Wang (2006). Spatial Data Mining of Colocation Patterns for Decision Support in Agriculture, *Asian Journal of Health and Information Sciences,* vol. 1, no. 1, 2006*,* pp. 61-72.

Huang, Yan; Zhang, Pusheng; Zhang, Chengyang (2008). On the relationships between clustering and spatial co-location pattern mining, *International Journal on Artificial Intelligence Tools*, vol. 17, no. 1, February 2008, pp. 55-70.

Huang, Yan; Pei, Jian; Xiong, Hui (2006). Mining co-location patterns with rare events from spatial data sets, *GeoInformatica*, vol. 10, no. 3, September 2006, pp. 239-260.

Huang, Yan; Zhang, Liqin; Yu, Ping (2005). Can we apply projection Based frequent pattern mining paradigm to spatial Co-location Mining? *Lecture Notes in Computer Science*, vol. 3518 LNAI, 2005, pp. 719-725.

Huang, Yan; Shekhar, Shashi; Xiong, Hui (2004). Discovering colocation patterns from spatial data sets: A general approach, *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 12, December 2004, pp. 1472-1485.

Huang, Yan; Xiong, Hui; Shekhar, Shashi; Pei, Jian (2003). Mining confident co-location rules without a support threshold, *Proceedings of the ACM Symposium on Applied Computing*, 2003, pp. 497-501.

Kervahut Tanguy and Jean-Yves Potvin (1996). An interactive-graphic environment for automatic generation of decision trees, *Decision Support Systems*, vol. 18, pp. 117-134.

Qian, Feng; He, Qinming; He, Jiangfeng (2009). Mining spread patterns of spatio-temporal co-occurrences over zones, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5593 LNCS, no. PART 2, 2009, pp. 677-692.

Qian, Feng; He, Qinming; He, Jiangfeng (2005). Mining spatial co-location patterns with dynamic neighborhood constraint, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5782 LNAI, no. PART 2, 2009, pp. 238-253.

Schetinin, Vitaly and Joachim Schult (2005). A neural-network technique to learn concepts from electroencephalograms, *Theory in Biosciences*, vol. 124, no. 1, 15 August 2005, pp. 41-53.

Sheng, C., Hsu, W., Li Lee, M., Tung, A. (2008): *Discovering Spatial Interaction Patterns. In: Haritsa, J.R., Kotagiri, R., Pudi, V. (eds.) DASFAA 2008. LNCS*, vol. 4947, Springer, Heidelberg 2008, pp. 95–109.

Verhein, Florian; Al-Naymat, Ghazi (2007). Fast mining of complex spatial co-location patterns using GLIMIT, *Proceedings of IEEE International Conference on Data Mining, ICDM*, 2007, pp. 679-684.

Wan, You; Zhou, Jiaogen; Bian, Fuling. CODEM: A novel spatial co-location and de-location patterns mining algorithm, *Proceedings in 5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008*, vol. 2, pp. 576-580, 2008.

Wang, Lizhen; Bao, Yuzhen; Lu, Joan; Yip, Jim (2008). A new join-less approach for co-location pattern mining, *Proceedings of IEEE 8th International Conference on Computer and Information Technology, CIT 2008*, pp. 197-202.

Xiao, X., Xie, X., Luo, Q., Ma, W. (2008). Density based co-location pattern discovery. In Proceedings of the *16th ACM SIGSPATIAL international conference on advances in geographic information systems*.

Xiong, Hui; Shekhar, Shashi; Huang, Yan; Kumar, Vipin; Ma, Xiaobin; Yoo, Jin Soung (2004). A framework for discovering co-location patterns in data sets with extended spatial objects, *Proceedings of the Fourth SIAM International Conference on Data Mining*, 2004, pp. 78-89.

Yoo, Jin Soung, Bow, Mark, (2009). Finding N-most prevalent colocated event sets; Source: *Lecture Notes in Computer Science*, vol. 5691 LNCS, *Data Warehousing and Knowledge Discovery - 11th International Conference, DaWaK 2009, Proceedings*, 2009, pp. 415-427.

Yoo, Jin Soung; Shekhar, Shashi (2006). A joinless approach for mining spatial colocation patterns, *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, October, pp. 1323-1337.

Yoo, Jin Soung; Shekhar, Shashi; Celik, Mete (2005). A join-less approach for co-location pattern mining: A summary of results, *Proceedings of IEEE International Conference on Data Mining, ICDM*, pp. 813-816.

Yoo, J., Shekhar, S., Smith, J., Kumquat, J. (2004). A partial join approach for mining collocation patterns. In: *Proceedings of the 12th annual ACM international workshop on geographic information systems*, pp. 241–249.

Zhang, X. N. Mamoulis, D. W. Cheung, and Y. Shou (2004). Fast mining of spatial collocations. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press-New York, pp. 384 – 393.

Zhou, G., and L. Wang (2010a). GIS and Data Mining to enhance pavement rehabilitation decision-making, *Journal of Transportation Engineering,* vol. 136, no. 4, February, 2010, pp. 332-341.

Zhou, G., L. Wang (2010b). Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation, *submitted to Transportation Research Part C,* revised, November 2010.

5. INITIAL EXPERIMENTS USING SPATIAL DECISION TREE MODELING[1]

5.1 Data Sources

In 1983, the Institute for Transportation Research and Education (ITRE) of North Carolina State University began working with the Division of Highways of the North Carolina Department of Transportation (NCDOT) to develop and implement a Pavement Management System for its 60,000 miles of paved state highways. At the request of several municipalities, NCDOT has made this Pavement Management System available for North Carolina municipalities. The ITRE modified this system for municipal streets in more than 100 municipalities in North and South Carolina. The data sources for this experiment are provided by ITRE of North Carolina State University. They conducted pavement distress surveys for several counties since January 2007 to determine whether or not the activity (rehabilitation treatment) for pavement needs to be carried out. This survey assessment was performed following the guidelines provided in the Pavement Condition Rating Manual (AASHTO, 2001; 1990). The collected 1285 records to be utilized in this empirical study a network-level survey, covering several-county roads including USA highway 1, and the rural road network. The provided pavement database is a spatial-based rational database, i.e., an ArcGIS software compatible database. In this database, 89 attributes including geospatial attributes (e.g., X,Y coordinates, central line, width of lane, etc.) and pavement condition attributes (e.g., cracking, rutting, etc.), and traffic attributes (e.g., shoulder, lane number, etc.), and economic attributes (e.g., initial cost, total cost) are recorded by engineers, who were carrying these surveys by walking or driving and recording the distress information and their corresponding maintenance and repair (M&R) strategy. Then, the data is integrated into database, as shown in Figure 5.1. The first through 19th column recorded road name, type, class, owner, etc. attributes; and the 31st through 43rd column recorded the pavement condition (distress) attributes; the 44th through 50th column recorded the different types of cost; others including the proposed activities, etc. As an initial experimental study, this Chapter will

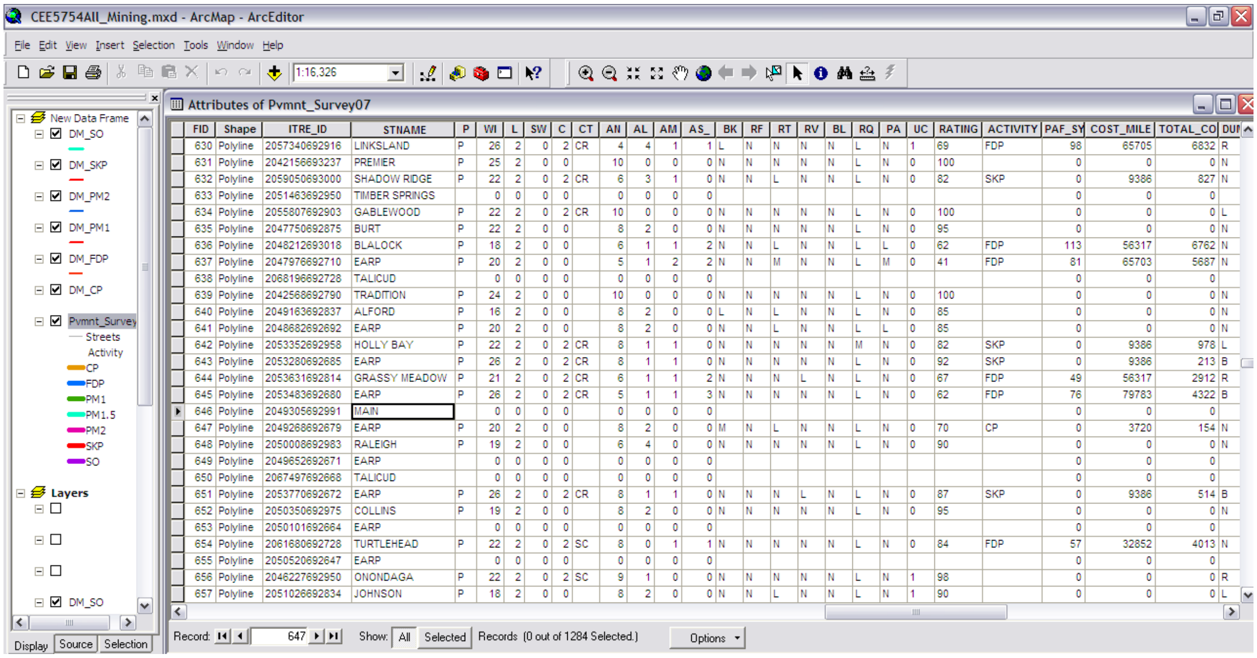
[1] This Chapter is based on the paper, Integration of GIS and Data Mining Technology to Enhance the Pavement Management Decision Making, published in Journal of Transportation Engineering, vol. 136, no. 4, pp. 332-341, April 2010.

first explore the decision tree and decision rules induction using the following nine common types of distresses, which are listed in Table 5.1 (Zhou et al., 2010a):

- Alligator Cracking,
- Block Cracking,
- Transverse Cracking,
- Bleeding ,
- Rutting,
- Utility Cut Patching,
- Patching Deterioration,
- Raveling.

Table 5.1: Nine common types of distresses for this study

Distress	**Rating**	
Alligator Cracking	Alligator None (AN)	Percentages of 1 = 10%, 2 = 20%, 3 = 30%, up to 10 = 100% indicate None, Light, Moderate, and Severe, respectively
	Alligator Light (AL)	
	Alligator Moderate (AM)	
	Alligator Severe (AS)	
Block/Transverse Cracking (BK)	This indicates the overall condition of the section as follows: • N-None • L-Light • M-Moderate • S-Severe	
Reflective Cracking (RF)	The same manner as BK	
Rutting (RT)	The same manner as BK	
Raveling (RV)	The same manner as BK	
Bleeding (BL)	The same manner as BK	
Patching (PA)	The same manner as BK	
Utility Cut Patching,	The same manner as BK	
Ride Quality (RQ)	The condition is designated as follows: • L–Average • M–Slightly Rough • S–Rough	



FID	Shape	ITRE_ID	STNAME	P	WI	L	SW	C	CT	AN	AL	AM	AS_	BK	RF	RT	RV	BL	RQ	PA	UC	RATING	ACTIVITY	PAF_SY	COST_MILE	TOTAL_CO	DU
630	Polyline	2057340692916	LINKSLAND	P	26	2	0	2	CR	4	4	1	1	L	N	N	N	N	L	N	1	69	FDP	98	65705	6832	R
631	Polyline	2042156693237	PREMIER	P	25	2	0	0		10	0	0	0	N	N	N	N	N	L	N	0	100		0	0	0	N
632	Polyline	2059050693000	SHADOW RIDGE	P	22	2	0	2	CR	6	3	1	0	N	N	L	N	N	L	N	0	82	SKP	0	9386	827	N
633	Polyline	2051463692950	TIMBER SPRINGS		0	0	0	0		0	0	0	0											0	0	0	
634	Polyline	2055807692903	GABLEWOOD	P	22	2	0	2	CR	10	0	0	0	N	N	N	N	N	L	N	0	100		0	0	0	L
635	Polyline	2047750692875	BURT	P	22	2	0	0		8	2	0	0	N	N	N	N	N	L	N	0	95		0	0	0	N
636	Polyline	2048212693018	BLALOCK	P	18	2	0	0		6	1	1	2	N	N	L	N	N	L	L	0	62	FDP	113	56317	6762	N
637	Polyline	2047976692710	EARP	P	20	2	0	0		5	1	2	2	N	N	M	N	N	L	M	0	41	FDP	81	65703	5687	N
638	Polyline	2068196692728	TALICUD		0	0	0	0		0	0	0	0											0	0	0	
639	Polyline	2042568692790	TRADITION	P	24	2	0	0		10	0	0	0	N	N	N	N	N	L	N	0	100		0	0	0	N
640	Polyline	2049163692837	ALFORD	P	16	2	0	0		8	2	0	0	L	N	L	N	N	L	N	0	85		0	0	0	N
641	Polyline	2048682692692	EARP	P	20	2	0	0		8	2	0	0	N	N	L	N	N	L	L	0	85		0	0	0	N
642	Polyline	2053352692958	HOLLY BAY	P	22	2	0	2	CR	8	1	1	0	N	N	N	N	N	M	N	0	82	SKP	0	9386	978	L
643	Polyline	2053280692685	EARP	P	26	2	0	2	CR	8	1	1	0	N	N	N	N	N	L	N	0	92	SKP	0	9386	213	B
644	Polyline	2053631692814	GRASSY MEADOW	P	21	2	0	2	CR	6	1	1	2	N	N	N	L	N	L	N	0	67	FDP	49	56317	2912	R
645	Polyline	2053483692680	EARP	P	26	2	0	2	CR	5	1	1	3	N	N	N	N	N	L	N	0	62	FDP	76	79783	4322	B
646	Polyline	2049305692991	MAIN		0	0	0	0		0	0	0	0											0	0	0	
647	Polyline	2049268692679	EARP	P	20	2	0	0		8	2	0	0	M	N	L	N	N	L	N	0	70	CP	0	3720	154	N
648	Polyline	2050008692983	RALEIGH	P	19	2	0	0		6	4	0	0	N	N	N	N	N	L	N	0	90		0	0	0	N
649	Polyline	2049652692671	EARP		0	0	0	0		0	0	0	0											0	0	0	
650	Polyline	2067497692668	TALICUD		0	0	0	0		0	0	0	0											0	0	0	
651	Polyline	2053770692672	EARP	P	26	2	0	2	CR	8	1	1	0	N	N	N	L	N	L	N	0	87	SKP	0	9386	514	B
652	Polyline	2050350692975	COLLINS	P	19	2	0	0		8	2	0	0	N	N	N	N	N	L	N	0	95		0	0	0	N
653	Polyline	2050101692664	EARP		0	0	0	0		0	0	0	0											0	0	0	
654	Polyline	2061680692728	TURTLEHEAD	P	22	2	0	2	SC	8	0	1	1	N	N	N	N	N	L	N	0	84	FDP	57	32852	4013	N
655	Polyline	2050520692647	EARP		0	0	0	0		0	0	0	0											0	0	0	
656	Polyline	2046227692950	ONONDAGA	P	22	2	0	2	SC	9	1	0	0	N	N	N	N	N	L	N	1	98		0	0	0	R
657	Polyline	2051026692834	JOHNSON	P	18	2	0	0		8	2	0	0	N	N	L	N	N	L	N	1	90		0	0	0	L

Record: 647 Show: All Selected Records (0 out of 1284 Selected.)

Figure 5.1: The spatial-based pavement database

5.2 Distress Rating

Experiments, accompanying with all the analyses and pavement condition evaluations presented in this Chapter, are based on the pavement performance measures. A common acceptable pavement performance measure is the Pavement Condition Index (PCI), which was first defined by the US Army (see Figure 5.2). In the PCI, the pavement condition is related to the factors such as structural integrity, structural capacity, roughness, skid resistance, and rate of distress. These factors are quantified in the evaluation worksheet that field inspectors use to assess and express the local pavement condition and damage severity. Mostly, inspectors use their own judgment to assess the distress condition. Usually, the PCI is quantified into 7 levels, corresponding to from Excellent (over 85) to Failed 0 (see Figure 1). Thus, PCI is an important index for maintenance and repair determination in which the overall conditions of the observed road surface are evaluated.

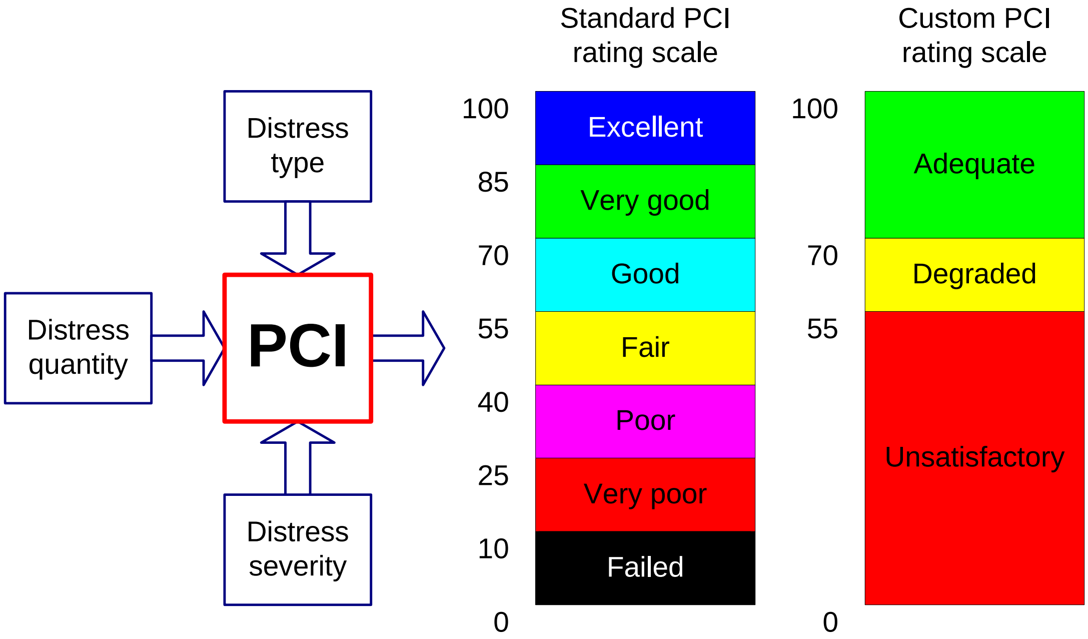


Figure 5.2: Pavement Condition Index standard and custom rating scales (Courtesy of Greene, J. and M. Shahin, 2010).

Table 5.1 presents the eight types of distresses that are evaluated for asphaltic concrete pavements. In Table 5.1, the severity of distress is rated in four categories, ranging from very slight to very severe. Extent (or density) is classified in five categories, ranging from few (less than 10 percent) to throughout (more than 80 percent). The identification and description of distress types, severity, and density are:

- The road conditions of the Alligator Cracking are rated as a percentage of the section that falls under the categories of None, Light, Moderate, and Severe. Percentages are shown as 1 = 10%, 2 = 20%, 3 = 30%, up to 10 = 100%. The appropriate percentages should be placed under None, Light, Moderate, and Severe. These percentages should always add up to 100%.
- The severity levels of distresses, Block Cracking, Transverse Cracking, Bleeding, Rutting, Utility Cut Patching, Patching Deterioration, and Raveling are rated 4 levels: None (N), Light (L), Moderate (M), and Severe (S), respectively.
- The severity levels of ride quality are classified: Average (L), Slightly Rough (M), and Rough (S).

5.3 Potential Rehabilitation Strategies

Based on the knowledge gained from experts, we have classified rehabilitation treatments for flexible pavements into three main categories according to the type of the problem to be corrected: cracking, surface defect problems, and structural problems. These problems can be treated using crack treatment, surface treatment, and nonstructural overlay (one- and two-course overlay), respectively.

