

KNOWLEDGE DISCOVERY AND DATA MINING FROM FREEWAY SECTION TRAFFIC DATA

A Dissertation

Presented to

The Faculty of the Graduate School

University of Missouri-Columbia

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

by

VANESSA AMADO

Dr. Mark R. Virkler, Dissertation Supervisor

MAY 2008

The undersigned, appointment by the Dean of the Graduate,
have examined the dissertation entitled

**KNOWLEDGE DISCOVERY AND DATA MINING FROM FREEWAY
SECTION TRAFFIC DATA**

Presented by

Vanessa Amado

A candidate for the degree of

Doctor of Philosophy

And hereby certify that in their opinion it is worthy of acceptance.

Dr. Mark R. Virkler
Department of Civil & Environmental Engineering UM-C

Dr. Carlos Sun
Department of Civil & Environmental Engineering UM-C

Mr. Charles Nemmers
Department of Civil & Environmental Engineering UM-C

Dr. C. Alec Chang
Department of Industrial and Manufacturing Systems
Engineering UM-C

Dr. Kristen L. Sanford Bernhardt
Department of Civil & Environmental Engineering
Lafayette College

*To my nephews and nieces,
Adrián, Charleen, Diego,
Gerardo, Miguel, Mya, and the rest to come
for them to know that sky is the limit*

ACKNOWLEDGEMENTS

This research was funded both by the Missouri Alliance for Graduate Education and the Professoriate (MAGEP) and the Department of Civil and Environmental Engineering of the University of Missouri-Columbia. I wish to specially thank Dr. Lenell Allen for her outstanding dedication in helping minority students achieve their dreams, Mr. Charlie Nemmers for his continuous enthusiasm of my educational and professional success, and the Midwest Transportation Consortium for all the funding support that allowed me to attend national and student conferences. The research was conducted with traffic generated data from the Puerto Rico Highway and Transportation Authority (PRHTA) and the National Climatic Data Center, for which I would like to thank these agencies and all the magnificent people that provided their assistance during my visits and many phone- and email communications. The IBM Intelligent Miner for Data was used for all the analyses performed, for which I would like to thank the IBM Corporation for providing the software as part of their scholars program. The wonderful group of people at the Department of Civil and Environmental Engineering of the University of Missouri-Columbia, I do not know how I could have finished my dissertation if it had not been for you. Thank you!

Many thanks to the great group of professors who make up my dissertation committee: I appreciate Dr. Sun, Mr. Charlie Nemmers, and Dr. Chang for taking time off your busy schedules to read, edit, and give great suggestions to my work. I have always looked up to Dr. Sanford Bernhardt. You are a great role model and excellent mentor. I hope we can work together again. And Dr.

Virkler, I think you are probably the most patient man I know. I am truly thankful for all the outstanding support throughout this entire process and all the excellent advise.

To the wonderful group of people that were on my side through thick and thin, my family and friends, I am in debt to you all. My friends from Columbia, you guys were my family for all those years. Thanks to the 2002 Eno Development Conference I met my two dearest friends, Juanfer and Carola. Thank you guys for always picking up the phone and replying to every email. It would have been impossible for me to finish my dissertation without my personal consultants! My dear Gustavo Adolfo - thank you for being my best friend and partner through this arduous process. My life is better because of you. Last but not least, thanks to my family for supporting me and for their constant love. Love you all!

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
TABLE OF CONTENTS	iv
LIST OF TABLES	vi
LIST OF FIGURES.....	vii
ABSTRACT.....	ix
CHAPTER	
1. INTRODUCTION	1
1.1 Motivation.....	4
1.2 Problem Statement	6
1.3 Objectives	10
1.4 Overview of Research Approach.....	10
1.5 Scope	12
1.6 Document Outline	13
2. BACKGROUND.....	14
2.1 Literature Review	15
2.1.1ITS Data Sources.....	16
2.1.2Agencies that Archive ITS Data	22
2.1.3Possible Users of Archived ITS Data	24
2.2Intelligent Transportation Interlinked Systems	26
2.3Measuring Traffic Flow on Expressway Facilities	29
2.4 Selecting a Case Study Site.....	37
3. CASE STUDY – LAS AMERICAS EXPRESSWAY (PR-18) SAN JUAN METROPOLITAN REGION	46
3.1 Characteristics of Facility	47
3.2Problems with Facility	51
3.3Type of Data Available.....	54
4. RESEARCH APPROACH AND METHODS.....	57
4.1 Building the Data Mining Database	61
4.1.1 Archived Data	61
4.1.1.1 File Formats	62

4.1.1.2 Data Elements	63
4.1.1.3 Counter Location Information	63
4.1.1.4 Data Collection Technology	65
4.1.1.5 Quality Control Checks	65
4.2 Examining and Preparing the Data	66
4.2.1 Data Analysis	66
4.2.2 Processing the Data	67
4.2.2.1 Data Quality and Validity Checks	69
4.3 Evaluating the Data Mining Application	69
4.3.1 Association Method	71
4.3.1.1 Sample Size	75
4.4 Building the Model	77
5. EVALUATION OF THE MODEL(S)	83
5.1 Study I – Association Mining to Analyze Average Vehicle Speeds ..	85
5.2 Study II – Association Mining to Analyze Density	90
5.3 Study III – Association Mining to Analyze Accident Data	94
5.4 Study IV – Association Mining to Analyze Work Zones Data	100
5.5 Study V – Association Mining to Analyze Weekday LOS.....	107
5.6 Study VI – Comparison of Traditional and Data Mining Approach for Analyzing ITS Generated Data	115
6. CONCLUSIONS AND RECOMMENDATIONS	123
APPENDIX A: GLOSSARY AND LIST OF ACRONYMS	137
A1. Glossary	137
A2. List of Acronyms	139
APPENDIX B: STATISTIC TEST RESULTS	143
APPENDIX C: MINING PREPARATION	152
C1. Data Preparation	155
C2. Association Mining Preparation	158
REFERENCES	172
VITA	178