In order to select an appropriate treatment for rehabilitation and maintenance to a specific road, seven potential rehabilitation and maintenance strategies have been proposed by the North Carolina Department of Transportation (NCDOT) (Table 5.2). Which treatment strategies will be carried out for a pavement segment is dependent on the comprehensive evaluation of all distresses. This used to be created by experts or a pavement engineer at North Carolina Department of Transportation. This research will experiment and test whether the decision tree and decision rule can produce an appropriate decision for M&R strategy using data mining technology, and then compare the differences of decisions made by manual method and data mining.

Table 5.2: Potential Rehabilitation Strategies

ID	**Rehabilitation Strategies**
0	Nothing
1	Crack Pouring (CP)
2	Full-Depth Patch (FDP)
3	1" Plant Mix Resurfacing (PM1)
5	2" Plant Mix Resurfacing (PM2)
6	Skin Patch (SKP)
7	Short Overlay (SO)

5.4. Decision Tree Induction-Based Maintenance and Repair (M&R)

5.4.1 Steps of Decision Tree Induction for Pavement M&R

Steps of decision tree-based decision-making usually involve the following five basic steps: problem identification, knowledge acquisition, knowledge representation, implementation and validation, and extension. For the first step, we have been much clear about what problem should be solved in this research, i.e., reveal the rule hidden in the pavement database in order to predict potential pavement condition and plan the rehabilitation. For Phase 2, we have obtained much knowledge from engineers, such as distress rating, PCI rating, and potential rehabilitation strategies. For Phase 3, the decision tree and decision rules to be induced by using traditional C5.0 algorithm are for knowledge representation. The theory part of decision tree and decision rules have been described in Chapter 3, and the experimental results will be presented in this Chapter, i.e., Chapter 5.4.1.1 will describe Step 4 (i.e., Implementation) and Chapter 5.4.1.2 will describe Step 5 (i.e., Validation).

5.4.1.1 Experiment of Decision Tree Induction

We used the CTree for Excel tool (Saha, 2003) to create the Decision Tree. This tool is based on the C5 algorithm, which lets user build a Tree-based Classification Model. The Classification Tree can generate the decision rules. The steps include

Step 1: Load Pavement Database

We first load the pavement database into the Data worksheet. The observations should be in rows and the variables should be in columns. In this data worksheet, each column, choose appropriate Type (Omit, Class, Cont, and Cat)

- **Omit** = To drop a column from model
- **Cat =** To treat a column as categorical predictor
- **Cont** = To treat a column as continuous Predictor
- **Output** = To treat a column as Class variable

Because any non-number in the **Cont** column is treated as missing value in CTree tool, we have to quantify the N, L, M and S into 100, 75, 50, and 25, manually. In this tool, the maximum predictor variables are 50, only a class variable is allowed. Application will treat the Class variable as categorical, and each of categorical predictors (including Class variable) is limited 20. After this data is uploaded, we need make sure that Class variable and data does not have blank rows or blank columns, which are treated as missing values.

Step 2: Data Inputs

This tool requires inputting some parameters to optimize the processes of decision tree generation. These parameters include:

(1) ***Adjust for # categories of a categorical predictor***: While growing the tree, child nodes are created by splitting parent nodes. Which is a predictor to use for this split is decided by a certain criterion. Because this criterion has an inherent bias towards choosing predictors with more categories, thus, input of adjust factor will be able to adjust this bias.

(2) ***Minimum Node Size Criterion***: While growing the tree, whether to stop splitting a node and declare the node as a leaf node will be determined by some criteria that we need choose. These criteria are:

- *Minimum Node Size*: A valid minimum node size is between 0 and 100.
- *Maximum Purity*: An effective values is between 0 and 100. Higher the value of this, LARGER will be the tree. Stop splitting a node if its purity is 95% or more, (e.g. Purity is 90% means). Also, stop splitting a node if number of records in that node is 1% or less of total number of records.
- *Maximum Depth:* a valid maximum depth is greater than 1 and less than 20. Higher the value of this, LARGER will be the tree. Stop splitting a node if its depth is 6 or more (Depth of root node is 1. Any node's depth is it's parent's depth + 1,)

(3) ***Pruning Option:*** This option allows us whether or not to prune the tree when tree is growing, which can help us to study the effect of pruning.

(4) ***Training and Test data:*** In this research, we used a subset of data to build the model and the rest to study the performance of the model. Also, we required the tool to randomly select the test set at a ratio of 10%.

5.4.1.2 Initial Experimental Results

1) Decision Tree

With the above data input, a decision tree is generated, as shown in Figure 5.3. Figure 5.3 displays the corresponding generated tree information, including the misclassified data percentage, the amount of time taken, total number of nodes, number of leaf nodes, and number of levels is listed in Table 5.3. The generated tree model is listed in Table 5.4. An example for a finally generated class, CP, and their predictor attribute values is listed in Table 5.5.

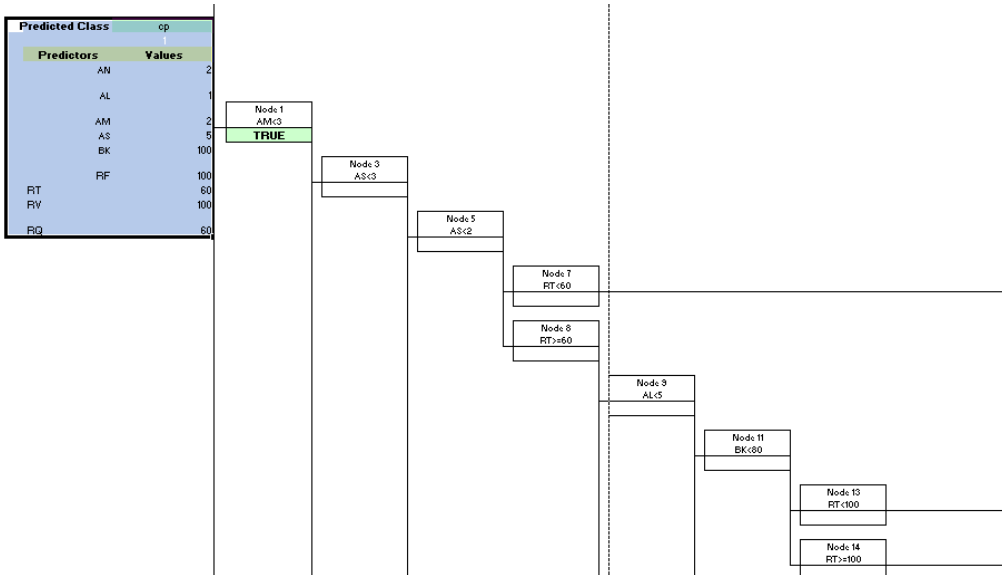


Figure 5.3: The decision tree created by C5.0 algorithm

Table 5.3: The decision tree model

Tree Information		**% Misclassified**		**Time Taken (Second)**	
Total Number of Nodes	72	Training Data	61.2%	Data Processing	1
Number of Leaf Nodes	37	Test Data	60.0%	Tree Growing	6
Number of Levels	20			Tree Pruning	1
				Tree Drawing	10
				Classification using final tree	1
				Rule Generation	35

Table 5.4: The decision tree analysis

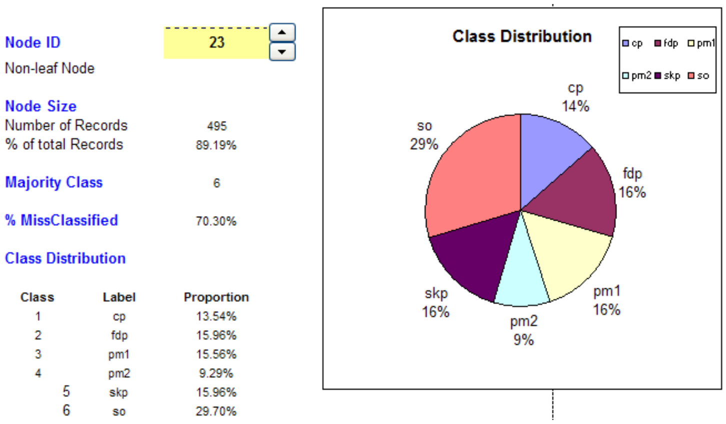
Decision Tree Model	
Number of Predictors	9
Class Variable	Activity
Number of Classes	6
Majority Class	PDK

Table 5.5: An example for a finally generated class, CP, and their predictor attribute values

Predictors	**Values**
AN	2
AL	1
AM	2
AS	5
BK	100
RF	100
RT	60
RV	100
RQ	60

2) Node View and Statistical Analysis

The statistical analysis of class distribution for any node can be overviewed from NodeView Sheet. An example, node_ID=23, is depicted in Figure 5.4, in which the class ID, node size, majority class, missed classified percentate, class proportion can be overviewed.



Node ID 23

Non-leaf Node

Node Size

Number of Records 495

% of total Records 89.19%

Majority Class 6

% MissClassified 70.30%

Class Distribution

Class	Label	Proportion
1	cp	13.54%
2	fdp	15.96%
3	pm1	15.56%
4	pm2	9.29%
5	skp	15.96%
6	so	29.70%

Figure 5.4: The detailed tree information at node 23

3) Rule Generation

After the tree is grown, the tree is further processed to generate decision rules. The decision rules are directly induced in this research by translating a decision tree in a top-down general-to-specific style. In other words, the decision rules are constructed by forming a conjunct of every test that occurs on a path between the root node and a leaf node of a tree. Thus, the decision rules are first induced by ordering all the classifications, and then using a fixed sequence, from the smallest to the largest class, to combine them together. When this rule is applied to new data set, the new data example is required in exactly the same sequence as they were generated in the training data. In total, 72 rules are generated and part of the rules is depicted in Figure 5.5., and the summary of induced rules with support, confidence, and capture is listed in Table 5.6.

```
Rule 0:
Activity = so


Rule 1:
IF AM >= 3
THEN Activity = pm2


Rule 2
IF AS >= 3
THEN Activity = cp
```

```
Rule 3
IF AS >= 2
THEN Activity = cp

Rule 4
IF RT < 60
THEN Activity = pm1

Rule 5
IF BK < 80
THEN Activity = pm1

Rule 6
IF AM >= 2
THEN Activity = pm2
... ... ... ... ... ... ... ...
```

Figure 5.5: The original rules induced by C5.0 algorithm at each node

Table 5.6: Support, confidence and capture for each generated rules

Rule ID	**Classes**	**Support**	**Confidence**	**Capture**
0	NO	100.0%	86.7%	93.0%
1	CP	60.7%	100.0%	75.6%
2	SKP	60.5%	66.7%	82.5%
3	FDP	71.6%	55.6%	66.2%
4	PM2	80.2%	100.0%	73.1%
5	PM1	81.3%	71.4%	85.3%
6	SO	81.6%	66.7%	73.5%

5.4.1.3 M&R Rules Verification

The generated decision tree organizes the obtained knowledge in a logical order. Whether or not the tree can provides a useful methodology for selecting a feasible and effective rehabilitation strategy from the 7 predetermined strategies. Thus, after the prototype of rules is generated using the algorithm above, it will have to be tested and validated, and then modified or extended, if necessary. In our study, 7 rules have been created to handle operations involved in the spatial knowledge of the pavement management system. With carefully checking the rules, these rules are all completely correct. For this reason, we used AIRA for Excel v1.3.3 tool to verify this rules. This tool is an add-in for MS-Excel and allows user to extract the 'hidden information' (i.e. discover rules) right from spreadsheets from small-/mid-range database files.

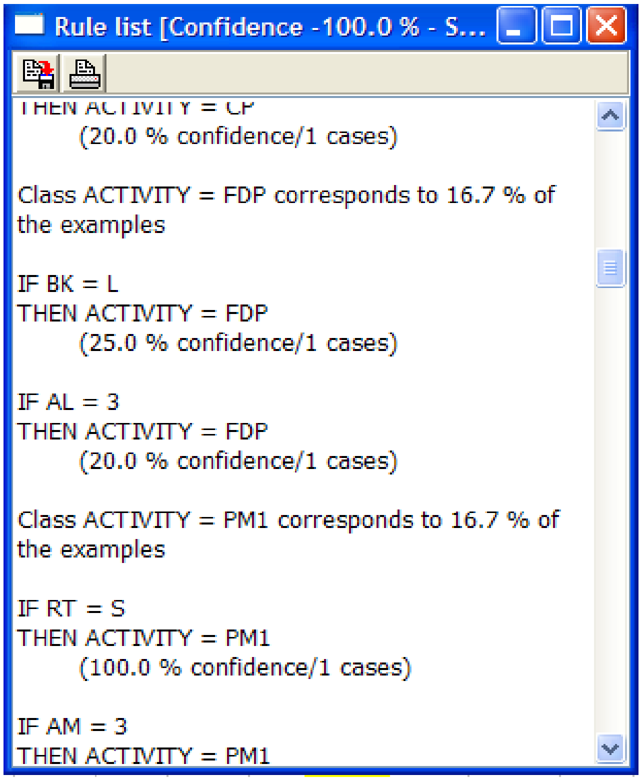
After successful operation of the AIRA tool, 41 rules are generated, part of which are listed in Figure 5.6. Obviously, so many rules will result in misclassification, thus have to be merged or deleted. To this end, the following schemes are suggested in this research:

(1) If the attribute values simultaneously match the condition of the rules induced by both C5.0 decision tree method and AIRA reasoning method, this rule would be retained;

(2) If attribute values simultaneously match the conditions of several rules, those rules with the maximum confidence will be kept;

(3) If attribute values simultaneously matching several rules with the same confidence values, those rules with the maximum coverage of learning samples will be kept; and

(4) If attribute values do not match any rule, this class of attribute is defined as the rule, nothing treatment.



```
THEN ACTIVITY = CP
      (20.0 % confidence/1 cases)

Class ACTIVITY = FDP corresponds to 16.7 % of
the examples

IF BK = L
THEN ACTIVITY = FDP
      (25.0 % confidence/1 cases)

IF AL = 3
THEN ACTIVITY = FDP
      (20.0 % confidence/1 cases)

Class ACTIVITY = PM1 corresponds to 16.7 % of
the examples

IF RT = S
THEN ACTIVITY = PM1
      (100.0 % confidence/1 cases)

IF AM = 3
THEN ACTIVITY = PM1
```

Figure 5.6: The decision rules induced by AIAR algorithm

With the schemes proposed above for decision rules reduction, 14 rules are still retained. However, only seven rehabilitation strategies in the study area were suggested by the ITRE. Thus, the following method is suggested to future reduce the number of rules:

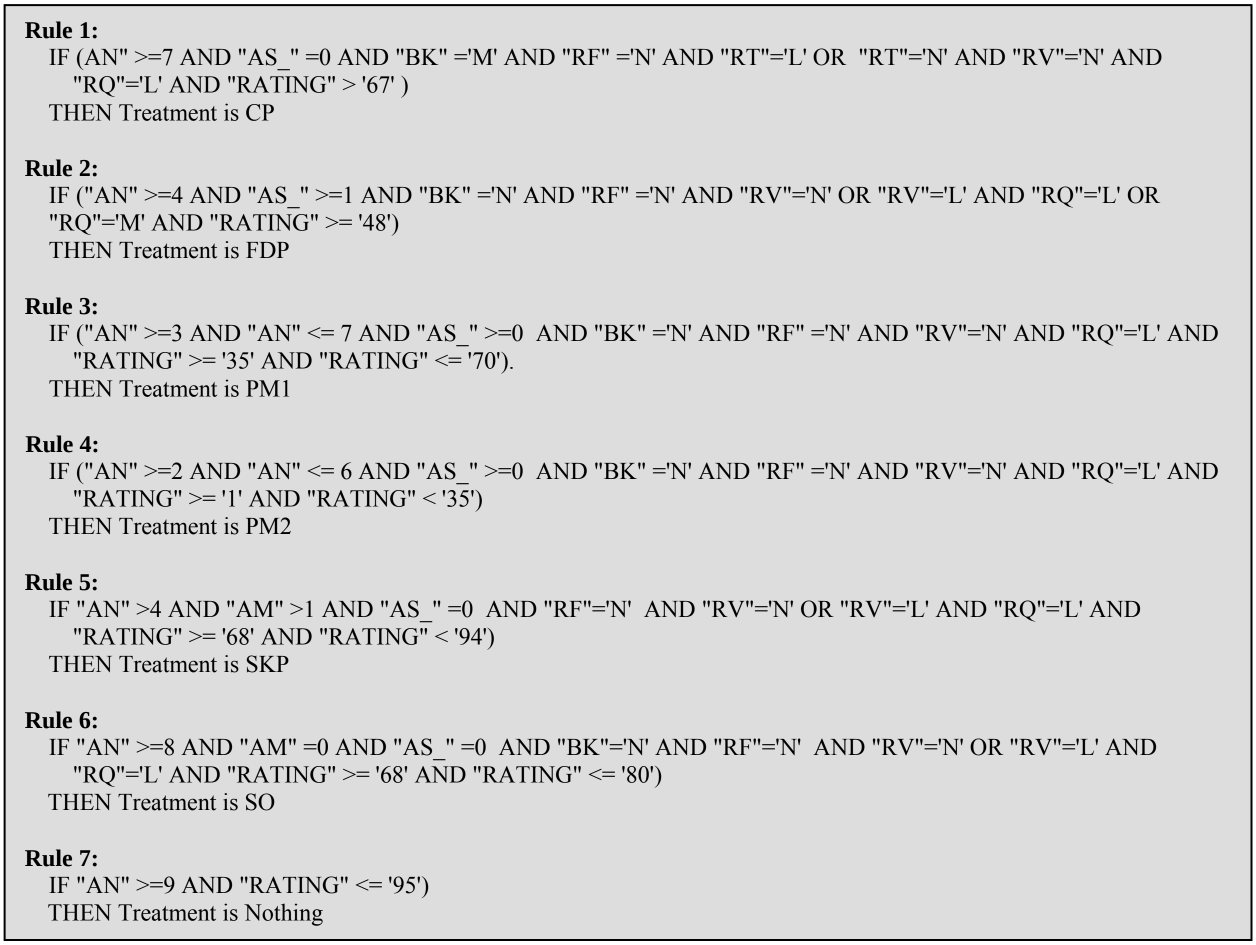
(1) Reduce the attribute data sets from Alligator Cracking family through :

$$AC = \{[AN],[AL],[AM],[AS]\}$$

(2) Reduce the attribute data through checking the PCI values. The principle is, e.g.,

- If $AC = M$ or S, reduce other attributes; and
- If $RT = S$, reduce other attributes

With the aforementioned reduction methods, the seven rules is finally refined (see Figure 5.7).



```
Rule 1:
   IF (AN" >=7 AND "AS_" =0 AND "BK" ='M' AND "RF" ='N' AND "RT"='L' OR  "RT"='N' AND "RV"='N' AND
      "RQ"='L' AND "RATING" > '67' )
   THEN Treatment is CP

Rule 2:
   IF ("AN" >=4 AND "AS_" >=1 AND "BK" ='N' AND "RF" ='N' AND "RV"='N' OR "RV"='L' AND "RQ"='L' OR
   "RQ"='M' AND "RATING" >= '48')
   THEN Treatment is FDP

Rule 3:
   IF ("AN" >=3 AND "AN" <= 7 AND "AS_" >=0  AND "BK" ='N' AND "RF" ='N' AND "RV"='N' AND "RQ"='L' AND
      "RATING" >= '35' AND "RATING" <= '70').
   THEN Treatment is PM1

Rule 4:
   IF ("AN" >=2 AND "AN" <= 6 AND "AS_" >=0  AND "BK" ='N' AND "RF" ='N' AND "RV"='N' AND "RQ"='L' AND
      "RATING" >= '1' AND "RATING" < '35')
   THEN Treatment is PM2

Rule 5:
   IF "AN" >4 AND "AM" >1 AND "AS_" =0  AND "RF"='N'  AND "RV"='N' OR "RV"='L' AND "RQ"='L' AND
      "RATING" >= '68' AND "RATING" < '94')
   THEN Treatment is SKP

Rule 6:
   IF "AN" >=8 AND "AM" =0 AND "AS_" =0  AND "BK"='N' AND "RF"='N'  AND "RV"='N' OR "RV"='L' AND
      "RQ"='L' AND "RATING" >= '68' AND "RATING" <= '80')
   THEN Treatment is SO

Rule 7:
   IF "AN" >=9 AND "RATING" <= '95')
   THEN Treatment is Nothing
```

Figure 5.7: The final rules after verification and post-processing

5.4.2 Mapping of Decision Rules-based Decision of M&R

With the rules induced above, the rehabilitation strategies can be predicted and decided for each road segment in the database using the rules. In other words, the operation using data mining and knowledge discovery (DMKD) only occurs in the database, and thus the results cannot be visualized and displayed on either map or screen. However, GIS system is capable of rapidly retrieving data from database and automatically generating customized maps to meet specific needs such as identifying maintenance locations, visualizing spatial and nonspatial data, linking data/information to its geographical location. Thus, this research employed ArcGIS software in combination with the above induced results to create the map of decision-making for maintenance and rehabilitation. The basic operation is: taking the above each rule as a logic query in ArcGIS software, and then queried results are displayed in the ArcGIS layout map. The results are listed in Figure 5.8 through Figure 5.13, corresponding to each preset rehabilitation strategies, respectively. The rehabilitations suggested by engineers at the ITRE of North Carolina State University are superimposed with the decisions made at this research. As seen from Figure 5.8 through Figure 5.13, each rehabilitation strategy derived in this research can be located with its geographical coordinates, and visualized with its spatial, non-spatial data and different colors.