LIST OF TABLES

	Page
Table 2.1 ITS generated data sources (Adapted from Margiotta 1998).....	20
Table 2.2 Summary of traffic management center loop detector data archiving practices (Adapted from Albert 1999, Smith 2003, and Dahlgren et al. 2001).....	23
Table 2.3 Summary of traffic management center AVI data archiving Practices (Adapted from Albert 1999)	23
Table 2.4 Metropolitan transportation systems infrastructure, functions, and benefits (Adapted from USDOT 2000)	27
Table 2.5 HCM levels of service thresholds for a freeway segment (Adapted from HCM 2000)	35
Table 2.6 Relationship between speed, flow, and density (Adapted from HCM 2000)	36
Table 2.7 Type of ITS generated data archived by the 78 largest Metropolitan areas in the US	39
Table 2.8 Metropolitan areas considered for case study	44
Table 3.1 Facility characteristics per highway segment for 2001	48
Table 4.1 Data Elements	63
Table 4.2 Variables, description, position, and category of data as used in the mining tool	80
Table 4.3 Work zone, accident, and weather variables used in the study	82
Table 5.1 Summary of studies	84
Table 5.2 Accidents northbound	96
Table B.1 Statistical test results from northbound PR18 SJ1999	144
Table B.2 Statistical test results from southbound PR18 CG1999	147

LIST OF FIGURES

	Page
Figure 1.1	Location of PR-18 in the San Juan metropolitan region 9
Figure 3.1	PR-18 overpassing PR-21 49
Figure 3.2	PR-18 underpassing ave. américo miranda and overpassing PR-17 50
Figure 3.3	PR-18 overpassing PR-23 51
Figure 3.4	Density versus time for a typical monday southbound 53
Figure 3.5	Flow versus time for a typical monday southbound 53
Figure 3.6	Speed versus time for a typical monday southbound 54
Figure 4.1	KDD seven-step process 58
Figure 4.2	Basic data processing using the KDD process 60
Figure 4.3	Counters location 64
Figure 4.4	Apriori Algorithm (Adapted from Agrawal and Shafer 1996) 74
Figure 5.1	Association rules – accident type versus average vehicle speeds friday - northbound 86
Figure 5.2	Association rules – accident type versus average vehicle speeds friday - northbound 87
Figure 5.3	Association rules – accident type versus average vehicle speeds during work zone activity - northbound 89
Figure 5.4	Association rules – accident type versus average vehicle speeds during work zone activity - southbound 90
Figure 5.5	Association rules – work zones versus density monday - southbound 92
Figure 5.6	Association rules – work zones versus density Monday - northbound 93
Figure 5.7	Association rules – accident duration vs accident type 95
Figure 5.8	Support of days in which accidents occurred 97
Figure 5.9	Patterns within the association rules 98
Figure 5.10	Relationship of accidents during work zones with Accidents between 3+ vehicles 101
Figure 5.11	Relationships of five accident 102
Figure 5.12	Support of accidents that occurred during a work zone 103
Figure 5.13	Accident types during work zone and no work zone activity 105
Figure 5.14	Relationship of accidents during no work zone activity With a tree 106
Figure 5.15	Rules comparison south- and northbound 5:45AM – 9:30AM 107
Figure 5.16	Rules comparison south- and northbound 9:30AM - 1:15PM 109
Figure 5.17	Rules comparison south- and northbound 5:30PM - 7:15PM 110

Figure 5.18	Worst level of service during work zone activity tuesday....	111
Figure 5.19	LOS as head rules	112
Figure 5.20	Relationship between work zone activities and LOS During rainy weather	114
Figure 5.21	Mining tool statistics from analyzing levels of service – 24-hour	116
Figure 5.22	Single item sets per model (north- and southbound).....	116
Figure 5.23	Monday northbound density, flow, and speed graphs	118
Figure 5.24	Monday southbound density, flow, and speed graphs	119
Figure 5.25	Combination of methodologies – northbound direction	120
Figure 5.26	Combination of methodologies – southbound direction	121
Figure C1.1	Main create data window.....	153
Figure C1.2	Database table and flat file view	154
Figure C1.3	Find and replace tool in ms words.....	154
Figure C1.4	Data format and settings	155
Figure C1.5	Flat files selection.....	156
Figure C1.6	Field parameters for flat files	157
Figure C1.7	Summary for flat file used.....	157
Figure C2.1	Main mining tools window.....	158
Figure C2.2	Associations settings.....	159
Figure C2.3	Associations input data.....	160
Figure C2.4	Associations input fields	161
Figure C2.5	Associations parameters setting.....	163
Figure C2.6	Associations results.....	164
Figure C2.7	Associations summary.....	165
Figure C2.8	Example of association numerical rules	166
Figure C2.9	Example of items sets	168
Figure C2.10	Example of graphical association rules	169
Figure C2.11	Rule color scale.....	170
Figure C2.12	Lift legend.....	170
Figure C2.13	Associations statistics.....	171

KNOWLEDGE DISCOVERY AND DATA MINING FROM FREEWAY SECTION TRAFFIC DATA

Vanessa Amado

Dr. Mark R. Virkler, Dissertation Supervisor

ABSTRACT

The rapid development of intelligent transportation systems (ITS) has generated large amounts of data for transportation professionals. Currently, operators, planners, researchers, air quality analysts, transit providers, consultants, media, and others are using archived data. Benefits generated from these systems have already been accounted for in many large cities in the United States. Benefits include more detailed temporal data; alternative data to the existing data allowing the costs of data collection to reduce; data with greater geographic coverage; data that meets unmet data gaps in the past; and data that are on electronic media allowing the expedition of data analysis and the dissemination of information. However, analysis of archived ITS generated data can provide additional benefits for highway users. Additional research of the data generated from intelligent transportation systems will provide transportation professionals the ability to make better decisions. Nonetheless, the better utilization of archived data will take time, but the more experimentation with data will allow greater benefits.

In this research, a set of archived traffic data from Las Américas Expressway (PR-18) in the San Juan Metropolitan Region (SJMR) in Puerto Rico was examined. The case study is a facility that has gone through many improvements in the last ten years with the purpose of reducing traffic congestion. However, it has been a major challenge for the Puerto Rico Highway and Transportation Authority (PRHTA) since, even with the installation of a moveable barrier on a large portion of PR-18, congestion remains a problem. Traffic flow data, accident data, and work zone data from this facility were studied by means of data mining.

The methodology used was a combination of association mining and the knowledge discovery in databases (KDD) process. Association mining was used to learn about the hidden patterns within each model created and the KDD process was used as the framework that guided the entire process. The KDD process used consisted of seven steps: building the data mining database, examining and preparing the data, evaluating the data mining application, building the models, evaluating the models, preparing a list of the conclusions/knowledge gained, and the decisions. A total of six specific studies were developed in which different variables were studied using the association mining tools of the IBM Intelligent Miner for Data software package.