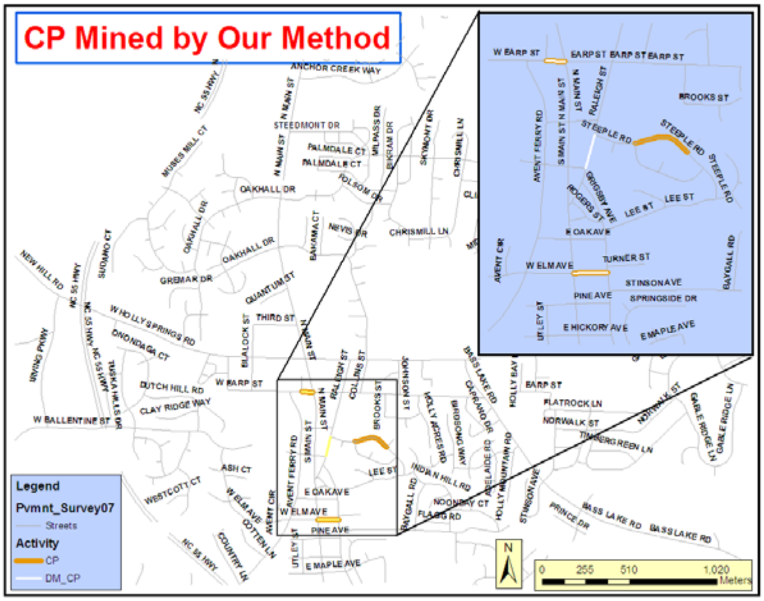


Figure 5.8: Comparison for the CP road rehabilitation made by the proposed method and by NCDOT (ITRC)

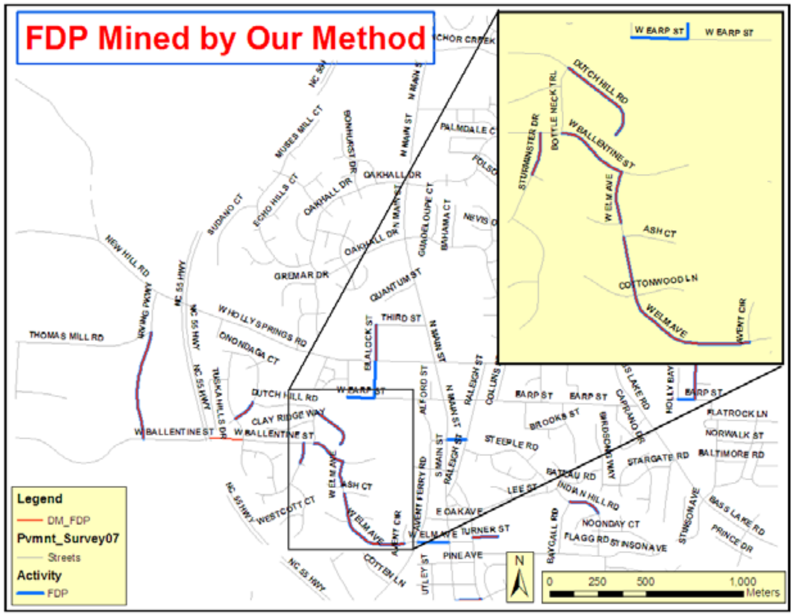


Figure 5.9: Comparison for the FDP road rehabilitation made by the proposed method and by NCDOT (ITRC)

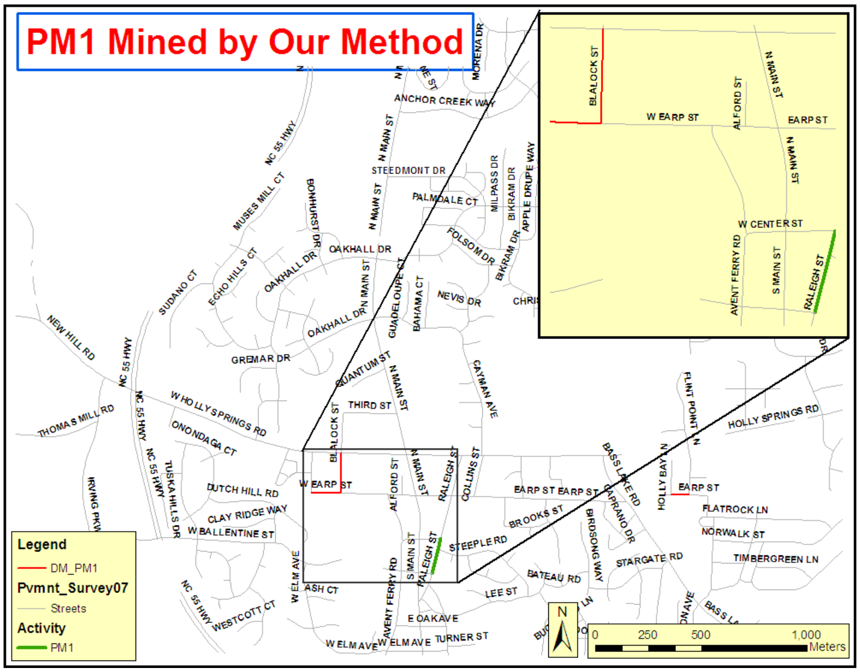


Figure 5.10: Comparison for the PM1 road rehabilitation made by the proposed method and by NCDOT (ITRC)

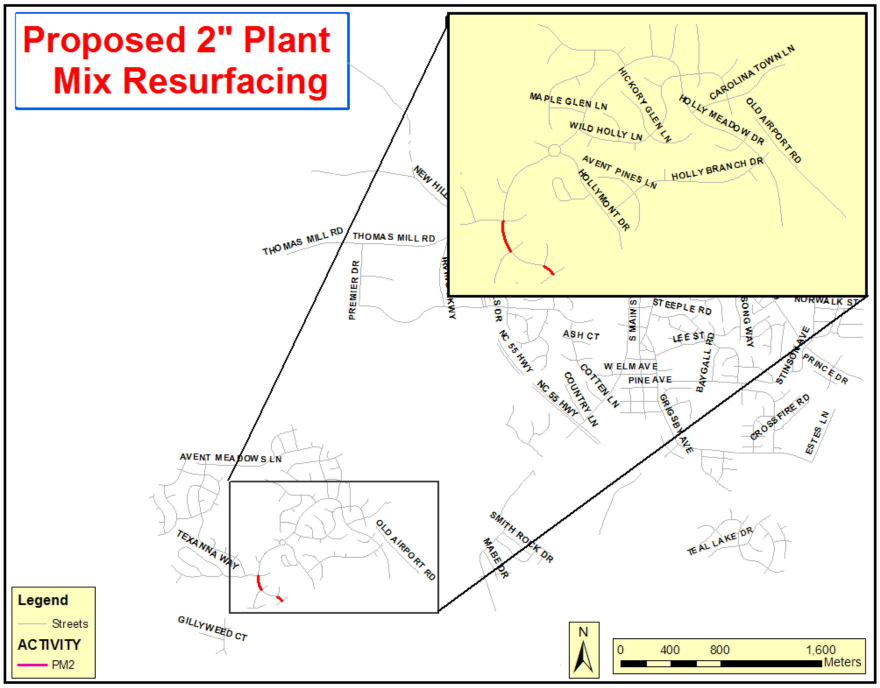


Figure 5.11: Comparison for the PM2 road rehabilitation made by the proposed method and by NCDOT (ITRC)

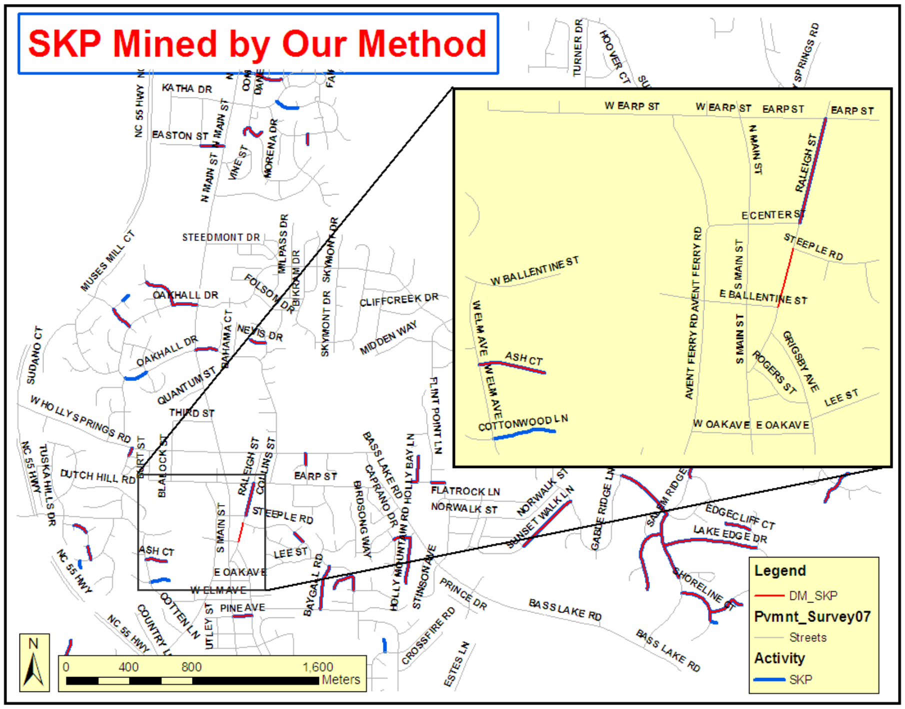


Figure 5.12: Comparison for the SKP road rehabilitation made by the proposed method and by NCDOT (ITRC)

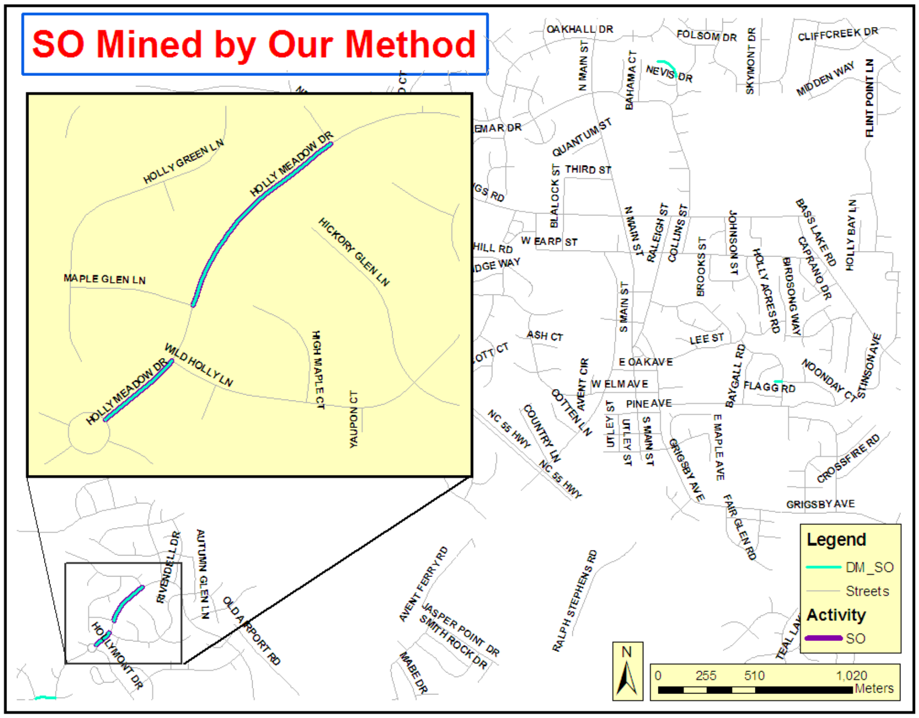


Figure 5.13: Comparison for the SO road rehabilitation made by the proposed method and by NCDOT (ITRC)

5.5. Comparison Analysis and Discussion

5.5.1 Comparison Analysis

In order to verify the correction of the proposed method for the decision-making of road maintenance and rehabilitation, the pavement segment treatments produced in this research were compared with those suggested by engineers at the ITRE at the NCDOT. All pavement treatments derived by this research and by NCDOT are displayed in Figure 5.14. A comparison analysis for both the quantity and the location of each treatment strategy derived by this research and by NCDOT in the study area is listed in Table 5.7 and Table 5.8. As seen from Table 5.7 and Table 5.8, the number of the crack pouring treatments derived by this research and by NCDOT is the same, i.e., 3, but the location of the three roads is not the same, i.e., the location of one road derived by this research is different from one derived by NCDOT. The number of the full-depth patch (FDP) treatments suggested by NCDOT is 34, but 29 by this research. The difference between 2 methods is 5. Moreover, the location of three road segments for FDP treatment is different for two methods. In the 1" plant mix resurfacing (PM1) strategy, NCDOT suggested six roads for PM1 treatment, but this research induced seven roads using decision method. Moreover, a road location is different for two methods. For the skin patch (SKP) rehabilitation strategy, 65 roads are suggested by NCDOT for treatment, but 56 roads by this research. Moreover 13 road locations are different between two methods. For the short overlay rehabilitation strategy, three roads are suggested for treatment, but 5 roads by this research, with which locations of two roads are different between two methods.

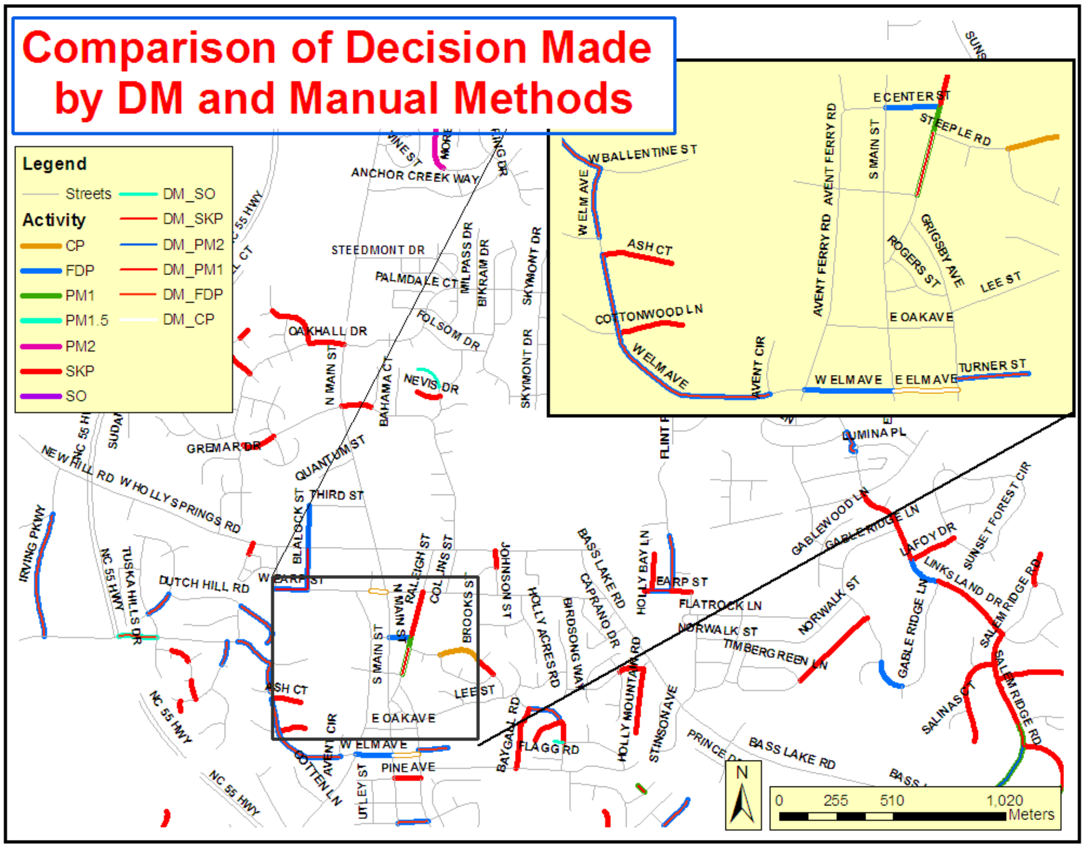


Figure 5.14: All decisions for the road rehabilitation made by ours and by NCDOT

Table 5.7: Comparison analysis of accuracy on the quantity of M&R decided by our method and by NCDOT

ID	Proposed Treatment Strategies	From	Number	Difference in number
1	Crack Pouring (CP)	NCDOT	3	0
		This research	3	
2	Full-Depth Patch (FDP)	NCDOT	34	5
		This research	29	
3	1" Plant Mix Resurfacing (PM1)	NCDOT	6	1
		This research	7	
4	2" Plant Mix Resurfacing (PM2)	NCDOT	3	1
		This research	4	

5	Skin Patch (SKP)	NCDOT	65	9
		This research	56	
6	Short Overlay (SO)	NCDOT	3	2
		This research	5	

Table 5.8: Comparison analysis of accuracy on the location of M&R decided by our method and by NCDOT

ID	**Proposed Treatment Strategies**	**From**	**Number**	**Difference in location**
1	Crack Pouring (CP)	NCDOT	3	1
		This research	3	
2	Full-Depth Patch (FDP)	NCDOT	34	3
		This research	29	
3	1" Plant Mix Resurfacing (PM1)	NCDOT	6	1
		This research	7	
4	2" Plant Mix Resurfacing (PM2)	NCDOT	3	1
		This research	4	
5	Skin Patch (SKP)	NCDOT	65	13
		This research	56	
6	Short Overlay (SO)	NCDOT	3	2
		This research	5	

5.5.2 Discussion

The decision-trees are based on the knowledge acquired from pavement management engineer for rehabilitation strategy selection. A decision-tree is in fact to organize the obtained knowledge in a logical order. Thus, the decision-trees can determine the technically feasible rehabilitation strategies for each road segment. On the other hand, different decision-trees can be built if the acknowledge changes. For example, the decision-tree in this research was based on severity levels of individual distresses. If the pavement layer thickness and material type are taken as knowledge, or work history, pavement type, and ride data are taken as knowledge for generating decision-tree, these decision-trees are different. This means that the decision rules induced by different knowledge are various.

On the other hand, our experiment has demonstrated that the decision rules induced from the decision-tree are inexact; and the decision-trees and rules generated by different tools (e.g., ACRT, CHIAD) are incompletely the same. Therefore, the post-processing for verification of rules is required.