The objective was to gain knowledge from the data about interrelationship between the variables. Thus the results obtained from the mining tool used were excellent for the purpose of this research. The approach was found to be a source of valuable information that could not have been detected by the use of

traditional statistical analysis alone. The approach allowed the identification of: “red flags” during work zone operations; similar patterns in levels of service (LOS) between Tuesdays and Wednesdays and similar patterns in LOS between Mondays, Thursdays, and Fridays; and it allowed the analysis of LOS over time. The major benefit learned from applying data mining to ITS generated data was that it allowed the analysis of numerous variables from multiple levels of information.

The new knowledge provides the basis for more advanced studies to be developed. In addition, the methodology could be used at other locations to increase the quality of information available for decision-making on similar facilities.

1. INTRODUCTION

At the beginning of the twentieth century, restrictive measures addressing the danger that vehicles imposed on the population began to diminish, allowing for a growing popularity and usage of automobiles among the driver population of the United States (US). However, this lifting of restriction led to generations of traffic problems. By 1931 the Institute of Transportation Engineers (ITE) established a specialized area for traffic analysis after realizing the need for those specific studies (Greenshields and Weida 1978). This was followed by a thirty-year period of intensive development of traffic control techniques and theories of traffic flow. By the 1970s, an era of energy restrictions came about and the need to operate traffic streams with the greatest possible efficiency became a priority for most government agencies. More intensive efforts were put to the continuing improvement of traffic control techniques (Greenshields and Weida 1978).

In the last twenty years computer equipment and software have evolved tremendously, from computers that took large rooms for storage to notebook size computers that people can carry under their arms. In the same manner, solutions to traffic related problems are being found through technological improvements and enhanced efficiency. The technology used to alleviate congestion and improve capacity on expressways has developed rapidly in the last twenty years. Nowadays there is more use of transponder tags, message boards, closed circuit cameras, and many in-vehicle systems. Embedded

computer chips are being used as standard equipment in vehicles to alert drivers of their perimeter (front, behind, to the side), avoid accidents, and report accidents automatically (Bureau of Transportation Statistics (BTS) 2000). The combination of the technology used and the integration of data from various equipment are captured under the term ITS (US Department of Transportation (USDOT) and Federal Highway Administration (FHWA 1999). Currently, these systems are widely deployed through many metropolitan regions in the US and US territories to improve the mobility and safety of our surface transportation systems (USDOT and FHWA 1999).

While ITS has and will keep providing benefits for motorists as they take advantage of the better service provided, some of the gains may be offset by the expected growth in highway travel by 2025 (BTS 2000). Trends of embedding new technology into operations and management of the transportation systems will keep accelerating as new and more advanced technology is developed, however congestion will remain an issue. Thus, not only should new technology be developed to address these problems, but ways to interpret the data that are archived by means of the many ITS interlinked systems should also be created. The interpretation of data will improve the decision-making process in state agencies and will result in an improved operational highway system for motorists. It would be a huge task to try to understand all of a system's data at once, given that the amount of data archived in some traffic management centers (TMCs) is immense. ITS generated data is not archived for every ITS system, even in cities like Los Angeles, CA where the ITS infrastructure is extremely advanced.

The Texas Department of Transportation defines ITS data archiving as “the systematic retention and re-use of transportation data that is typically collected to fulfill real-time transportation operation and management data” (Texas Transportation Institute (TTI) 2001). Some studies have been conducted that provide a good basis for others interested in researching this area. A few examples are the evaluation of potential partnerships for the management of archived ITS data by Albert (1999); addressing issues such as the collection, process, archival and dissemination of traffic data by Dahlgren et al. (2002); analysis of archived ITS data in San Antonio, TX by the TTI and Texas A&M University System (1999); and addressing issues of data quality in archived ITS data in San Antonio, TX by Turner et al. (2000). All of these provide very interesting findings, however more research could yield significant benefits in the area of archived ITS generated data.

The purpose of this research was to provide a means to improve the monitoring of system performance by means of data mining. The data mining method chosen for this research was association mining, which was applied to traffic counts, accident, and work zones data to study the frequent traffic patterns in an expressway facility. Several specific studies were performed which allowed for a better description of these patterns. From some of the studies, the quality of traffic flow was measured for specific instances during a 24-hour period. The research gave some insights into the performance of an expressway system and it was concluded that the methodology could be of great importance to traffic

analysts. The methodology could also be used by state agencies to concentrate on improving the quality of traffic on specific locations.

1.1 Motivation

The growth of automated data collection by means of ITS, the failure to reuse archived data, and the need to ensure appropriate uses of the data archived are major concerns for state Departments of Transportation (DOTs). Generally, the data are being maintained by TMC personnel whose equipment collects the data, but who may have few resources or little motivation to make the data easily accessible to researchers (Dahlgren et al. 2002). Dahlgren et al. agree that the key to effective use of archived data is the “clear assignment of responsibility and adequate funding,” however funding is usually scarce for most government agencies (Dahlgren et al. 2002). Turner (2001) suggests that the reuse of archived traffic data could be of economical benefit to state agencies by avoiding the recollection of data. These already collected data could be used for other areas of transportation system development (Turner 2001).

Archived traffic data should be used to improve traffic management center performance (ITS Data Archiving Five-Year Program Description 2000). That is, given that the purpose of these centers is to provide motorists with real-time information on the condition of traffic including issues such as accidents, work zones, and detours for hazardous spills the information obtained from the data collected should be maximized. Researchers agree that there is a need for

better utilization of archived data, but that it is a time consuming task (Dahlgren et al. 2002). KDD has the potential to systematically approach this problem.

The concept of KDD has been used for the last decade to address the problems in understanding large datasets. In 1989, the term KDD was assigned to refer to the process of finding and interpreting patterns from data (Fayyad et al. 1996). KDD is a new name given to a practice that has been occurring for years. People have been generating large banks of data and extracting portions of it to create statistical models and obtain useful information. However, the framework provided by the KDD process allows researchers to systematically proceed through the exploration of these large banks of data. Carvalho (2007) describes KDD as an interactive and iterative process that consists of four steps that one plans before executing the extraction of knowledge from the data, whereas Nassar (2006) describes it as the “whole process of extraction of knowledge from data.” Nonetheless, the most popular definition has been proposed by Fayyad et al. (1996): “KDD is a process with many stages, non-trivial, interactive, for the identification of comprehensive, valid, and potentially useful standards from large data sets.”