5.6 Some Remarks

This Chapter conducts an initial research and analysis of applying the decision tree technology in pavement treatment strategies. The main purpose of the research is to utilize decision tree techniques to find some interesting knowledge hidden in the pavement database. The C 5.0 algorithm has been employed to generate decision-trees and rules. The induced rules have been used to predict which maintenance and rehabilitation strategy should be selected for each road segment. A pavement database covering four counties, which are provided by the ITRC at NCDOT, has been used to test the proposed method. The comparison of two decisions for rehabilitation treatment suggested by NCDOT and by the methodology presented in this research has been conducted. From the experimental results, it was found that the rehabilitation strategies derived by the rules, i.e., C5.0 method, are different from those suggested by NCDOT. After combining other technologies, e.g., AIRA method, and post-processing, seven rules are finally refined. Using the final rules, mapping for different types of pavement rehabilitation strategies is created using ArcGIS v. 9.3. When compared with the results from NCDOT, the quantity and location of the suggested road rehabilitations are different. The maximum error for the number of the suggested road rehabilitations is 9, and for the location is 13 out of 65 (see Table 5.7 and Table 5.8).

Through this initial exploration on the decision tree applied in decision-making of pavement treatment strategies, it has been concluded that (Zhou et al., 2010a; 2010b; 2008):

(1) The use of data mining and knowledge discovery method for road maintenance and rehabilitation can largely increase the speed of decision-making, saving time and money, and shorten the project period.

(2) The use of data mining and knowledge discovery method for pavement management can make a consistent decision for road-network treatment strategies, avoiding any human factors for decision-making of treatment.

(3) A decision tree is used to organize the obtained knowledge from experts in a logical order. Thus, decision trees can determine the technically feasible rehabilitation strategies for each road segment at a reasonable manner.

On the other hand, application of decision tree in decision-making of pavement treatment strategies also discovered many shortcomings as follows:

(1) *Post-processing*: the DMKD method is not quite as smart as people imagine, since it is based on severity levels of individual distresses. Consequently, the induced decision rules for pavement treatment rehabilitation and maintenance are not completely correct. So, post-processing for verification is quite needed.

(2) *Many leaves and nodes, and decision rules*: The current algorithms of decision tree induction produce many tree nodes and leaves, resulting in redundant individual decision rules. The organization of individual rules into a logically ordered decision rules is time-consuming, sometime is incorrect.

(3) *Attribute selection*: The current algorithms of decision tree induction, such as C4.5, produce a decision tree through selecting each of attribute data. This implies that the algorithm does not consider relationship among the attribute data, such as co-location, co-occurrence, and cross-correlation.

(4) *Spatial data:* The data set of pavement database includes geospatial data in addition to the attribute data. As known, these geospatial elements basically have three characteristics: attributes, geographical location, and topological relationship. The non-spatial (attribute) data is basically the same as those used in any traditional database, e.g., condition of pavement, and history of construction and maintenance. Spatial data that link the geospatial elements to its geodetic position in a give map-based coordinate system, such as State Plane Coordinate System, to uniform all data sets in the same reference.

The topological data structure or topology relationship describes the spatial relationships between adjacent features, and uses x, y coordinates to identify the location of a particular point, line, or polygon. Using such data structures enforces planar relationships, and allows GIS specialists to discover relationships between data layers, to reduce artifacts from digitization, and to reduce the file size required for storing the topological data. Unfortunately, the two major characteristics of spatial data in current decision tree induction methods have not been considered.

References for Chapter 5

AASHTO (2001): Pavement Management Guide. AASHTO, Washington, D.C., 2001. 64 Paper, No. 02-3100 *Transportation Research Record,* vol. 1816.

AASHTO (1999): AASHTO guidelines for pavement management system. (1990). *American Association of State Highway and Transportation Officials,* Washing ton, D.C.

Greene, J. and M. Shahin, Airfield Pavement Condition Assessment**,** www.cecer.army.mil/.../AirfieldPavementConditionAssessment_r3.doc., accessed on November 2010.

Metropolitan Transportation Commission and ERES Consultants, Inc., (1986). Pavement Condition Index Distress Identification Manual For Asphalt and Surface Treatment Pavements, prepared by Metropolitan Transportation Commission and ERES Consultants, Inc., published by *Metropolitan Transportation Commission*, no. 94607, February 1986, 3rd printing July 1988, 4th printing January, 1993.

Pavement Condition Index (PCI) Guidance. Ser ESC63/103 of 16 Feb 2000.

Saha, A. (2003). CTree Software for Excel. http://www.geocities.com/adotsaha/ (25.02.2003), accessed on November 2010.

Zhou, G., and L. Wang (2010a). GIS and Data Mining to enhance pavement rehabilitation decision-making, *Journal of Transportation Engineering,* vol. 136, no. 4, February, 2010, pp. 332-341.

Zhou, G., L. Wang (2010b). Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation, *submitted to Transportation Research Part C,* Revised November 2010.

Zhou, G., and L. Wang (2008). Integrating GIS and Data Mining to Enhance the Pavement Management Decision-Making, *The 8th International Conference of Logistics and Transportation,* Chengdu, China July 31 –August 2, 2008.

6. IMPROVED EXPERIMENTS USING CO-LOCATION DECISION TREE MODELING

6.1 Experimental Design

6.1.1 Flowchart of Experimental Design

The Co-location decision tree consists of two major steps, co-location mining and decision tree induction. Non-spatial data used for decision trees is the same as those used in Chapter 5, and maintenance and rehabilitation strategies are the same as those used in Chapter 5. A flowchart is depicted in Figure 6.1, in which the data selection, including spatial data, is the first step, and co-location mining is critical.

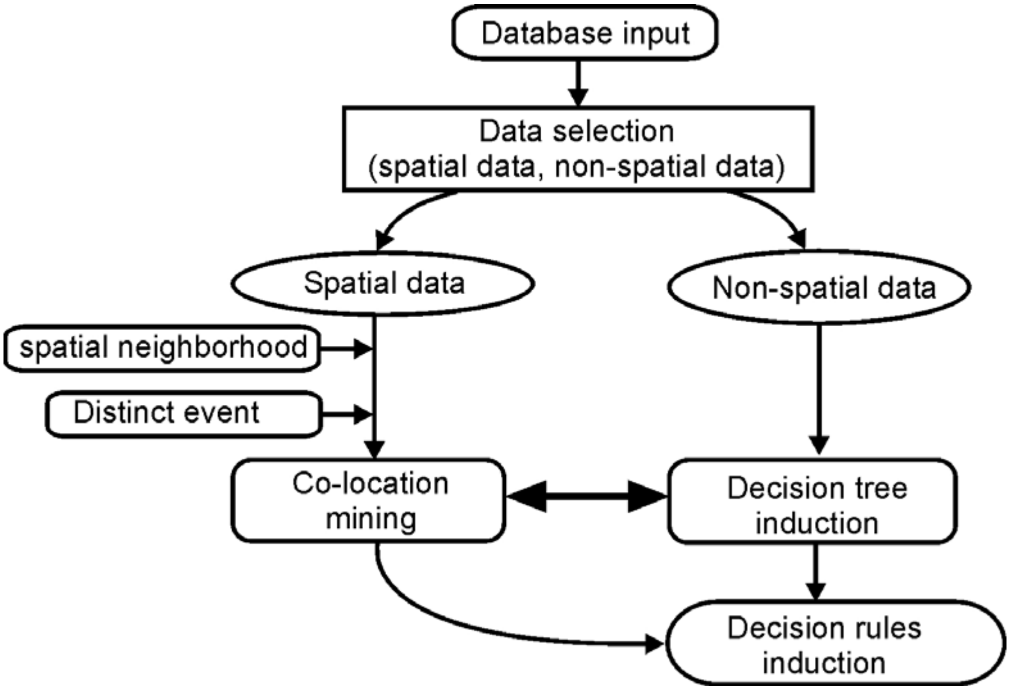


Figure 6.1: Flowchart of experimental design

6.1.2 Data Selection

As mentioned above, the provided pavement database is a geospatial rational database, i.e., an ArcGIS software compatible database. In this database, 89 attributes are collected including geospatial data (e.g., X, Y coordinates, central line, width of lane, etc.), pavement condition data (e.g., cracking, rutting, etc.), traffic data (e.g., shoulder, lane number, etc.), and economic data (e.g., initial cost, total cost). The data was recorded in different columns. The 31st through 43rd columns recorded the pavement condition; the proposed activities are recorded in 42nd column, and spatial data are recorded in different columns. Traditionally, only non-spatial attribute data was considered for this purpose. However, as mentioned early, it is incorrect for decision-making without considering spatial data. Thus, two types of data sets, *spatial data* and *non-spatial data*, should be considered simultaneously for computer to automatically make decisions for pavement maintenance and rehabilitation.

6.1.3 Non-Spatial Attribute Data Selection

In order to keep a consistent comparison with the results in Chapter 5, this research would select eight pre-defined common types of distress, and ride quality, which were proposed by experts for experimental analysis to be conducted in this Chapter. The data sets are exactly the same as the ones considered in Chapter 5 (see Table 6.1), i.e.,

- Alligator Cracking,
- Block Cracking,
- Transverse Cracking,
- Bleeding ,
- Rutting,
- Utility Cut Patching,
- Patching Deterioration,
- Raveling, and
- Ride quality (RQ).

Table 6.1: Eight common types of distresses plus ride quality for this study

#	Distress	Rating	
1	Alligator Cracking (four types of rates are given)	Alligator None (AN)	Percentages of 1 = 10%, 2 = 20%, 3 = 30%, up to 10 = 100% indicate None, Light, Moderate, and Severe, respectively
		Alligator Light (AL)	
		Alligator Moderate (AM)	
		Alligator Severe (AS)	

2	Block/Transverse Cracking (BK)	This indicates the overall condition of the section as follows: • N-None • L-Light • M-Moderate • S-Severe
3	Reflective Cracking (RF)	The same rating as BK's
4	Rutting (RT)	The same rating as BK's
5	Raveling (RV)	The same rating as BK's
6	Bleeding (BL)	The same rating as BK's
7	Patching (PA)	The same rating as BK's
8	Utility Cut Patching	The same rating as BK's
9	Ride Quality (RQ)	The condition is designated as follows: • L−Average • M−Slightly Rough • S−Rough

6.1.4 Spatial Attribute Data Selection

The spatial data in the database includes X,Y coordinates, central line, width of lane, number of travel lanes, length of street segment, first-left, to-left, first-right, to-right, etc. This research only considers the following two spatial attribute data sets. The metadata of two spatial data are explained in Table 6.2

Table 6.2: Selected spatial data for this study

Attributes	Explanation
• X coordinate • Y coordinate	Datum: NAD_1983_StatePlane_North_Carolina_FIPS_3200_Feet Coordinate system name: GCS_North_American_1983 Map Projection Name: Lambert Conformal Conic Standard Parallel: 34.333333 Standard Parallel: 36.166667 Longitude of Central Meridian: -79.000000 Latitude of Projection Origin: 33.750000 False Easting: 2000000.002617 False Northing: 0.000000
Length	GIS length of street segment (in feet)

6.2 Maintenance and Rehabilitation (M&R) Strategies

Also, in order to keep the consistent comparison with the results derived in Chapter 5, seven potential rehabilitation and maintenance strategies proposed by the North Carolina Department of Transportation (NCDOT) (see Table 6.3) are selected. These treatments include crack treatment, surface treatment, and nonstructural overlay (one- and two-course overlay), which correspond to the cracking, surface defect problems, and structural problems, respectively.

This Chapter will test how the proposed co-location decision tree method described in Chapter 4 can be used to select a M&R strategy for a pavement segment, which is traditionally dependent on the comprehensive evaluation of all distresses by experts. A comparison analysis with respect to those proposed by experts at NCDOT will be conducted.

Table 6.3: Potential Rehabilitation Strategies

#	**Rehabilitation Strategies**
0	Nothing
1	Crack Pouring (CP)
2	Full-Depth Patch (FDP)
3	1" Plant Mix Resurfacing (PM1)
4	2" Plant Mix Resurfacing (PM2)
5	Skin Patch (SKP)
6	Short Overlay (SO)

6.3 Induction of Co-Location Mining Rules

6.3.1 Determination of Candidate Co-Locations

As proposed in Chapter 4, the candidate instances with co-location relationship will be determined using the spatial neighborhood criterion with a given threshold, D_θ. In this research, the spatial neighborhoods for all instances are computed by:

$$Dist_{i,j} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2} \qquad \forall \mathrm{i,j} = 1,2,\cdots,1285 \qquad (6.1)$$

Where X amd Y is the spatial data of the pavement database. With the given database at a dimension of 1285 instances, the spatial distances of any two instances produce a matrix with the dimension of 1285 x 1285, i.e.,

$$\underset{1285\times1285}{Dist} = \left\{ \begin{matrix} 0 & d_{12} & \cdots & \cdots & d_{1\times1285} \\ & 0 & \cdots & \cdots & d_{2\times1285} \\ & & \cdots & \cdots & \cdots \\ & & & \cdots & d_{1284\times1285} \\ & & & & 0 \end{matrix} \right\} \qquad (6.2)$$

With the given database, a statistic analysis, including average and standard deviation, for the length of street segment is conducted. It is found that the length with approximately 25000 feet is appropriate as threshold. Thus, the threshold of spatial distance of two instances is selected $D_\theta = 25000$ (feet). Combining the generated spatial neighborhood matrix (Eq. 6.2) and threshold, the elements of spatial neighborhood matrix, $\underset{1285\times1285}{Dist}$, is re-calculated by:

$$d_{i,j} = \begin{cases} d_{i,j} & \text{if } \mathrm{d_{i,j}} \le 25000 \\ 0.0 & \text{if } \mathrm{d_{i,j}} > 25000 \end{cases} \qquad d_{i,j} \subseteq Dist \qquad (6.3)$$

With the above filtering, the potential of co-location instances can be determined by the spatial neighborhood matrix, which is a spare matrix.

6.3.2 Determination of Table Instances of Candidate Co-Locations

6.3.2.1 Determination of Distinct Events

In addition to the above geospatial distance constraint, another constraint condition for the determination of candidate co-location is the distinct event-type constraint. This implies that if two instances are co-located, they must be distinct event. The constraint condition of distinct event is mathematically expressed by:

$$\Gamma_i = \sum_{k=1}^{K} \left(\| f_i - v_k \| \right)^2 \qquad \mathrm{i} = 1,\cdots,1285 \qquad (6.4)$$

where $\|x_i - v_k\|$ represents the Euclidean distance between f_i and v_k; $V = \{v_1, v_2, \cdots, v_k\}$ is a set of corresponding clusters center of attributes, $\{a_1, a_2, \cdots, a_k\}$; Γ is a squared error clustering criterion; and K is number of event. $v_k, \forall k = 1, 2, \cdots K$ can be calculated by:

$$v_k = \sum_{i=1}^{N} f_i / K, \qquad \forall k = 1,2,\cdots,K \qquad (6.5)$$

So, the distinct event can be determined by:

$$\Gamma_i = \begin{cases} \text{event} & \text{if } \Gamma_i \le \Gamma_\theta \\ N/\text{A} & \text{if } \Gamma_i \le \Gamma_\theta \end{cases} \qquad (6.6)$$

where Γ_θ is threshold. If the Γ_i is greater than a given threshold, i-th instance is assumed the distinct event.

For the given database, only one attribute, ride quality (RQ) is selected for evaluating the distinct event, with which the clusters center of attributes of ride quality, v is 85, which is calculated by Eq. 6.5.

Eq. 6.4 can be rewritten by:

$$\Gamma_i = \sum_{i=1}^{K} (f_i - 85)^2 \qquad (6.7)$$

With Eq. 6.7, the values of Γ_i for 1285 instances is depicted Figure 6.2. Further, the distinct events can be determined by Eq. 6.6.

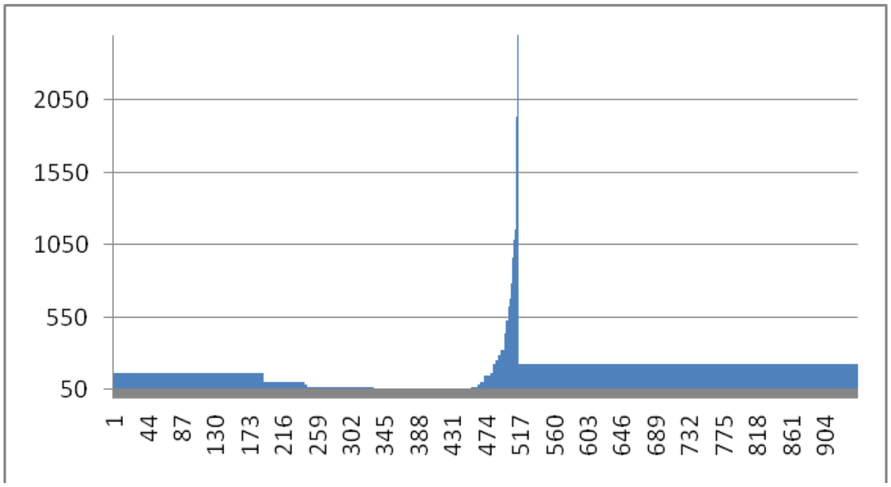
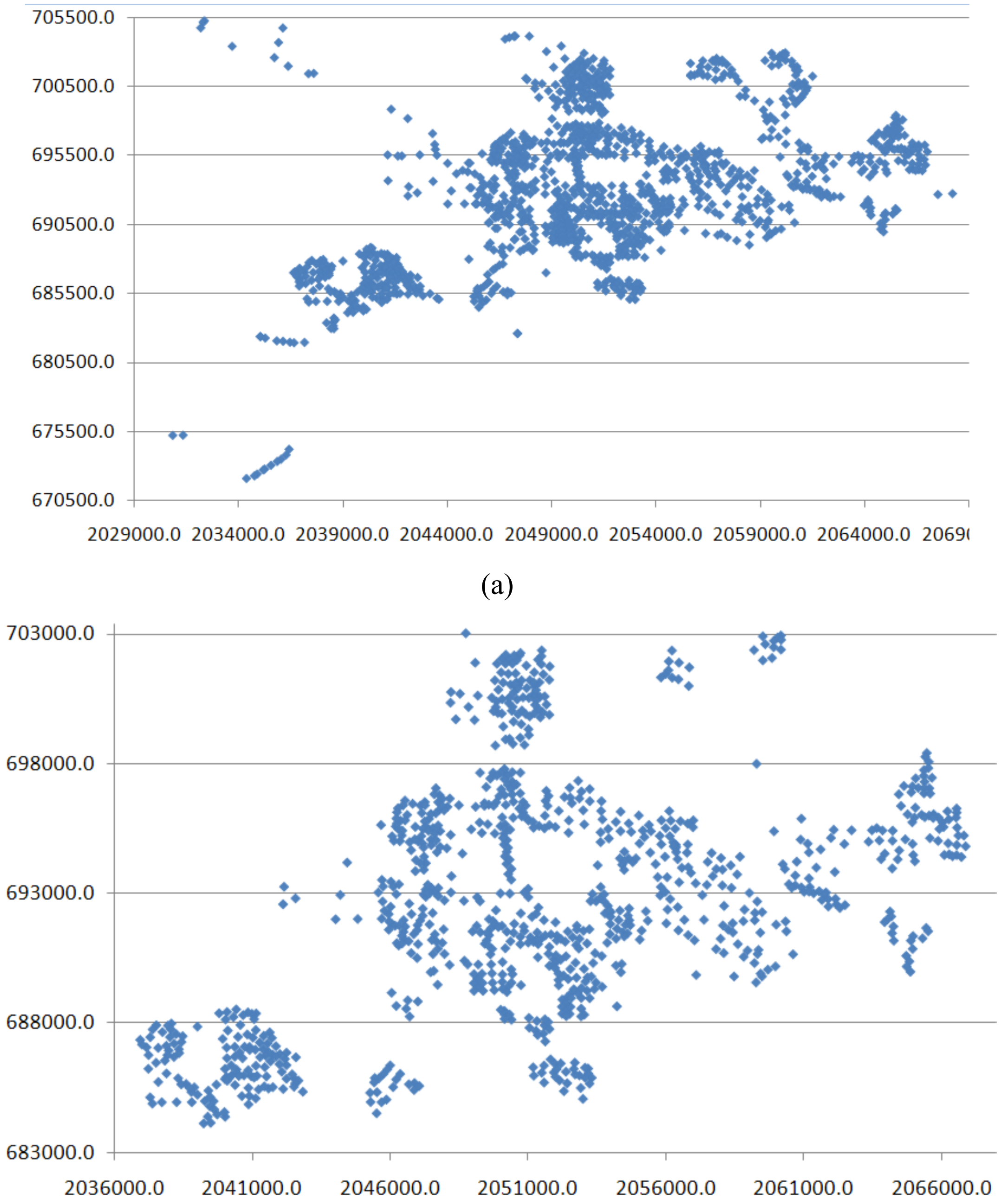


Figure 6.2: Determination of distinct events using ride quality (RQ)

The computational process of the above two constrain conditions can be illustrated by Table 6.4. For example, for a given distinct event, PM1, the geosptial distance criterion first produces 11 instances, which are co-located with PM1. With the second criterion condition of distinct event, only 7 instances are co-located with PM1 event, since the other 4 instances have no records of rating of ride quality. Figure 6.3 depicts the distributions from the original 1285 instances (Figure 6.3a) to 946 instances (Figure 6.3b) after two constraint conditions are used. Finally, a total of 946 distinct events are found.