It is clear that TMCs could make better use of the large amounts of traffic data they collect on a daily basis. Therefore, a set of applications to assist in the process of discovering useful knowledge by extracting patterns from raw data must be applied. The application of the KDD and data mining algorithms allow the latter to be performed.

1.2 Problem Statement

Archived ITS generated data are a rich resource that might improve a broad range of transportation decisions. However, these are rarely fully utilized. There has been a major focus on the use of archived ITS data to learn about congestion and incident management, especially in Texas (Turner 1997). Benefits have been obtained from these studies, and some cities have been able to use these results to reduce their congestion. However, the difficulties of accessing the data due to improper or incompatible format and the tediousness of the task have hindered the development of prediction models and/or algorithms to analyze archived ITS data. Given the amount of ITS generated data that is archived on a yearly basis (approximately 2,015 miles worth of ITS data), there is significant potential for new information to be obtained from such data (refer to Tables 2.2 and 2.3 for details of the ITS data archived).

The number of states that have accomplished the development of an ITS, using the latest technology, for most of their large metropolitan areas is approximately 12 states (refer to Table 2.7). It is actually a major challenge for most states due to various reasons, such as inconsistencies in data collection practices through the years; lack of people and/or equipment to collect the proper data; lack of electronic databases; lack of much needed infrastructure; lack of engineers, technicians, and specialized personnel; and lack of hardware and software among others. All of the latter items put together generate the major drawback, which is the cost of the project. However, defining ITS as a combination of old and new technology used to collect transportation related

data, we can easily say that every DOT has some type of ITS already established in their organization. Thus, the existence of archived ITS data is found in every DOT across the nation.

Due to the large amounts of clean and complete data collected every year by means of ITS in cities like Los Angeles, Houston, and Dallas the majority of the research work related to archived ITS data is being concentrated on these few metropolitan areas. However, there is much needed work to be done with the less than perfect data that it is currently being collected on the majority of the metropolitan areas in the US and US territories. These DOTs could benefit greatly from the research conducted using data that it is much similar to their own as opposed to future uncertain data.

The main focus of this research was to use archived ITS generated data in measuring quality of traffic flow and relating this to various data patterns. The latter was achieved by conducting several specific studies in which the relationship of different variables was mined using association algorithms. The better use of the archived ITS data to address particular problems, like improving system operation, quality of flow, and planning for system improvements were examined.

Archived traffic data from PR-18, Las Américas Expressway in the SJMR in Puerto Rico, was used as the case study. The San Juan Metropolitan Region was listed as one of the 78 largest metropolitan areas with ITS deployments in the US and US territories (Gordon and Trombly 2000). PR-18 is one of the four main highways that connect the heart of the SJMR from North to South. Figure

1.1 shows how PR-22 connects Las Américas Expressway (PR-18) in the North and Luis A. Ferré Expressway (PR-52) in the South part of the SJMR (Guía Urbana del Área Metropolitana 2000). PR-18 consists of five lanes in each direction and a moveable barrier allowing for one extra lane in the higher volume direction during the morning and afternoon commute. This expressway has gone through many renovations throughout the years in attempts to reduce congestion, but it has been a challenging task. Currently, the congestion has reduced somewhat compared to that in 1998, before the moveable barrier and the addition of one permanent lane in each direction, but it is still a major challenge for the PRHTA. There are currently over a dozen traffic counter stations collecting traffic data continuously on PR-18 (García 2002).

The San Juan Metropolitan Region has some interesting transportation characteristics. Currently, there are 146 vehicles per square mile in the central SJMR, making it one of the most congested urban roadway networks in the world and one of the highest vehicle density rates in the US (Pesquera and González 1996). More than one third (37%) of Puerto Rico's total population (1.3 million) resides in the SJMR and 63% of all jobs in the region are concentrated in its main centers. Pesquera, a former Secretary of the PRHTA, has said that the population of the SJMR generates more than 3.2 million vehicular trips per day, "giving the SJMR the highest traffic density in the world." (Pesquera and González 1996). Further, the number of daily vehicular trips are expected to increase by 45% between the year 1996 and 2010 (Pesquera and González 1996).

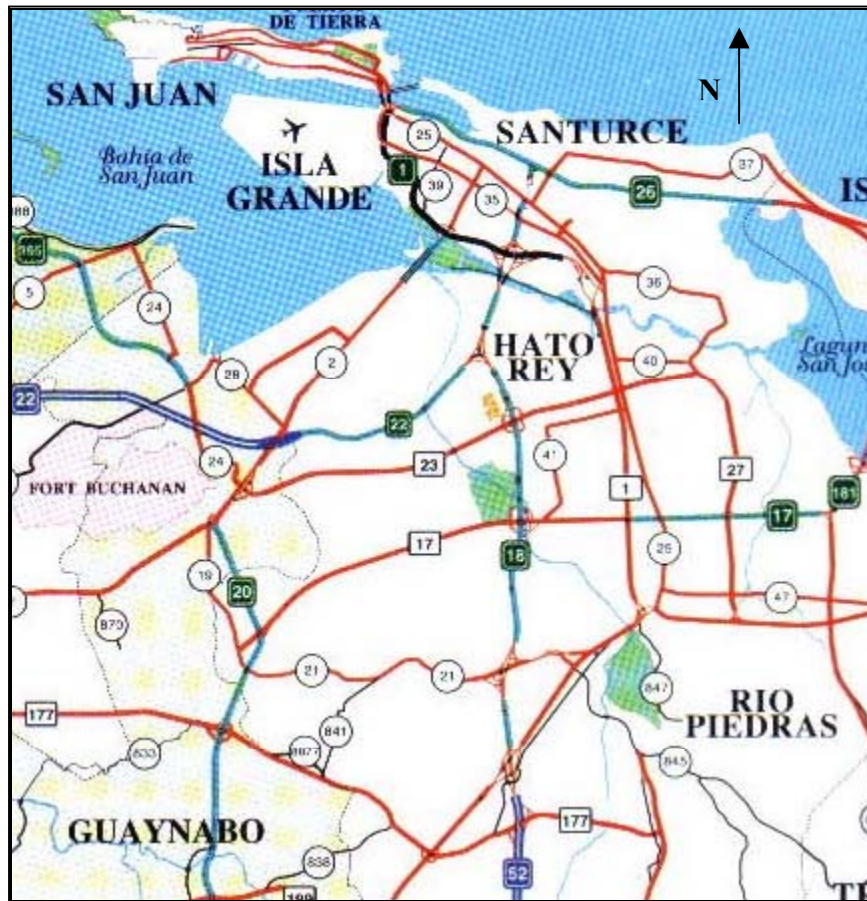


Figure 1.1 Location of PR-18 in the San Juan metropolitan region

(Source: Guía Urbana del Área Metropolitana, 2000 San Juan Metropolitan Area Metro Data. CD-ROM)

Because of the regular congestion experienced, this location provided a wealth of opportunities to use ITS data to measure the quality of traffic flows. In addition, the availability of data provided a great opportunity for this research project. As part of the research and to learn about the type of ITS data that are being collected and the amount of archived data, a study of the TMCs around US was conducted. This will allow the lessons learned here to be applied in other locations around the US that are similar to the SJMR.

1.3 Objectives

This research focused on the application of data mining algorithms to archived ITS generated data with the purpose of learning from the hidden patterns in the data in order to improve transportation decision-making for freeway and expressway facilities. Several statistical models were created from the frequent patterns generated by the traffic data. Even though this application was developed with data from a particular case study (PR-18), it could be readily transferable to archived data from any expressway system.

The approach integrated the KDD process and data mining algorithms. The KDD process was used as a framework to build the database and prepare the data, and the data mining algorithms were used to gain information from the data. The combination of these methods (KDD and data mining) was applied to each specific study.

1.4 Overview of Research Approach

The approach required several steps in order to accomplish the objectives of the research. These steps were:

1. Obtain several years worth of data from several traffic count stations within Las Américas Expressway (PR-18) from the PRHTA.
2. Obtain (from the PRHTA) and process work zone operations and accident data for PR-18 for the same year as the traffic count data. Processing these two datasets began rapidly once they were obtained because these were provided in hard copies. These two datasets were incorporated into

the working database as categorical data with the purpose of examining the operational conditions of PR-18 during these activities.

3. Conduct research on the characteristics of PR-18 (number of lanes and width, construction work conducted during the time of data that was collected, special events in the Old San Juan area (e.g., 1992 Regatta), operational hours of the moveable barrier, date when moveable barrier began its operations, natural disasters (e.g., hurricanes) during the time of data we will be working with, etc.
4. Create the working database.
5. Examine the data for missing and/or erroneous values.
6. Create and evaluate the model through statistical analysis: use statistical tests to gain insight into the distribution of the data, use association rules to mine frequent patterns generated in the working database, focus on particular patterns for further analysis, and learn about extreme behavior of drivers by examining the outliers.
7. Determine potential applications for the newly generated knowledge. For example, state agencies will be able to learn about the quality of traffic conditions by time and location within a particular expressway facility. Additionally, in the process of applying the methodology proposed in this document, state agencies will establish needs for data collection, data storage, and will be able to develop procedures for data management.
8. Prepare a set of conclusions and recommendations based upon the results.

1.5 Scope

Nowadays very sophisticated equipment, such as closed circuit television, has been developed for the collection of traffic related data. However, old technology, such as rubber pneumatic tube counters, is still used for collecting traffic count data. Furthermore, some states have more advanced equipment for collecting and integrating traffic data than others, and have taken more advantage of the data collected as well. But something that every state and US territory has in common is that they all archive traffic data, some more than others. Due to the relative ease of data collection with modern equipment, the amount of data collected has increased tremendously, causing the ITS data sources' databases to increase in size.

According to Turner et al. (2000) there are three major issues involving ITS data archiving: erroneous data, missing data, and data accuracy (Turner et al. 2000). Erroneous data is detected at most TMCs by comparing reported volume, occupancy, and speed values to minimum or maximum threshold values. Missing data is a common attribute of ITS traffic monitoring due to the continuous operation and occasional malfunctions of the collection equipment. Data accuracy is the traffic monitoring equipment's ability to reflect the actual traffic conditions, and it is a major concern when archiving ITS data because of the difference in accuracy requirements for each ITS data source (e.g. TMCs, Metropolitan Planning Organizations (MPOs), and DOTs) (Turner et al. 2000). Another major problem of archiving ITS data is the issue of format inconsistency causing difficulties in coordinating retrieval of archived ITS data (Winick 2002).

The latter is a common problem confronted by researchers interested in using archived ITS data from various sources.

These are issues that are associated with the assessment of archived ITS generated data but were touched upon only briefly in the research. The research addressed specific techniques for understanding and analyzing archived ITS generated data and developed a methodology for using data to improve the decision-making process for expressway systems, such as Las Américas Expressway (PR-18). The research also described various components of a metropolitan intelligent transportation system and the sources of ITS data.

1.6 Document Outline

This Chapter describes the research problem, the background and methods, and the expected results of the research. Chapter 2 contains a literature review and relevant background information from the areas of ITS and traffic flow. Chapter 3 gives a detailed description of the case study site, including characteristics and problems of the facility, and types of data available. Chapter 4 describes the approach taken and methods used in the research. The six specific studies that were conducted to evaluate the data mining tool as a means to measure the quality of flow are included in Chapter 5. The conclusions and recommendations are presented in Chapter 6. A glossary and list of acronyms have been included in the Appendix. Also in the Appendix is a simple step-by-step example for mining preparation to help the users that are willing to try the data mining tools of the IBM Intelligent Miner (IBM).

2. BACKGROUND

In the last twenty years, ITS have evolved rapidly due to the evolution in electronics and information technology; agencies' concerns over traffic congestion, traffic safety, and air quality; and governments realizing that new construction alone would not solve these problems (BTS 2000). This rapid development of ITS has generated large amounts of data for transportation professionals. These data are mostly collected by TMCs and used primarily for real-time applications. However, ITS generated data includes data that has been collected by planners, operators, and researchers on a traditional basis for years. Benefits generated from ITS have already been accounted for in many large cities in the US. For example, the use of variable message signs (VMSs) to alert drivers of accidents ahead have provided drivers enough time for them to make decisions about alternate routes. Another example is the adjustment of ramp meter timing based on freeway flow conditions to regulate the amount of traffic entering a freeway. However, additional analysis of the archived data is needed would provide additional benefits for highways users. Furthermore, additional research of the ITS generated data will provide transportation professionals with the ability to make better decisions. For the purpose of this document, "ITS data" will refer to data generated using various ITS interlinked systems.