Table 6.4: The process of co-location mining using both the geospatial distance criterion and distinct event criterion

#	X	Y	Activity	Rating
1	2049671.1	691641.5	PM1	69
2	2049518.9	691461.1		98
3	2049600.3	691368.6		90
4	2049673.2	690247.2	CP	68
5	2049643.1	697413.1		100
6	2049600.3	691368.6		90
7	2049646.1	690634.4		88
8	2049632.7	702497.6		
9	2049615.2	699440.8		
10	2049660.7	693303.0		
11	2049652.9	692671.9		



(a)

(b)

Figure 6.3: Spatial distributions of the 1285 original instances (a) and 946 instances (b) after co-location algorithm

6.3.2.2 Co-Location Mining for Individual Rehabilitation and Maintenance Strategy

As mentioned earlier, seven potential rehabilitation and maintenance strategies have been proposed by the North Carolina Department of Transportation (NCDOT). In order to find the co-location events for each R&M strategy, we take each strategy as a distinct-event, and then find the co-location using co-location mining algorithm, which has been described above, respectively. For each of R&M treatment strategy, the results of co-locating mining are as follows.

A) Crack Pouring (CP)

The ITRC at the NCDOT indicated three crack pouring (CP) treatment strategies. Two of them are chosen to illustrate the results of the proposed co-location mining method. As seen in Figure 6.4a and Figure 6.4b, 5 instances are clustered with the first CP event (Figure 6.4a), and three instances are clustered with another CP event (Figure 6.4b). Other events are not clustered due to far distances.

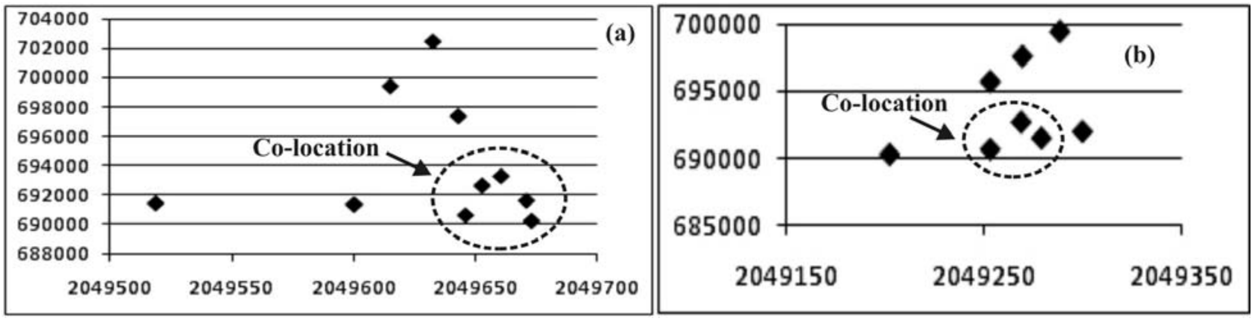


Figure 6.4: Spatial distributions of CP instances after initial determination of co-location algorithm

B) Full-Depth Patch (FDP)

The ITRC at the NCDOT indicated 34 full-depth patch (FDP) treatment strategies. Four of them are chosen to illustrate the results of the proposed co-location mining algorithm. As seen in Figure 6.5a and Figure 6.5d, no instances are clustered around the two FDP events, but there are clusters in Figure 6.5b and 6.5c for the other two FDP events.

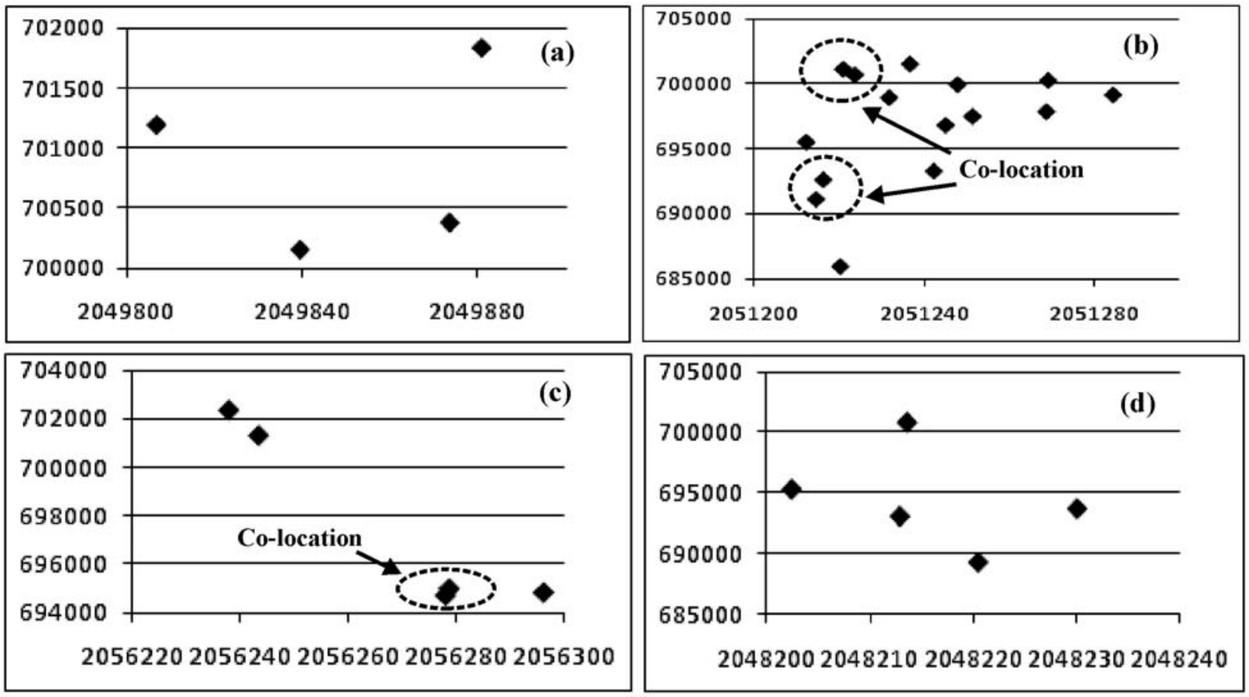


Figure 6.5: Spatial distributions of FDP instances after initial determination of co-location algorithm

C) 1" Plant Mix Resurfacing (PM1)

The ITRC at the NCDOT indicate six 1" plant mix (PM1) treatment strategies. All of them are chosen to illustrate the results of the proposed co-location mining algorithm. As seen from Figure 6.6a, Figure 6.6c, Figure 6.6d and Figure 6.6f, no instances are clustered with the four PM1 events, but there are clusters for the two PM1 events in Figure 6.6b and 6.5e.

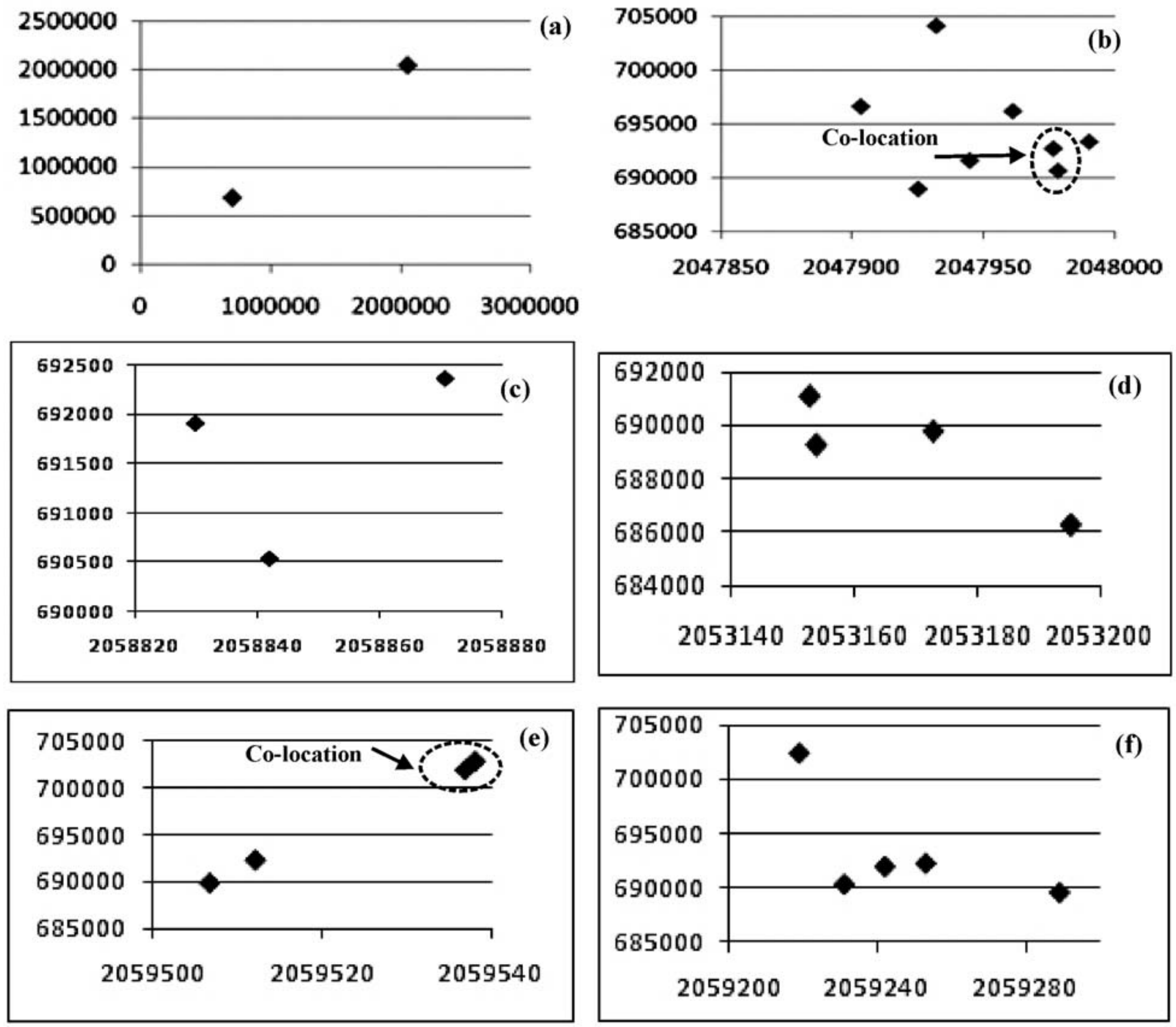


Figure 6.6: Spatial distributions of PM1 instances after initial determination of co-location algorithm

D) 2" Plant Mix Resurfacing (PM2)

The ITRC at the NCDOT indicated three 2" plant mix (PM2) treatment strategies. All of them are chosen to illustrate the results of the proposed co-location mining algorithm on the basis of event of PM2 treatment strategy. As seen in Figure 6.7a, and Figure 6.7b, no instances are clustered with the two PM2 events, but there is a cluster for another PM2 event in Figure 6.7c.

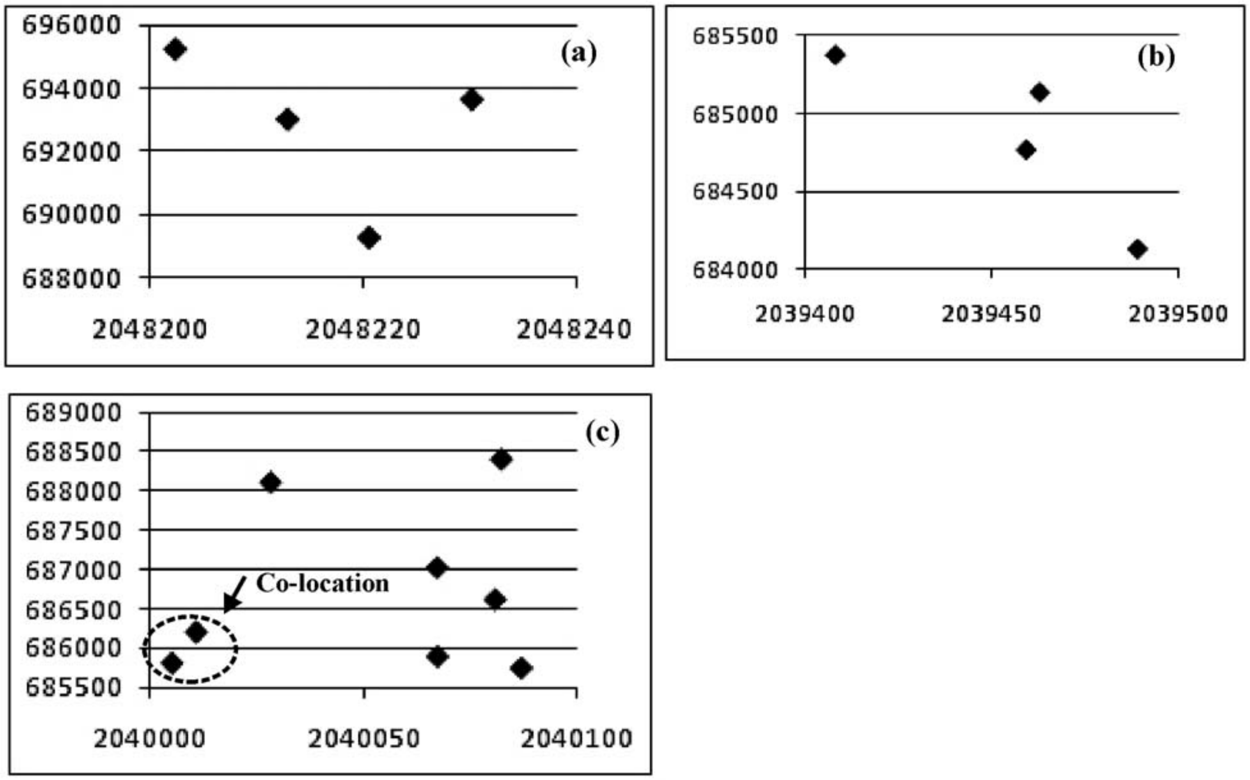


Figure 6.7: Spatial distributions of PM2 instances after initial determination of co-location algorithm

E) Skin Patch (SKP)

The ITRC at the NCDOT indicated 56 skin patch (SKP) treatment strategies. 9 representatives of them are chosen to illustrate the results of the proposed co-location mining method with each SKP treatment strategy. As seen in Figure 6.8a through Figure 6.8d, Figure 6.8f, and Figure 6.8g through Figure 6.8i, several candidate events are clustered surrounding the individual SKP treatment event, but no candidate event is clustered surrounding one SKP treatment event in Figure 6.8e.

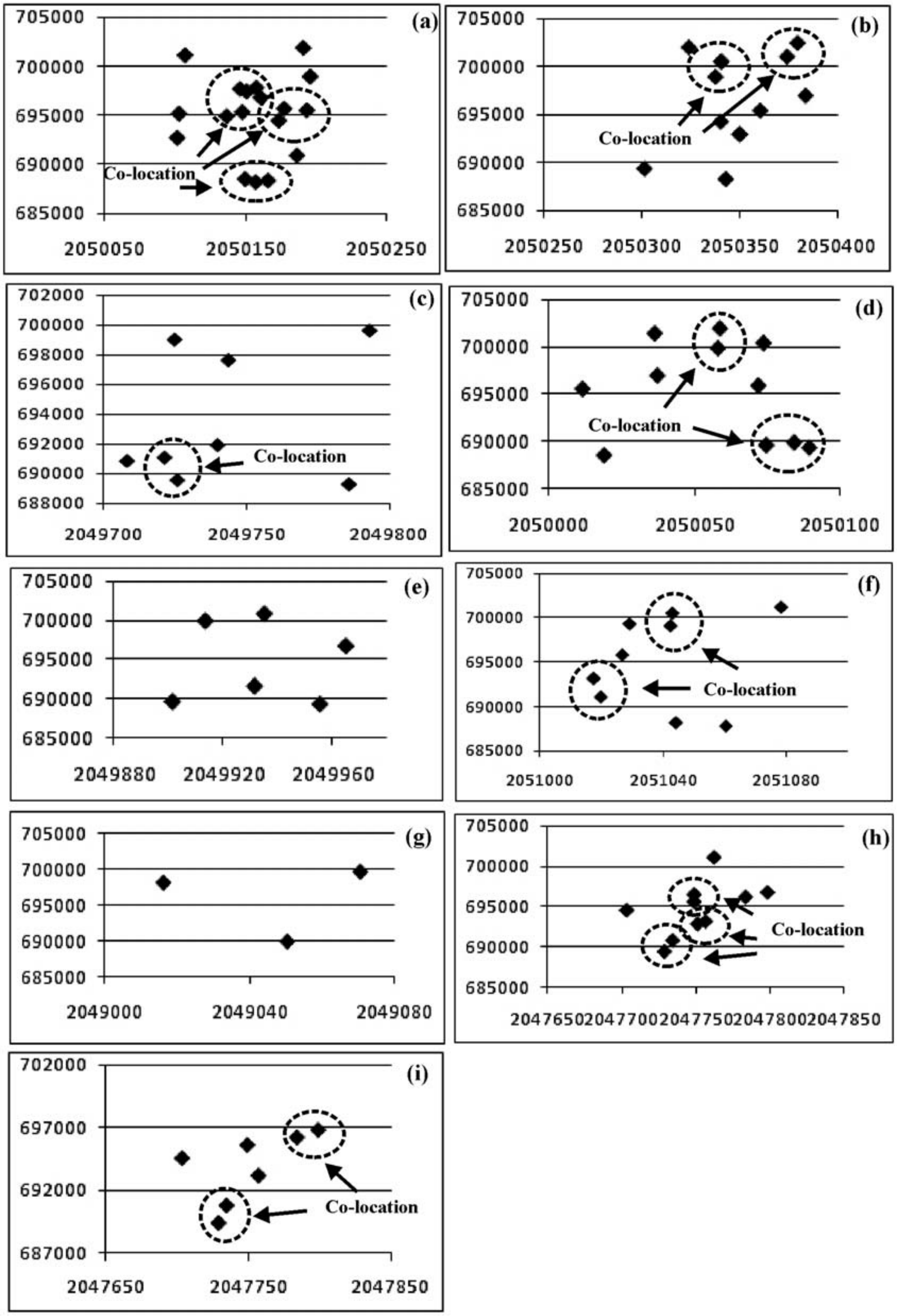


Figure 6.8: Spatial distributions of SKP instances after initial determination of co-location algorithm

F) Short Overlay (SO)

The ITRC at the NCDOT indicated three short overlay (SO) treatment strategies. All of them are chosen to illustrate the results of the proposed co-location mining algorithm for each SO treatment strategy. As seen in Figure 6.9a and Figure 6.9b, there are clusters surrounding the SO treatment, but no cluster surrounding another SO treatment in Figure 6.9c.