The following Sections give a summary of the previous work conducted using archived ITS data, an overview of the concept of ITS, a brief description of the

ITS interlinked systems and ITS data sources based on a metropolitan ITS system, and possible users of ITS data. In addition, some background information on the theory of traffic flow is included as a basis for the understanding of traffic flow in expressway facilities.

2.1 Literature Review

To understand the present state of knowledge, previous studies in the area of archived ITS generated data were reviewed. These include a study conducted using clustering and regression to identify outliers in weigh-in-motion data from Mn/ROAD data (Buchheit et al. 2002), the application of the KDD and data mining to large amounts of construction project data to identify the novel patterns in construction fields (Soibelman and Kim 2000), the application of data mining to analyze traffic incidents (Lee et al. 2004), the use of data mining and KDD to predict railroad demand (Carvalho 2007), the potential use of data mining in the construction industry (Nassar 2007), and the application of classification analysis to large sets of pavement condition data to predict the present serviceability rating (PSR) of pavements from the state of Missouri (Amado 2001). Many more examples of data mining applications within the broad field of civil engineering can be found.

A brief history of how the need to archive data became an important issue for the FHWA was provided to give way to the following Sections of this Chapter (Smith 2003).

- ✓ 1996 – the need to archived data was initially expressed at the Highway Performance Monitoring System (HPMS) Steering Committee.
- ✓ 1998 – the FHWA sponsored a meeting to discuss the uses of archived data.
- ✓ 1998 – the FHWA revised the National Architecture to include a new user service, Archived Data User Service (ADUS).
- ✓ 1999 – ADUS was officially added to the National Architecture.
- ✓ 2000 – the FHWA developed an ITS Data Archiving Five-Year Program Description to explain the need for a Federal Program that addresses the archiving and multi-agency use of data generated from ITS applications.

A few studies addressing different aspects of archived ITS data were found. They focused on ITS data sources, potential partnerships for the management of archived ITS data, and possible users of archived ITS data. These studies are summarized in Sections 2.1.1 through 2.1.3.

2.1.1 ITS Data Sources

There are several sources of ITS data, such as MPOs, DOTs, TMCs, and the highway patrol. Although there are other sources of ITS generated data, for the purpose of this work the discussion only includes the previously mentioned sources.

Metropolitan Planning Organizations (MPOs) – maintain and manage a data archive for their own use and the use of other agencies in the region. These organizations often perform quality control measures; provide access to the data (e.g. by internet or compact disc (CDs) by request); provide information and

documentation on the data; and provide software applications to help analyze the data and/or provide data formats that allow easy data analysis by other software (Dahlgren et al. 2002). These organizations are concerned with identifying multimodal passenger transportation improvements (long- and short-range), congestion management, performing air quality planning, and developing and maintaining forecasting and simulation models. Some of the applications these organizations have regarding ITS generated data include (Margiotta 1998):

- ✓ Congestion monitoring
- ✓ Link speeds for TDF and air quality models
- ✓ Demand estimation
- ✓ Temporal traffic distributions
- ✓ Truck travel estimation by time of day
- ✓ Macroscopic traffic simulation
- ✓ Parking utilization and facility planning
- ✓ High occupancy vehicle (HOV), paratransit, and multimodal demand estimation
- ✓ Congestion pricing policy

These organizations usually collect traffic volume and travel time data.

Departments of Transportation (DOTs) – design, construct, maintain, and operate the state highway systems, including the Interstate highway system within the state. Generally, DOTs work together with TMCs and/or traffic operation centers (TOCs) to share information obtained from the surveillance cameras located in various highways and intersections within the state.

However, the most common practice of collecting traffic data is by means of counting stations. Currently, DOTs are providing much more information than they used to through the Internet. Many DOTs maintain a user friendly website for anyone to access. Among the information provided are: real-time video of various intersections of highways and intersections; real-time information of incidents; dynamic maps with current construction projects and/or incidents; information about future construction projects; and route study polls.

Traffic Management Centers (TMCs) – maintain and manage day-to-day operations of deployed ITS in its geographic area of responsibility. TMCs depend in part on automated systems to accomplish their goals. These systems monitor transportation resources, provide control from the TMC and distribute transportation information. Pearce (1999) states that TMC systems are typically separated into two categories: those found within the TMC and those found outside, referred to as “field equipment” or vehicle systems (Pearce 1999). The latter are the equipment used to collect the data. Some of the applications these centers have regarding ITS generated data are (Margiotta 1998):

- ✓ Pre-planned control strategies (ramp metering and signal timing)
- ✓ Highway capacity analysis
- ✓ Saturation flow rate determination
- ✓ Microscopic traffic simulation (historical, short-term prediction of traffic conditions)
- ✓ Dynamic traffic assignment
- ✓ Incident management

✓ Congestion pricing operations

An excellent example of a TMC is the San Antonio TransGuide® in Texas (TransGuide® 2002). San Antonio provides a user-friendly website available to anyone. It has real-time data for traffic conditions and travel times for most of the highway system in the San Antonio area. In addition, data is not only available to see on the web but also to download. Data is provided in unzipped folders. For more information, access the TransGuide® website provided in the reference Section of this document.

Highway Patrols – ensure safety and provide service to the public as they utilize the highway transportation system and assist local governments during emergencies when requested. The highway patrol collects large amounts of information that are useful for transportation professionals. Information from traffic incidents is especially interesting for transportation managers to try to improve the system in order to avoid or reduce incidents. A typical set of information collected by highway patrol officials when handling a traffic incident include variables like: identification number, time, type, location, and area. In some cases, such as the California Highway Patrol (California Highway Patrol 2002), the information is put on the web for users to access and make decisions on detour routes. In addition, a zoom-in map of locations is provided, allowing easy access to everyone (e.g., drivers, planners, operators, researchers).

Generally, ITS generated data are the same type of data that have been traditionally collected by planners, operators, and researchers but are much more detailed in their temporal and spatial coverage (Margiotta 1998). Table 2.1 lists

some specific ITS generated data sources. Even though it is not a complete list, it provides an idea of how much data is generated by various sources and, thus, how much data is available for research.

Table 2.1 ITS generated data sources (Adapted from Margiotta 1998)

ITS Data Source	Primary Data Elements	Typical Collection Equipment	Spatial Coverage	Temporal Coverage
Freeway traffic flow surveillance data	Volume, Speed, Occupancy	Loop detectors, Video imaging, Acoustic, Radar/microwave	Usually spaced at ≤ 1 mile; by lane	Sensors report at 20-60 second intervals
	Vehicle classification, vehicle weight	Loop detectors, WIM equipment, video imaging, acoustic	Usually 50-100 per state; by lane	Usually hourly
Ramp meter and traffic signal preemptions	Time of preemption, location	Field controllers	At traffic control devices only	Usually full-time
Ramp meter and traffic signal cycle lengths	Begin time, end time, location, cycle length	Field controllers	At traffic control devices only	Usually full-time
Visual and video surveillance data	Time, location, queue length, vehicle trajectories, vehicle occupancy	CCTV, aerial videos, image processing technology	Selected locations	Usually full-time
Vehicle counts from electronic toll collection	Time, location, vehicle counts	Electronic toll collections equipment	At instrumented toll lanes	Usually full-time
TMC generated traffic flow metrics (forecasted or transformed data)	Link congestion indices, stops/delays estimates	TMC software	Selected roadway segments	Hours of TMC operation
Arterial traffic flow surveillance data	Volume, speed, occupancy	Loop detectors, video imaging, acoustic, radar/microwave	Usually midblock at selected locations only (system detectors)	Sensors report at 20-60 second intervals
Traffic signal phasing and offsets	Begin time, end time, location, up/down-stream offsets	Field controllers	At traffic control devices only	Usually full-time
Parking management	Time, lot location, available spaces	Field controllers	Selected parking facilities	Usually day time or special events
Transit usage	Vehicle boarding (by time and location), station O/D, paratransit O/D	Electronic fare payment systems	Transit routes	Usually full-time
Transit route deviations and advisories	Route number, time of advisory, route segments taken	TMC software	Transit routes	Usually full-time
Rideshare requests	Time of day, O/D	CAD	Usually areawide	Day time, usually peak periods
Incident logs	Location; begin, notification, dispatch, arrive, clear, depart	CAD, computer-driven logs	Extent of incident management program	Extent of incident management

ITS Data Source	Primary Data Elements	Typical Collection Equipment	Spatial Coverage	Temporal Coverage
	times; type; extent (blockage); HazMat; police accident reports ref.; cause			program
Train arrivals at highway rail intersections	Location, begin time, end time	Field controllers	At instrumented HRIs	Usually full-time
Emergency vehicle dispatch records	Time; O/D; route; notification, arrive, scene, leave times	CAD	Usually areawide	Usually full-time
Emergency vehicle locations	Vehicle type, time, location, response type	AVI or GPS equipment	Usually areawide	Usually full-time
Construction and work zone identification (ID)	Location, date, time, lanes/shoulders blocked	TMC software		
HazMat cargo identifiers	Type, container/package, route, time	CVO systems	At reader and sensor locations	Usually full-time
Fleet activity reports	Carrier, citations, accidents, inspection results	CVO inspections	N/A	Usually summarized annually
Cargo identification	Cargo type, O/D	CVO systems	At reader and sensor locations	Usually full-time
Border crossings	Counts by vehicle type, cargo type, O/D	CVO systems	At reader and sensor locations	Usually full-time
On-board safety data	Vehicle type, cumulative mileage, driver log (hrs. of service), subsystem status (e.g., breakers)	CVO systems	At reader and sensor locations	Usually full-time
Emissions management system	Time, location, pollutant concentrations, wind conditions	Specialized sensors	Sensor locations	Usually full-time
Weather data	Location, time, precipitation, temperature, wind conditions	Environmental sensors	At sensor locations	Usually full-time
Probe data	Vehicle ID, segment location, travel time	Probe readers and vehicle tags, GPS on vehicles	GPS is areawide; readers restricted to highway locations	Usually full-time
VMS messages	VMS location, time of message, message content	TMC software	VMS locations	Hours of TMC operation
TMC and information service provider generated route guidance	Time/date, O/D, route segment, estimated travel time	TMC/information service provider software	Usually areawide	Hours of TMC operation
Parking and roadway pricing changes	Time/day, route segment/lot ID, new price	TMC software	Facilities subject to variable pricing	Hours of TMC operation

2.1.2 Agencies that Archive ITS Data

Albert (1999) provides a list of agencies that archive ITS data (presented here in Tables 2.2 and 2.3). The TTI and The Texas A&M University System listed four additional agencies: the Montgomery County DOT in Maryland, NORPASS Kentucky Transportation Center, CALTRANS in California mentioned in the final report on “ITS Data Archiving” (TTI and The Texas A&M University System 1999), and the Kentucky Transportation Cabinet (KTC) (Turner 2002). Smith (2003) lists Dallas, TX as archiving few types of data on an irregular basis and the TransPort Program in Oregon as archiving specific loop data independent of the Oregon DOT for the purpose of proving the utility of archiving data for future applications. Four additional agencies (Virginia DOT, Maryland DOT, TransVISION in Fort Worth, TX, and Arizona DOT) were listed by Smith (2003) as collecting data. Table 2.2 shows loop detector data that most of these agencies archive. Information about the amount of centerline miles that the system covers, the approximate spacing between consecutive detectors, the level of aggregation, and the media to which the data are aggregated are also included (See Table 2.2). Table 2.3 summarizes of the same type of information for the automatic vehicle identification (AVI) systems.