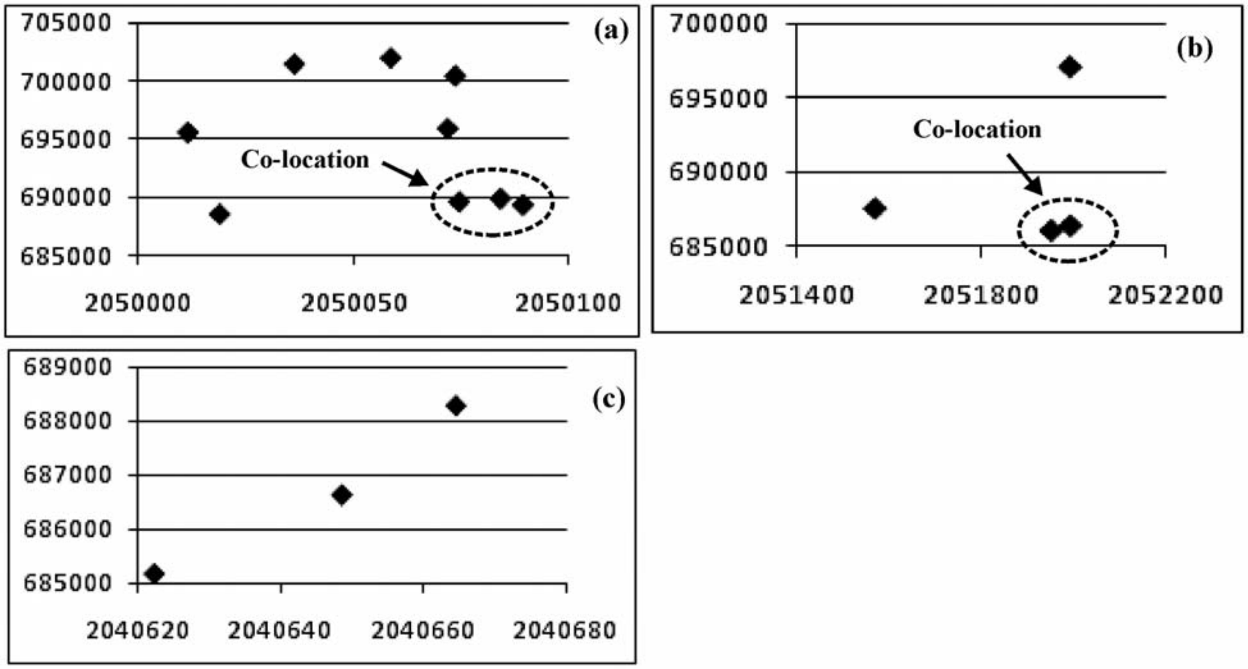


Figure 6.9: Spatial distributions of SO instances after initial determination of co-location algorithm

6.3.2.3 Pruning

The above generated candidates of co-location events for each treatment strategy may include incorrect determination. The purpose of pruning is to remove the non-prevalent co-locations from the candidate prevalent co-location set so that the further co-location mining rule induction is reliable. To this end, cross-correlation criterion of spatial attributes is applied to eliminate those non-prevalent co-location instances. The computation of cross-correlation is modeled in Eq. 4.5, with which we have co-correlation matrix as follows:

$$\underset{946\times946}{\Sigma} = \left\{ \begin{matrix} \rho_{11} & \rho_{12} & \cdots & \cdots & \rho_{1\times946} \\ & \rho_{22} & \cdots & \cdots & \rho_{2\times946} \\ & & \cdots & \cdots & \cdots \\ & & & \cdots & \rho_{945\times946} \\ & & & & \rho_{946\times946} \end{matrix} \right\} \qquad (6.8)$$

With observation to the coefficient of cross-correlation matrix, cross-correlation coefficients threshold is set at 0.95, i.e.,

$$p_{i,j} = \begin{cases} \text{non}-\text{correlated} & \text{if } \mathrm{p_{i,j}} \le 0.95 \\ \text{correlated} & \text{if } \mathrm{d_{i,j}} > 0.95 \end{cases} \qquad d_{i,j} \subseteq \sum_{946\times946} \qquad (6.9)$$

With the above given threshold, all of candidates prevalent co-location events are kept without pruning.

6.3.3 Generating Co-location Rules

Accompanying with the generation of co-location set, the co-location rules with the user defined constrain conditions (threshold) from the prevalent co-locations and their table instances can be generated (see Figure 6.10), i.e.,

IF (Dist_ij <= 25000 **AND** Gramma <=85 **AND** cross-correlation<=0.95) Co-location
ELSE Non_Co-location

Figure 6.10: Generated co-location rules

6.4 Experiment of Co-Location Decision Tree (CL-DT) Induction

6.4.1 Basic Steps of CL-DT Induction

With the above generated prevalent co-location events, the CTree for Excel tool is applied to create the decision tree. The steps include (Zhou et al., 2010a; 2010b):

Step 1: Load Pavement Database

As described in Chapter 5.2 and Chapter 5.3, pavement database has to be first loaded into the software. In this experiment, the loaded data is a prevalent co-location database, i.e., they have been “pre-processed” using co-location mining algorithm. Similarly, the distress data (non-spatial data) in the database, such as N, L, M and S, will be quantified into 100, 75, 50, and 25.

Step 2: Data Inputs

Similarly, some parameters to optimize the processes of decision tree generation will be input. These parameters include:

(1) ***Adjust factor of categorical predictor***: While growing the tree, child nodes are created by splitting parent nodes. Which is a predictor to use for this split is decided by certain criterion. Because this criterion has an inherent bias towards choosing predictors with more categories, thus, input of adjust factor will be able to adjust this bias.

(2) ***Minimum node size criterion***: While growing the tree, whether to stop splitting a node and declare the node as a leaf node will be determined by some criteria that we need choose. These criteria are the same as those adopted in Chapter 5, i.e.:

   a. *Minimum node size*: A valid minimum node size is between 0 and 100.

   b. *Maximum purity*: An effective value is between 0 and 100. Stop splitting a node if its purity is 95% or more. Also, stop splitting a node if number of records in that node is 1% or less of total number of records.

   c. *Maximum depth:* a valid maximum depth is greater than 1 and less than 20. Stop splitting a node if its depth is 6 or more.

(3) ***Pruning option:*** This option allows us to decide whether or not to prune the tree when tree is growing, which can help us to study the effect of pruning.

(4) ***Training and test data:*** In this research, a subset of data is used to build the model and the rest to study the performance of the model. Also, a random selection of the test set at a ratio of 10% is adopted.

6.4.2 Experimental Results

A) Induced Decision Tree

With the above data input, a decision tree is generated. The corresponding information for decision tree, including misclassified data percentage, time taken, total number of nodes, number of leaf nodes, and number of levels is listed in Table 6.5.

Table 6.5: Information of the induced decision tree using CL-DT algorithm

Tree Information		**% Misclassified**		**Time Taken (Second)**	
Total Number of Nodes	22	Training Data	21.7%	Data Processing	1
Number of Leaf Nodes	14	Test Data	15.3%	Tree Growing	2
Number of Levels	13			Tree Pruning	1
				Tree Drawing	5
				Classification using final tree	1
				Rule Generation	19

B) Induced Decision Rules

After the decision tree is induced, the tree is further processed to induce decision rules. The decision rules are directly induced in this research by forming a conjunct of every test that occurs on a path between the root node and a leaf node of a tree, i.e., top-to-bottom mode. Thus, the decision rules are first induced by ordering all the classifications, and then using a fixed sequence to combine them together. After the above processing, 12 rules are generated. Finally 6 rules are induced and depicted in Figure 6.11. The quality of the individual rules is measured by Support, Confidence, and Capture (see Table 6.6).

Table 6.6: Support, confidence and capture for each generated rules

Rule ID	**Classes**	**Support**	**Confidence**	**Capture**
0	NO	100.0%	89.2%	92.2%
1	CP	80.0%	95.0%	79.9%
2	SKP	79.2%	83.6%	83.8%
3	FDP	83.2%	84.4%	77.7%
4	PM2	83.5%	94.1%	79.2%
5	PM1	89.0%	88.8%	90.2%
6	SO	83.3%	76.3%	89.5%

```
Rule 1:
   IF (AN" >=5 AND "AS_" =0 AND "BK" ='50' AND "RF" ='100' AND "RT"='75' OR  "RT"='100' AND "RV"='100' AND
      "RQ"='75' AND "RATING" > '73' )
   THEN CP

Rule 2:
   IF ("AN" >=3 AND "AS_" >=1 AND "BK" ='100' AND "RF" ='100' AND "RV"='100' OR "RV"='75' AND "RQ"='75'
   OR "RQ"='50' AND "RATING" >= '43')
   THEN FDP

Rule 3:
   IF ("AN" >=2 AND "AN" <= 6 AND "AS_" >=0  AND "BK" ='100' AND "RF" ='100' AND "RV"='100' AND "RQ"='75'
      AND "RATING" >= '45' AND "RATING" <= '65').
   THEN PM1

Rule 4:
   IF ("AN" >=2 AND "AN" <= 6 AND "AS_" >=0  AND "BK" ='100' AND "RF" ='100' AND "RV"='100' AND "RQ"='75'
      AND "RATING" >=1 '1' AND "RATING" < '45')
   THEN PM2

Rule 5:
   IF "AN" >3 AND "AM" >1 AND "AS_" =0  AND "RF"='100'  AND "RV"='100' OR "RV"='75' AND "RQ"='75' AND
      "RATING" >= '63' AND "RATING" < '90')
   THEN SKP

Rule 6:
   IF "AN" >=6 AND "AM" =0 AND "AS_" =0  AND "BK"='100' AND "RF"='100'  AND "RV"='100' OR "RV"='75' AND
      "RQ"='75' AND "RATING" >= '60' AND "RATING" <= '75')
   THEN SO

Rule 7:
   IF "AN" >=9 AND "RATING" <= '100')
   THEN Nothing
```

Figure 6.11: The final rules after verification and post-processing

6.5 Mapping of CL-DT-based Decision of M&R

With the rules induced above, the M&R strategies can be predicted and decided for each road segment in the database using the rules. In other words, the operation using the co-location decision tree only occurs in the database, and thus the results cannot be visualized and displayed on either map or screen. Thus, this research employed ArcGIS software in combination with the above induced results to create the map of decision-making for maintenance and rehabilitation. The basic operation is the same as that described in Chapter 5, i.e., taking the above each rule as a logic query in ArcGIS software, and then queried results are displayed in the ArcGIS layout map. In order to compare the results, the rehabilitations suggested by engineers at the ITRE of North Carolina State University are superimposed with the decisions made at this research. As seen from Figure 6.12 through Figure 6.17, each rehabilitation strategy derived in this research can be located with its geographical coordinates, and visualized with its spatial, non-spatial data and different colors.

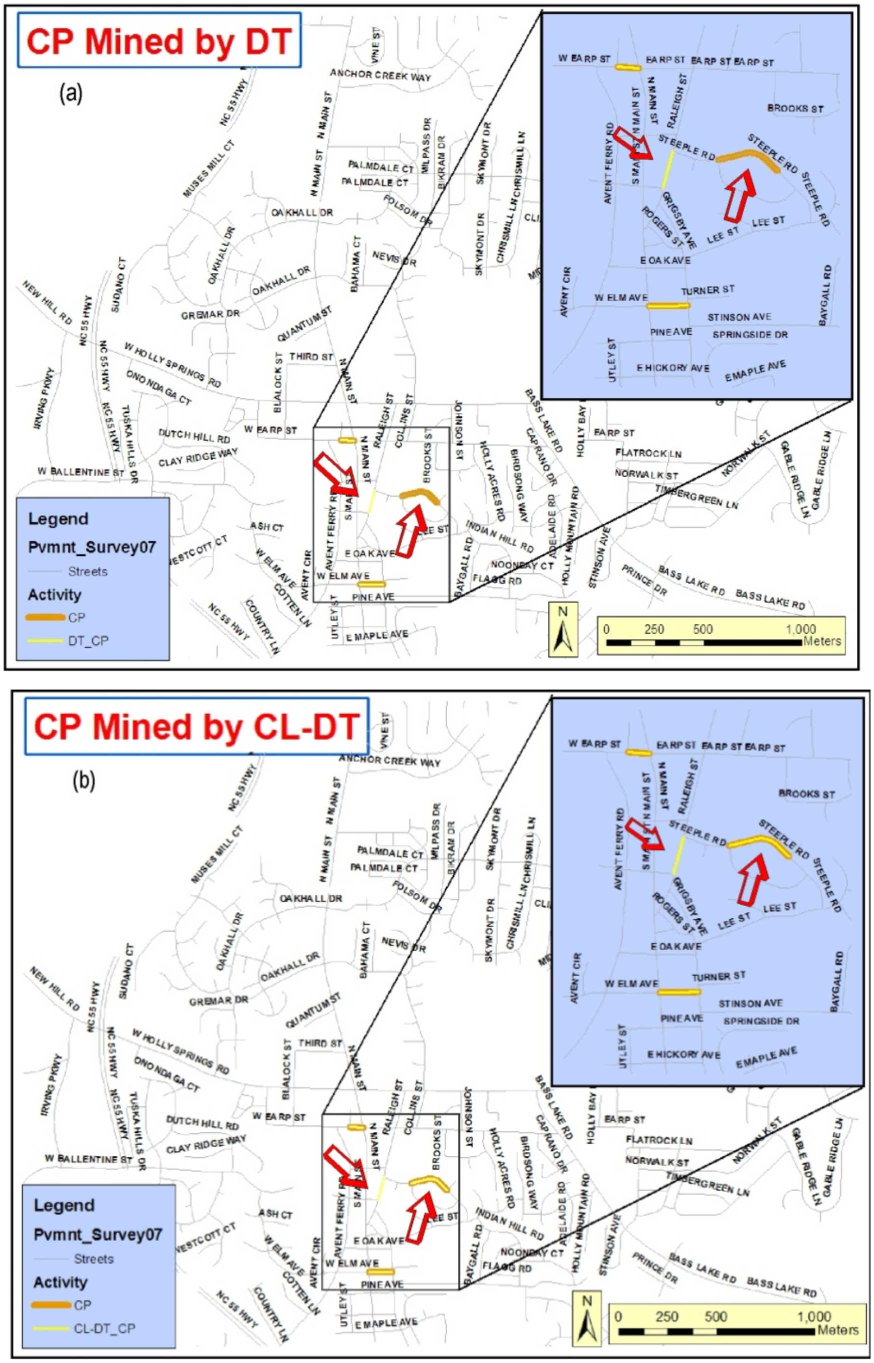


Figure 6.12: Comparison analysis of the CP decision of road rehabilitation made by DT (described in Chapter 5) and the proposed CL-DT (described in Chapter 6), both of which are compared to the CP decision made (provided) by the ITRC at the NCDOT

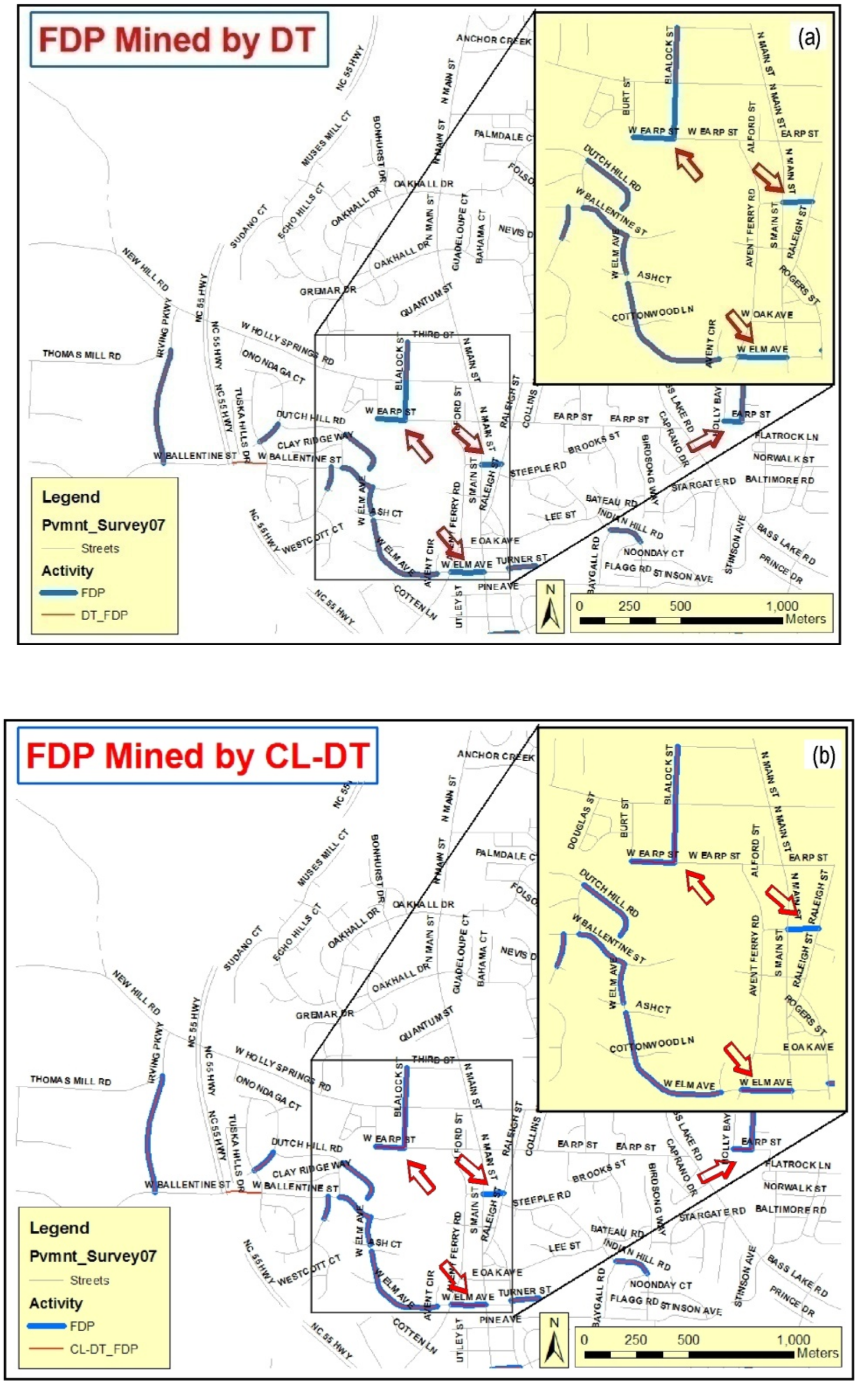


Figure 6.13: Comparison analysis of the FDP decision of road rehabilitation made by DT (described in Chapter 5) and the proposed CL-DT (described in Chapter 6), both of which are compared to the FDP decision made (provided) by the ITRC at the NCDOT

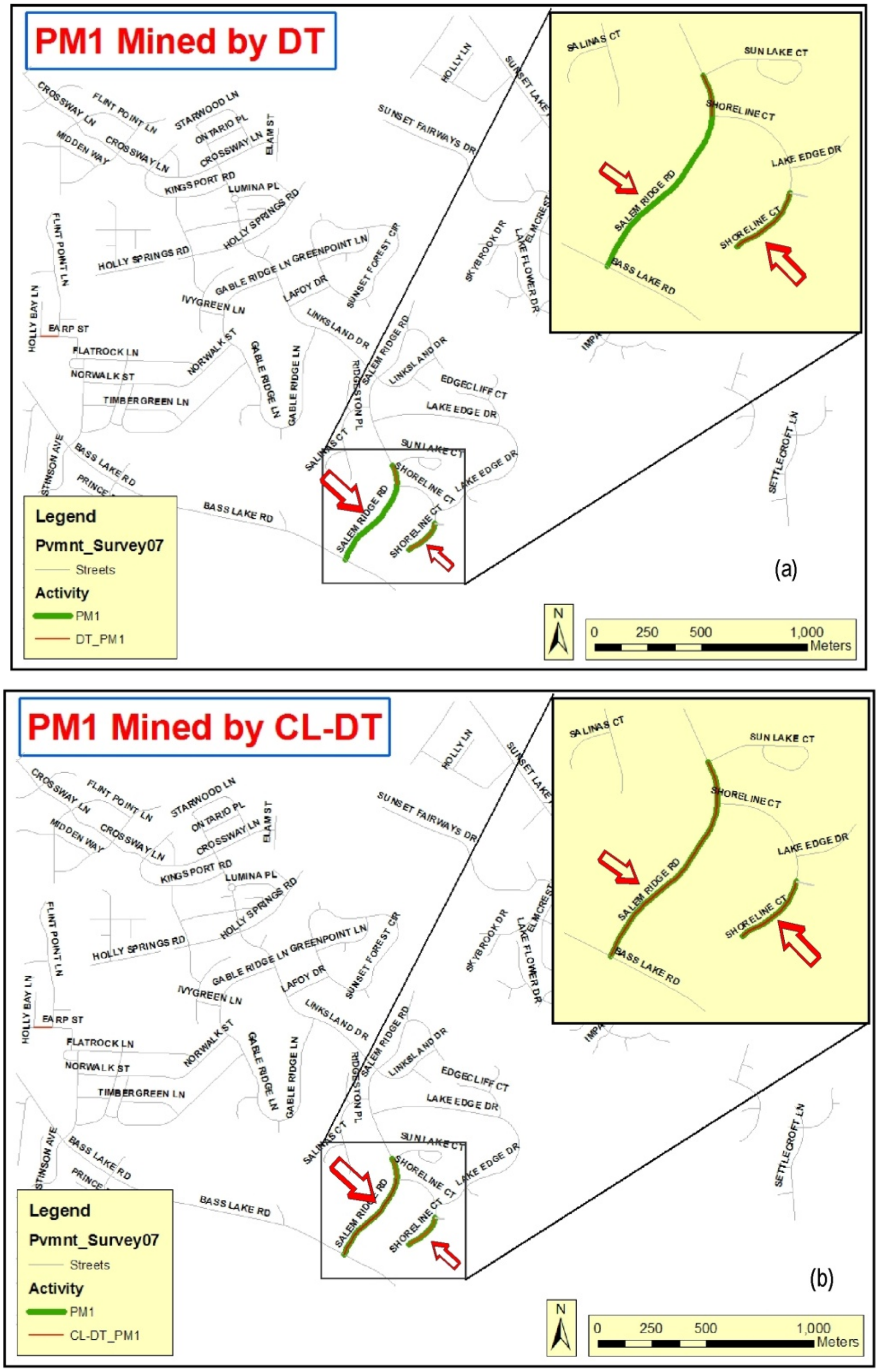


Figure 6.14: Comparison analysis of the PM1 decision of road rehabilitation made by DT (described in Chapter 5) and the proposed CL-DT (described in Chapter 6), both of which are compared to the PM1 decision made (provided) by the ITRC at the NCDOT

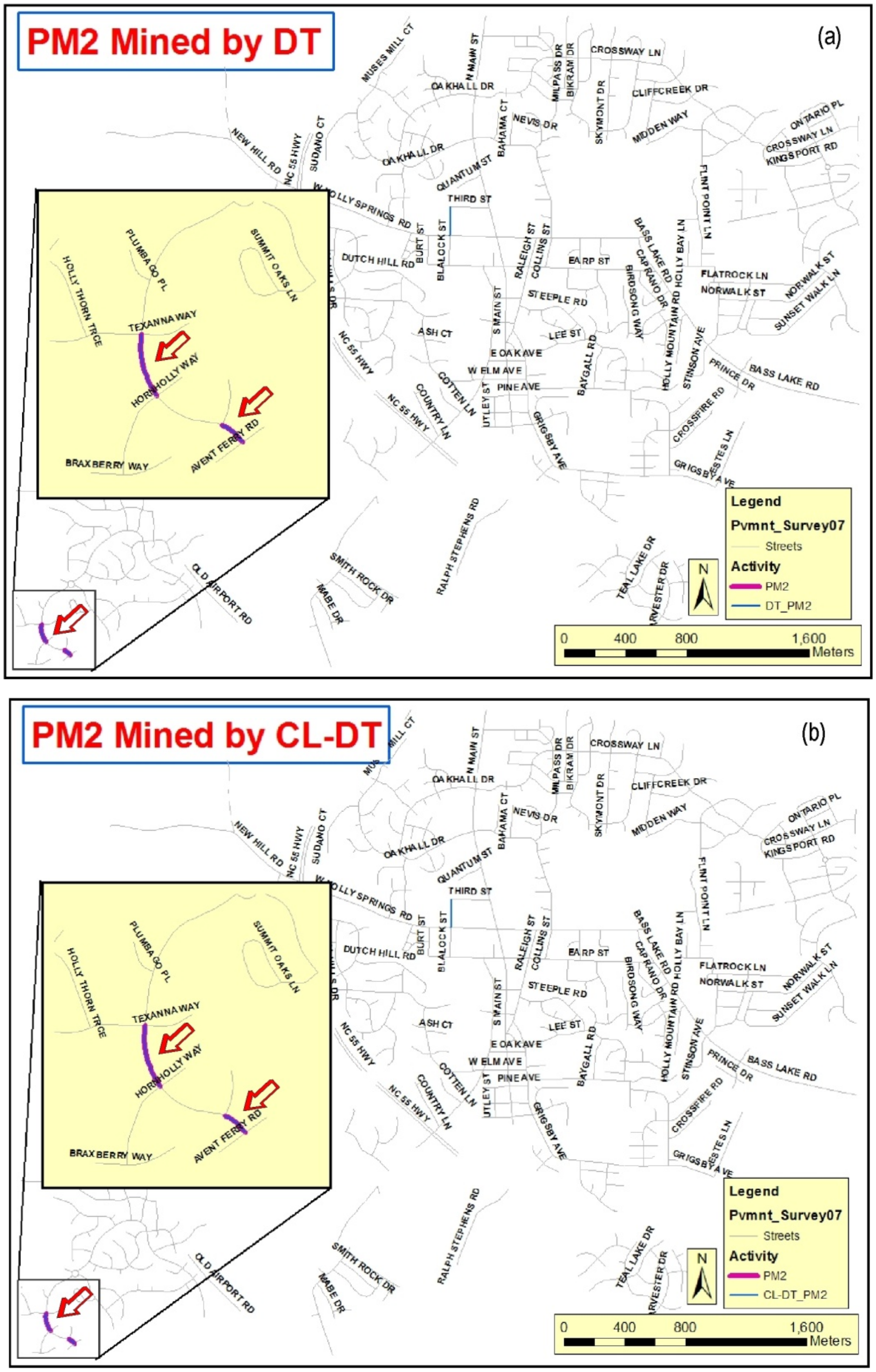


Figure 6.15: Comparison analysis of the PM2 decision of road rehabilitation made by DT (described in Chapter 5) and the proposed CL-DT (described in Chapter 6), both of which are compared to the PM2 decision made (provided) by the ITRC at the NCDOT

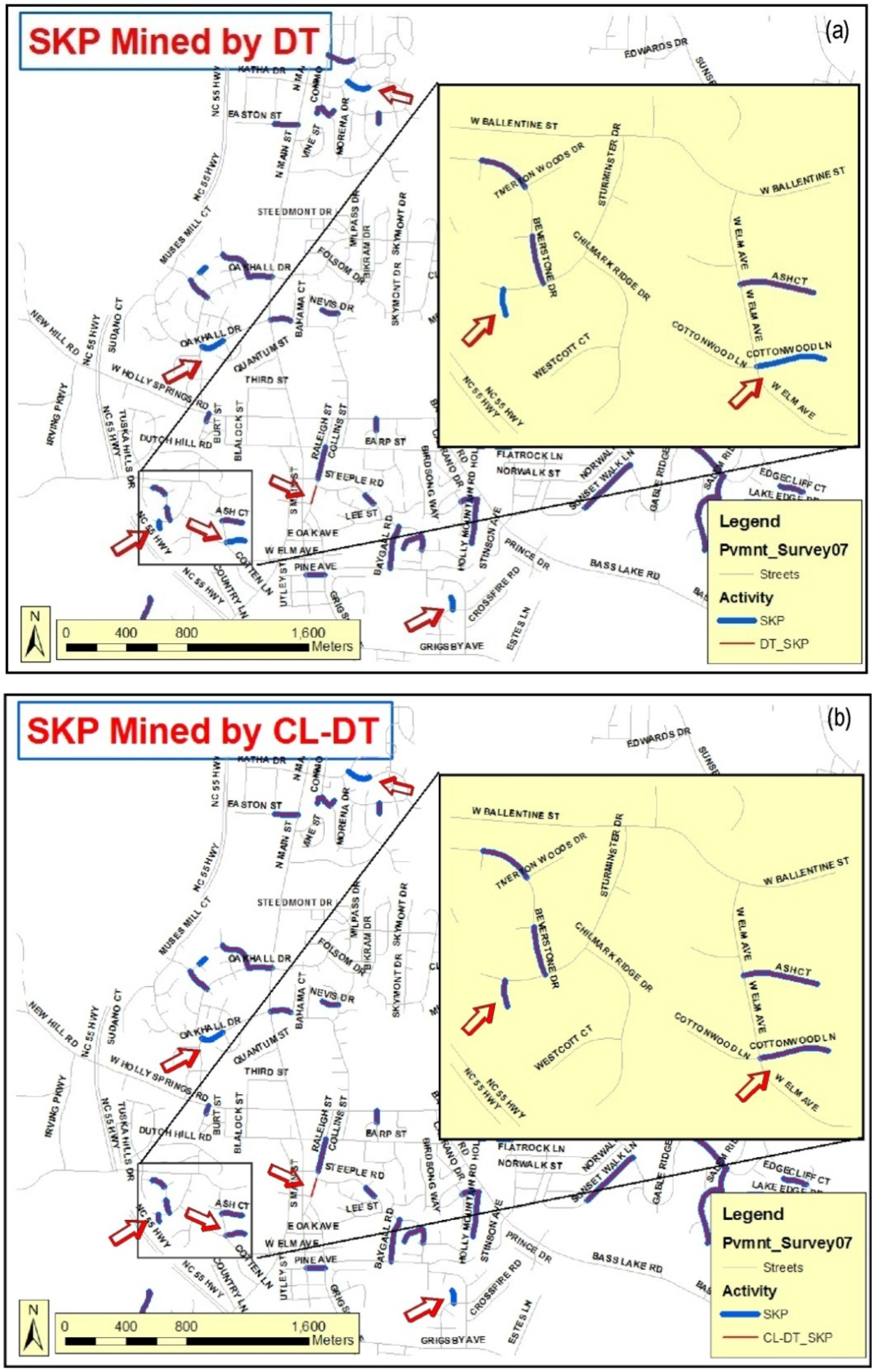


Figure 6.16: Comparison analysis of the SKP decision of road rehabilitation made by DT (described in Chapter 5) and the proposed CL-DT (described in Chapter 6), both of which are compared to the SKP decision made (provided) by the ITRC at the NCDOT

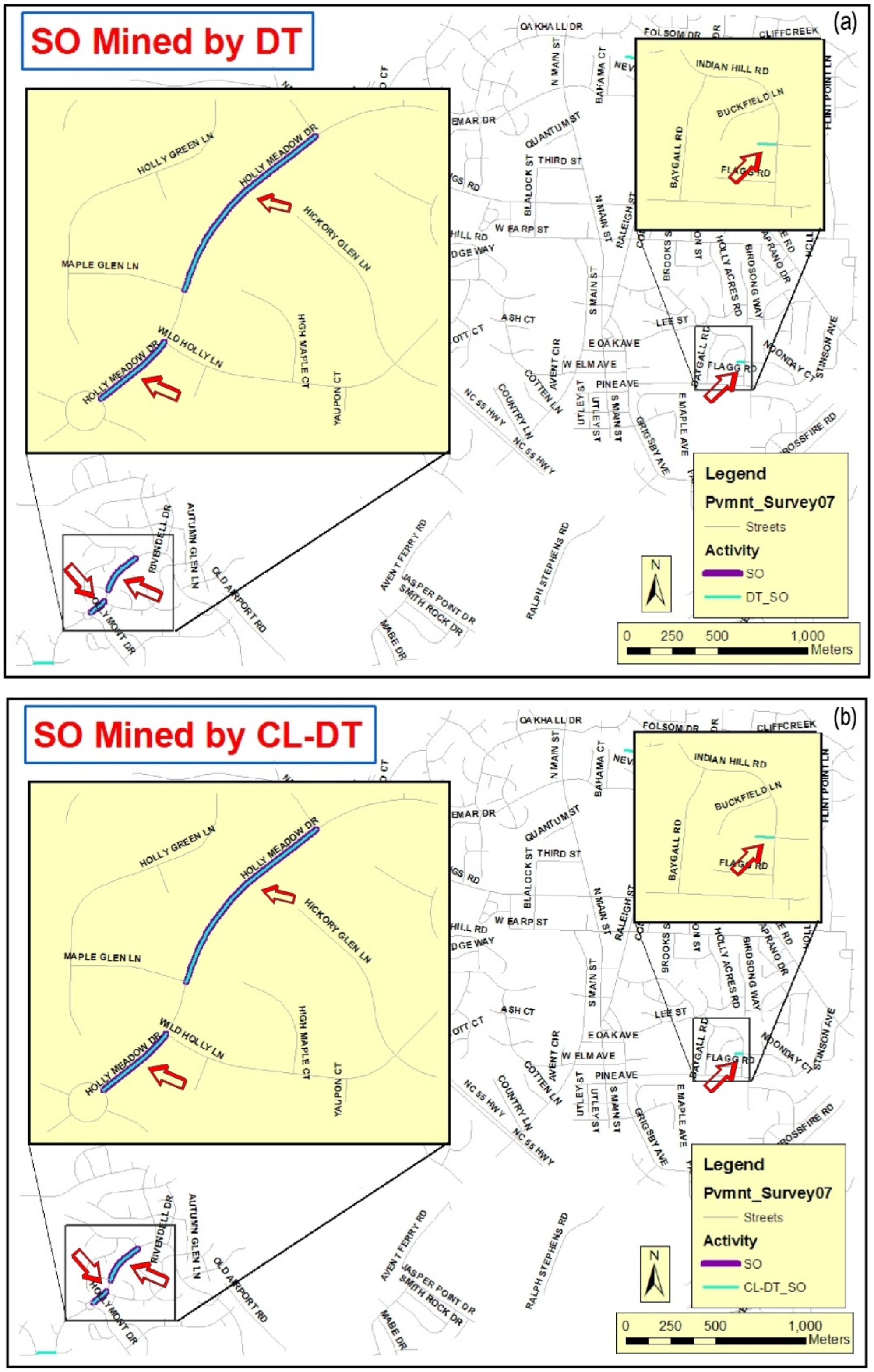


Figure 6.17: Comparison analysis of the SO decision of road rehabilitation made by DT (described in Chapter 5) and the proposed CL-DT (described in Chapter 6), both of which are compared to the SO decision made (provided) by the ITRC at the NCDOT

6.6 Comparison Analysis and Discussion

6.6.1 Comparison Analysis for the Induced Decision Tree Parameter

The proposed co-location decision tree (CL-DT) method should have many advantages over the traditional decision tree method in the effectiveness and accuracy of decision tree (decision rules) generation when applied in the decision-making of road maintenance and repair. In order to validate this conclusion, we compare the tree induction information for the two methods and the results are listed in Table 6.7. As seen in Table 6.7, the total number of nodes, number of leaf nodes, and number of levels decreases 51%, 62% and 35%, respectively. Thus, computational time will largely decrease. Accuracy of decision tree increases.

Table 6.7: Comparison of tree information parameters between DT and CL-DT algorithm

Tree Information	**Methods**		**Decreasing percentage**
	DT	**CL-DT**	
Total Number of Nodes	72	35	51%
Number of Leaf Nodes	37	14	62%
Number of Levels	20	13	35%

6.6.2 Comparison Analysis for the Misclassified Percentage

Also, we check the misclassified percentage, and the results are listed in Table 6.8. As seen in Table 6.8, the misclassified percentage for the training data decreases from 61.2% to 9.7%. This is probably caused by the fact that we used co-location mining technology to delete any non-prevalent candidate co-location instances. As a result, the training data contributed to the decision tree induction.

Table 6.8: Comparison of misclassified percentage between DT and CL-DT algorithm

Misclassified percentage	**Methods**	
	DT	**CL-DT**
Training Data	61.2%	9.7%
Test Data	60.0%	8.3%

6.6.3 Comparison Analysis for the Computational Time

Theoretically, the proposed CL-DT method should save much computational time, since the “pre-processing” method uses co-location mining technology, which deletes the non-prevalent co-location events. In order to verify this conclusion, we retrieved the computational time of data processing, tree growing, tree pruning, tree drawing, classification using final tree, and rule generation from the computer for the two methods. The results are listed in Table 6.9. As observed in Table 6.9, the time taken for the tree growing, tree drawing and rule generation is largely decreased. The time taken for rule generation decreases by 20%.

Table 6.9: Comparison of the computation time between DT and CL-DT algorithm

Items	Time taken for two methods (second)		% decreasing
	DT	CL-DT	
Data Processing	1	1	Rounded to 1"
Tree Growing	6	2	66%
Tree Pruning	1	1	Rounded to 1"
Tree Drawing	10	4	60%
Classification using final tree	1	1	Rounded to 1"
Rule Generation	35	15	20%

6.6.4 Comparison Analysis of Support, Confidence and Capture for Rule Induction

Another comparison analysis is for support, confidence and capture of training data when inducing the decision rules. The results for the two methods are listed in Table 6.10. As observed in Table 6.10, the percentage of support, confidence and capture for training data in FDP treatment strategy increase from 71.6%, 55.6% and 66.2% to 83.2% , 84.4% and 77.7%, respectively. This means that most of training data actively contributes the decision rule induction, which demonstrates that the co-location mining method can largely increase the effectiveness of decision tree/rules induction.

Table 6.10: Comparison for support, confidence and capture for two methods in each generated rules

Rule ID	Strategies	Support		Confidence		Capture	
		DT	CL-DT	DT	CL-DT	DT	CL-DT
0	NO	100.0%	100.0%	86.7%	89.2%	93.0%	92.2%
1	CP	60.7%	80.0%	100.0%	95.0%	75.6%	79.9%
2	SKP	60.5%	79.2%	66.7%	83.6%	82.5%	83.8%
3	FDP	71.6%	83.2%	55.6%	84.4%	66.2%	77.7%
4	PM2	80.2%	83.5%	100.0%	94.1%	73.1%	79.2%
5	PM1	81.3%	89.0%	71.4%	88.8%	85.3%	90.2%
6	SO	81.6%	83.3%	66.7%	76.3%	73.5%	89.5%

6.6.5 Verification of the Quantity of Each Treatment Strategy

As mentioned earlier, the ITRC at the NCDOT has indicated quantity of six treatment strategies at different road segments in the study area (four counties at North Carolina). Theoretically, the proposed CL-DT method should find the same quantity and location of each treatment strategy as those proposed by the ITRC at the NCDOT, since the proposed CL-DT applied the expert's knowledge from the ITRC. In order to verify this result, Table 6.11 lists the comparison for each treatment strategy proposed by ITRC at the NCDOT, and discovered by the proposed CL-DT method. Meanwhile, the quantity of each treatment strategy discovered by DL method (Zhou et al., 2010a) is also listed in Table 6.11.

As observed in Table 6.11, the quantity discovered by CL-DT is very close to those proposed by the ITRC for each treatment strategy. Thus, the traditional decision tree method mines 56 skin patch (SKP) strategies, which is 9 differences from those proposed by the ITRC, while the CL-DT method mined 62 SKP treatments, which is only 3 differences from those proposed by the ITRC.

Table 6.11: Quantity comparison of different treatment strategies made by three methods

ID	Proposed Treatment Strategies	Methods	Quantity	Differences in quantity referred to NCDOT
1	Crack Pouring (CP)	NCDOT	3	
		DT	3	0
		CL-DT	3	0
2	Full-Depth Patch (FDP)	NCDOT	34	
		DT	29	5
		CL-DT	32	2
3	1" Plant Mix Resurfacing (PM1)	NCDOT	6	
		DT	7	1
		CL-DT	6	0
4	2" Plant Mix Resurfacing (PM2)	NCDOT	3	
		DT	4	1
		CL-DT	5	2
5	Skin Patch (SKP)	NCDOT	65	
		DT	56	9
		CL-DT	62	3
6	Short Overlay (SO)	NCDOT	3	
		DT	5	2
		CL-DT	4	1

6.6.6 Verification of the Location of Each Treatment Strategy

Also, the ITRC at the NCDOT has indicated locations of 6 treatment strategies at different road segments in the study area (four counties at North Carolina). Theoretically, the proposed CL-DT method should find the same location for each treatment strategy as those proposed by the ITRC at the NCDOT, since the proposed CL-DT applied the expert’s knowledge (distress for each road segment) from the ITRC. In order to verify this conclusion, Table 6.12 lists the comparison for each treatment strategy proposed by ITRC at the NCDOT, and discovered by the proposed CL-

DT method. Meanwhile, the locations of each treatment strategy discovered by DL method (Zhou et al., 2010a) are also listed in Table 6.12.

As observed in Table 6.12, the location differences referred to those proposed by CL-DT for skin patch (SKP) strategies is significant. In other words, 13 road segments for SKP strategy are different from those proposed by the traditional decision tree method, but only 3 differences by CL-DT method, when referred to those by the ITRC (also see Figure 6.16a and 6.16b).

Table 6.12: Location comparison of different treatment strategies made by three methods

ID	Proposed Treatment Strategies	From	Number	Difference in location referred to NDCOT
1	Crack Pouring (CP)	NCDOT	3	
		DT	3	1
		CL-DT	3	1
2	Full-Depth Patch (FDP)	NCDOT	34	
		DT	29	3
		CL-DT	32	1
3	1" Plant Mix Resurfacing (PM1)	NCDOT	6	
		DT	7	1
		CL-DT	6	0
4	2" Plant Mix Resurfacing (PM2)	NCDOT	3	
		DT	4	1
		CL-DT	5	1
5	Skin Patch (SKP)	NCDOT	65	
		DT	56	13
		CL-DT	62	3
6	Short Overlay (SO)	NCDOT	3	
		DT	5	2
		CL-DT	4	1

6.7 Discussion and Remarks for Co-Location Decision Tree Algorithm

With the existing shortcomings of the decision tree induction method discovered in Chapter 5, this chapter presented the theory and algorithm of a new decision tree induction, called *co-location decision tree (CL-DT).* The main purpose of the proposed algorithm is to utilize the characteristics of attribute co-location (co-occurrence) to find the co-occurrence rules. These rules are used to enhance the traditional decision tree induction algorithm.

With the above experimental results and comparison analysis, it can be concluded that the proposed CL-DT algorithm can better make a decision for pavement treatment maintenance and rehabilitation when compared to the traditional decision tree method (e.g., C5.0 algorithm), since the new proposed method considers the co-occurrence distinct events. This Chapter especially makes a comparison analysis for the induced decision tree parameter, the misclassified percentage, the computational time taken, support, confidence and capture for rule induction. This chapter also verified the quantity and location of each treatment strategy referred to those proposed by the ITRC at the NCDOT.

With the above experimental results and comparison analyses, it can be concluded that:

(1) The proposed CL-DT method has many advantages over the traditional decision tree method in the effectiveness and accuracy of decision tree (decision rules) generation, when applied in the decision-making of road maintenance and repair. With comparing the analyses of two methods, DT and CL-DT, it is concluded that the total number of nodes, number of leaf nodes, and number of levels decrease 51%, 62% and 35%, respectively.

(2) With comparison analysis of two methods, DT and CL-DT, it is concluded that the misclassified percentage for the training data decrease from 61.2% to 9.7%, which demonstrated that the training data can be fully played roles in contribution to decision tree induction.

(3) With the comparison of the two methods, DT and CL-DT, it is concluded that the time taken by data processing, tree growing, tree pruning, tree drawing, classification using final tree, and rule generation is largely decreased, which can achieve 20% for rule generation.

(4) With the comparison of two methods, it is concluded that the percentage of support, confidence and capture for the FDP treatment strategy increase from 71.6%, 55.6% and 66.2% to 83.2% , 84.4% and 77.7%, respectively. This means that most of training data contributes the decision rule induction.

(5) With comparison of the quantity of six treatment strategies proposed by the ITRC at different road segments in the study area and by CL-DT method, it is concluded that that the quantity discovered by CL-DT is much close to those proposed by the ITRC for each treatment strategy. For example, 56 skin patch (SKP) strategies were mined by the traditional decision tree method, which is 9 differences from those proposed by the ITRC, while only 3 differences for the proposed CL-DT method when compared to those proposed by the ITRC.

(6) With comparison of the locations of six treatment strategies at different road segments in the study area proposed by CL-DT method and by the ITRC at the NCDOT, it is found that there are 13 road segments for SKP strategy different from those proposed by the traditional decision tree method, but only 3 differences from CL-DT method, when compared to those by the ITRC.

References for Chapter 6

Zhou, G., and L. Wang (2010a). GIS and Data Mining to enhance pavement rehabilitation decision-making, *Journal of Transportation Engineering,* vol. 136, no. 4, February, 2010, pp. 332-341.

Zhou, G., L. Wang (2010b). Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation, *submitted to Transportation Research Part C,* Revised November 2010.

7. CONCLUSIONS

A) Main Contributions

The main contribution of this research is the development of the theory and algorithm of a new decision tree induction algorithm, called *co-location-based decision tree (CL-DT).* This idea stems from the fact that the shortcomings of the existing traditional decision tree induction algorithm have been discovered by Chapter 5 when applied in the decision-making of pavement treatment strategies. The proposed algorithm utilizes the co-location (co-occurrence) characteristics of spatial attribute data of the pavement database, i.e., one distinct event occurrence can associate two or multiple attribute value changes simultaneously in spatial and temporal domains. That is,

- In the spatial domains: This implies that the presence of two or more spatial objects is at the same location or at significantly close distances from each other. Co-location patterns indicate interesting associations among spatial data objects with respect to their non-spatial attributes.
- In the temporal domains: The event occurrence should be distinct, thus is called *distinct event-type.*

This research dissertation has given the detailed descriptions of algorithms and steps of realizing the proposed algorithm. First, the research gave the detailed co-location mining algorithm, including spatial attribute selection in pavement database, determination of candidate co-locations, determination of table instances of candidate co-locations, pruning the non-prevalent co-locations, and co-location rule induction. In this step, a hybrid constraint, i.e., spatial geometric distance constraint condition and distinct event-type constrain condition is developed. The spatial geometric distance constraint condition is a neighborhood relationship-based spatial joining of table instances of many prevalent co-locations with one prevalent co-location; and the distinct event-type constraint condition is a Euclidean distance between a set of attributes and its corresponding clusters center of attributes. This research dissertation also developed the spatial features pruning method using the multi-resolution pruning criterion, i.e., the cross-correlation criterion of spatial features is used to remove the non-prevalent co-locations from the candidate prevalent co-location set with a given threshold.

This research dissertation is especially focused on the development of the co-location decision tree (CL-DT) algorithm, which includes the attribute (non-spatial) data selection of the pavement database, co-location algorithm modeling, node merging criteria, and co-location decision tree induction. In this step, co-location mining rules are used to guide the decision tree generation and induce decision rules.

For each step, this research dissertation gave the detailed flowchart or outline, such as flowchart of co-location decision tree induction, co-location/co-occurrence decision tree algorithm, co-location/co-occurrence decision tree (CL-DT) algorithm, and outline of steps of SFS algorithm. Finally, this research used a pavement database covering four counties, which are provided by NCDOT, to verify and test the proposed method. The comparison analyses of different rehabilitation treatment decisions proposed by ITRC at the NCDOT, by the traditional decision tree induction algorithm and by the proposed new method were conducted. Some conclusions are drawn up and some findings are found (see the descriptions below).

B) Conclusions through This Research

Through this research, the following conclusions can be drawn.

(1) Advantages of applying traditional DT method for pavement M&R strategy decision-making are:

a) The DT technology can make a consistent decision for a pavement M&R strategy under the same road conditions, i.e., less interference from human factors.

b) The DT technology can greatly increase the speed of decision-making because the technology automatically generates decision-tree and decision rules if the expert knowledge is given, thus, saving time and cost of pavement management.

c) Integration of the DT and GIS can provide the PMS with the capabilities of graphically displaying treatment decisions; visualize the attribute and non-attribute data, and link data and information to the geographical coordinates.

(2) *Disadvantages of applying traditional DT method for pavement M&R strategy decision-making are*

a) Traditional DT induction methods are not as quite intelligent as people's expectation. In other words, the DT inducted by DMKD are not completely exact, thus, the post-processing and refinement are necessary.

b) Traditional DT induction methods for pavement M&R strategy decision-making only used the non-spatial attribute data. It has been demonstrated that the spatial data is very useful for enhancing decision-making of pavement treatment strategies.

c) A DT induction method is based on the knowledge acquired from pavement management engineer for strategy selection. A decision tree is used to organize the obtained knowledge in a logical order. Thus, decision trees can determine the technically feasible rehabilitation strategies for each road segment.

(3) *Significances of the proposed CL-DT method for pavement M&R strategy decision-making*

Since the DT induction methods are based on the knowledge acquired from pavement management engineer for rehabilitation strategy selection, different decision-trees can therefore be built if the knowledge changes. For example, the decision-trees were based on severity levels of individual distresses in this research. If the pavement layer thickness, and/or material type, is taken as knowledge, these decision-trees are different. This means the decision rules generated by different knowledge are different. Thus, successful implementation of this proposed CL-DT method is able to develop an "optimal" decision tree (decision rules), which greatly enhance the decision-making in pavement treatment strategies.

This research dissertation has verified the advantages through the experimental results and several comparison analyses including the induced decision tree parameters, the misclassified percentage, the computational time taken, support, confidence and capture for rule induction, the quantity and location of each treatment strategy. It can be concluded that:

a) The proposed CL-DT algorithm can make a good decision for pavement M&R strategy when compared to the traditional decision tree method (e.g., C5.0 algorithm);

since the new proposed method considers the co-location (co-occurrence) distinct events of spatial data in the pavement database.

b) The proposed CL-DT method has higher accuracy and effectiveness than the traditional decision tree method does. With comparison of the tree induction information, the total number of nodes, number of leaf nodes, and number of levels decrease to 51%, 62% and 35%, respectively.

c) With comparison of the misclassified percentage, it is found that the misclassified percentage for the training data using CL-DT method decreased from 61.2% to 9.7%. As a result, the training data can play roles in contribution to decision tree induction.

d) With the comparison of the computational time taken, it is concluded that the computational time taken for the tree growing, tree drawing and rule generation is largely decreased for CL-DT method, especially, computational time taken for the rule generation decreased to 20%.

e) The percentages of support, confidence and capture of the FDP treatment strategy increased from 71.6%, 55.6% and 66.2% to 83.2%, 84.4% and 77.7%, respectively. This means that most of training data actively contributes the decision rule induction.

f) With comparison of the quantity of six treatment strategies proposed by the ITRC and by CL-DT method at different road segments in the study area, it is concluded that the quantity discovered by CL-DT is much close to those proposed by the ITRC for each treatment strategy. For example, the traditional DT method mines 56 skin patch (SKP) strategies, which is 9 differences from those proposed by the ITRC; while the CL-DT method mined 62 SKP treatments, within which only 3 treatments are different from those proposed by the ITRC.

g) With comparison of the locations of six treatment strategies proposed by CL-DT method and by the ITRC at different road segments in the study area, it is concluded that 13 road segments for SKP strategy are different from those proposed by the traditional DT method, but only 3 differences by CL-DT method.

8. FUTURE WORK

With the initial effort in and obtained accomplishments from this research, future work may place the emphases on the following fields:

A) Assessment of Sensitivity of various Attributes

The decision-making of treatment strategies for a given pavement database in this research selected eight attributes and three geometric data. How to choose an attribute is a critical issue because a most appropriate choosing will result in partitioning the training set in an *optimized* manner. When a decision node relative to this attribute is created after a test, this node becomes the root of the decision tree. This means that the sensitivity of selecting various attributes is significant on decision making. Thus, the future research work is recommended on the *Attribute selection*. A rigorous model should be developed to optimally select attribute data and geometric data, i.e., considering the relationship such as co-location, co-occurrence, and cross-correlation. In addition, when selecting an attribute, the attribute selection measure should be developed as well*,* such as the existing measure of "information gain".

B) Pavement knowledge discovery using diverse data types

This research only uses ride quality (RQ) data as a control for distinct event occurrence. In fact, the pavement management treatment strategies used other distress data, such as alligator cracking (alligator none (AN), alligator light (AL), alligator moderate (AM), and alligator severe (AS)), block/transverse cracking (BK), reflective cracking (RF), rutting (RT), raveling (RV), bleeding (BL), patching (PA), and utility cut patching. Thus, future work should consider all of these distress data for distinct-event types.

On the other hand, the research in this dissertation only considers geographic coordinates of event, i.e., XY coordinates. Other spatial data such as pavement width of the section measured in feet from edge of pavement to edge of pavement, and the number of through travel lanes that exist on the section, etc. should be considered. Thus, the future work will consider these spatial data simultaneously.

C) Better spatio-temporal representations in geographic knowledge discovery

The current knowledge discovery (GKD) techniques for pavement treatment strategies generally use very simple representations of distress data and spatial relationships. The future work in the pavement decision tree techniques should recognize more complex geographic objects (lines and polygons) and relationships (non-Euclidean distances, direction, connectivity and interaction through attributed geographic space such as terrain). On the other hand, the time dimension will also need to be more fully integrated into these geographic representations and relationships.

D) User interfaces for geographic knowledge discover

The research in this dissertation has combined GIS, decision-tree, co-location (co-occurrence). However, the data mining and knowledge discovery needs to move beyond technically-oriented research to the broader GIScience and pavement management fields. Lastly, we need to build discovered pavement knowledge into GIS and spatial analysis, and require effective representations of discovered pavement knowledge that are suitable for GIS and spatial analysis. This may include GIS interfaces and intelligent tools for guiding pavement knowledge discovery and GIS spatial analysis.

List of Published Papers in Transportation Infrastructure Engineering During Ph.D. Study

a) Chapters in Books

(1) Analysis of Flexible Pavement Distresses on IRI Model, *Pavements and Materials: Modeling, Testing, and Performance,* pp. 150-160, [ed] Zhanping You, Ala R. Abbas, and Linbing Wang, *American Society of Civil Engineering (ASCE), Geo Institute* (978-0-7844-1008-0), 2009, Reston, VA, ISBN: 498704679

b) Published Papers

(1) **Zhou, G.,** L. Wang, GIS and Data Mining to enhance pavement rehabilitation decision-making, *Journal of Transportation Engineering,* vol. 136, no. 4, pp. 332-341*,* Feb., 2010.

(2) Lin, B.,' Z. Wang, **G. Zhou**, Optimizing the Freight Train Connecting Service Network of a Large-Scale Rail System**,** *The Transportation Research Board (TRB) 89th Annual Meeting,* Washington, D.C., January 9–13, 2010.

(3) **Zhou, G.** and M. Abbas, Integration RTK-GPS and remotely sensed imagery for travel time measurement, *American Society for Photogrammetry and Remote Sensing (ASPRS) Annual Meeting*, Baltimore, MD, March 9 – 13, 2009.

(4) Chen, P., and **G. Zhou**, Detecting and counting vehicles from small low-cost UAV images, *American Society for Photogrammetry and Remote Sensing (ASPRS) Annual Meeting*, Baltimore, MD, March 9 – 13, 2009.

(5) **Zhou, G.** and Jingyu Wei, Survey and analysis of land satellite remote sensing applied in highway transportations infrastructure and system engineering, *American Society for Photogrammetry and Remote Sensing (ASPRS) Annual Meeting*, Baltimore, MD, March 9 – 13, 2009.

(6) Wang, Linbing; Harris, C.; **Zhou, G**.; Cooley, L., Effect of Permeameter Size and Anisotropy on Field Pavement Permeability Measurements, *The Transportation Research Board (TRB) 88th Annual Meeting,* Washington, D.C., January 11–15, 2009.

(7) Zang, D. and **G. Zhou.** Linear Spatial Object Co-Registration between Imagery and GIS data for Spatial-Temporal Change Analysis of Transportation Network, *Journal of Spatial Information Engineering* , vol. 3, no. 3, 2008, pp. 345-352.

(8) Xie, X., and **G. Zhou**, A method on the quickly loading transportation themes using ArcObject, *Journal of Geospatial Science,* Vol. 33, no. 3, p. 169-171 (Chinese).

(9) **Zhou, G.,**, and L. Wang, Effective Analysis of Flexible Pavement Distresses on IRI Model Using LTPP Data*, Inaugural International Conference of the Engineering Mechanics,,* Minneapolis, Minnesota, May 18-21, 2008.

(10) **Zhou, G.,** and L. Wang, Integrating GIS and Data Mining to Enhance the Pavement Management Decision-Making, *The 8th International Conference of Logistics and Transportation,* Chengdu, China July 31 –August 2, 2008

(11) **Zhou, G.,** and L. Wang, IRI Model Enhancement For Flexible Pavement Design Using LTPP Data, *87th TRB Annual Meeting***,** Washington DC*,* January 13-17, 2008.

(12) **Zhou, G.,** and L. Wang, 3D In-Vehicle Navigation Using Photorealistic Urban Model For Intelligent Transportation System, *87th TRB Annual Meeting***,** Washington DC*,* January 13-17, 2008**.**

(13) Wang, D., L. Wang, **G. Zhou**, and G. Flintsch, Binder Film Thickness Effect on Aggregate Contact Behavior, *87th TRB Annual Meeting***,** Washington DC*,* Jan. 13-17, 2008.

(14) **Zhou, G.,** and D. Wei, Traffic Spatial Measures and Interpretation of Road Network Using Aerial Remotely Sensed Data. *2008 IEEE International Geoscience and Remote Sensing Symposium* (*IGARSS 2008*), Boston, MA, USA, July 7-11, 2008

(15) **Zhou, G.,** and D. Wei, Survey and Analysis of Satellite Remote Sensing Applied in Transportations Infrastructure and System Engineering, *2008 IEEE International Geoscience and Remote Sensing Symposium* (*IGARSS 2008*), Boston, MA, USA, July 7-11, 2008

(16) Zang, D. and **G. Zhou**. Road Network Spatial Data Co-Registration using Imagery-to-GIS Mining, *2007 IEEE International Geoscience and Remote Sensing Symposium* (*IGARSS 2007*), Barcelona, Spain , July 23 – 28, 2007.

c) <u>Under Peering Papers</u>

(1) **Zhou, G.,** and L. Wang. Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation, *submitted to Transportation Research Part C, revised,* November 2010.