Table 2.2 Summary of traffic management center loop detector data archiving practices (Adapted from Albert 1999, Smith 2003, and Dahlgren et al. 2001)

TMC	Coverage	Spacing	Archival/Availability
San Antonio TransGuide®	55 mi	0.5 mi	20-second data available on internet
Houston TranStar®	30 mi	0.5 mi	None
Minneapolis TMC	175 mi	0.5 mi	5-min data available on CD
Phoenix TOC	41.5 mi	0.33 mi	20-sec data saved. 5-min data easily accessed
Michigan ITS	32 mi (1) 150 mi (2)	0.33 mi (1) 2 mi (2)	1-min data saved to tape. One week available online.
Illinois TSC	130 mi	3 mi	1-hr data saved to tape
Los Angeles TMC	748 mi	0.5 mi	30-sec data saved to tape
North Seattle ATMS	100 mi	0.5 mi	5-min data saved to CD
Toronto Compass	22 mi	0.5 mi	5-min data saved to CD since 1997
INFORM	35 mi	0.5 mi	15-min data saved for 3 months
TransVISION	39 mi		Data will be provided upon request in the future
Virginia DOT	47 mi		2-min detector and station data, and incident data saved in an Oracle database
Maryland DOT	≈ 200 intersections		5-min saved in an Oracle server
Arizona DOT	75 mi	0.33 mi	20-sec stored on-line in compressed text formats, old data in stored on CDs
TransPort	11 mi	1.1 mi	20-sec data (Oct. 30 – Nov. 3, 2000 as an experiment)

Table 2.3 Summary of traffic management center AVI data archiving practices (Adapted from Albert 1999)

TMC	Coverage	Spacing	Archival/Availability
San Antonio TransGuide®	97.5 mi	1.0 mi – 2.0 2.0 mi	24-hour data saved
Houston TranStar®	227 mi	0.9 mi – 6.7 mi	15-min data saved
TRANSCOM		1.5 mi	15-min data saved

2.1.3 Possible Users of Archived ITS Data

Researchers and planners have been the primary users of archived ITS generated data. However, Albert (1999) provided a list of other potential users of ITS data. Those potential users are: academic institutions, transportation agencies, information service providers, insurance companies, wireless telephone providers, and private consulting companies (Albert 1999).

The common practice for obtaining the data is to acquire the information from TMCs in a raw form and archive the data based on the user's needs (Albert 1999). Currently, there are several methods to archive ITS generated data: on-line in compressed text formats, CDs, Oracle servers (or relational databases), and magnetic tape cartridges. Some agencies store the data for a period of time (e.g., 6 months to 1 year) and then destroy it (Dahlgren et al. 2001). ITS data are also archived using electronic spreadsheets and paper, however these are becoming less common with the accessibility users have to computers.

Some of the uses of archived ITS generated data are: monitoring of system performance, traffic impact assessment, improved travel demand models, prediction of future traffic conditions in the short-term, improvement of wireless telephone networks, incident information, and congestion information (Albert 1999). Some examples of these applications include the following:

Monitoring of System Performance – using detector data to study daily trends in a transportation system.

Traffic Impact Assessment – using data collected for short periods of time to assess the effects that land development has had on traffic and parking patterns.

These studies are done comparing traffic conditions before and after the land was developed.

Prediction of Future Traffic Conditions in the Short-Term – using the real-time data from advanced traveler information systems (ATIS) to develop short-term predictions of future traffic conditions based on historical traffic information.

Improvement of Wireless Telephone Networks – using ITS technology to track wireless telephone users with the purpose of following their travel patterns. These studies allow wireless telephone companies to expand their networks where it supplies the users' needs.

Incident Information – using incident information archived by TMCs, automobile insurance companies can acquire accident information.

Congestion Information – using closed circuit television to provide users with immediate information of current congestion and to study patterns of congestion in specific sites.

Margiotta (1998) conducted a survey in which ITS generated data stakeholders were asked about current ITS data availability and possible applications of each type of ITS data. As part of the study, traffic management operators stated that predictive traffic flow algorithms could be developed using ITS generated data. Using a data mining technique, such as the Bayesian approach, a prediction of traffic conditions 15 minutes into the future could be achieved, for example. However, the author stated that, at the time of the study, the ITS data needed to conduct this type of research was extremely limited (Margiotta 1998). Another group of stakeholders surveyed by Margiotta (1998)

were transportation researchers. This group had interests similar to traffic management operators, in that ITS generated data could be used to develop models. The two models proposed by transportation researchers were travel behavior and traffic flow models. This group of stakeholders stated that travel behavior models could be developed by measuring traveler response to system conditions using system detectors, probe vehicles, or monitoring in-vehicle and personal device used. For traffic flow models, transportation researchers stated that roadway surveillance data could provide continuous volume counts, densities, truck percents, and speeds at very small time increments, and global positioning systems (GPS) instrumented vehicles could provide second-by-second performance characteristics for microscopic model development and validation (Margiotta 1998).

2.2 Intelligent Transportation Interlinked Systems

ITS consist of a wide range of tools for managing transportation networks, as well as services for travelers (2000). ITS tools are based on three main components: information, communications, and integration. The purpose of ITS is to collect, process, integrate, and supply information for state agencies to obtain information on system conditions and choices to the public.

ITS technologies can be divided into four functional areas: metropolitan ITS, rural ITS, intelligent vehicle initiatives (IVI), and commercial vehicle operations (CVO). Each of the areas has a set of interlinked systems that are composed of equipment and software (BTS 2000). For this research we will work with

metropolitan ITS, given that our interest is to investigate the quality of traffic flow in expressway facilities on urban metropolitan areas.

Metropolitan ITS' infrastructure includes nine major components: arterial management systems, freeway management systems, transit management systems, incident management systems, emergency management, electronic toll collection (ETC), electronic fare payment, highway-rail intersections, and regional multimodal traveler information (BTS 2000). Table 2.4 shows the functions and benefits of the metropolitan ITS infrastructure.

Table 2.4 Metropolitan transportation systems infrastructure, functions, and benefits (Adapted from USDOT 2000)

Infrastructure	Functions	Benefits
Arterial Management	Monitor arterial network traffic, Implemented range of adaptive control strategies, Manage area-wide signal coordination	Safety, decreased travel times, increased capacity, fuel savings/lower emissions, customer satisfaction
Freeway Management	Monitor freeway conditions, Identify flow impediments, Control ramp metering and lane control, Central highway advisory radios	Safety, decreased travel times, increased capacity
Incident Management	Incident detection, Incident response/clearance	Safety, decreased travel times, fuel savings/lower emissions
Transit Management	Monitor transit vehicle position, Disseminate real-time schedules, Provide computer-aided dispatch, Provide vehicle condition monitoring	Safety, decreased travel times, lower costs, customer satisfaction
Electronic Fare Payment	Provide payment at station/stop or in-vehicle	Decreased travel times, customer satisfaction
Electronic Toll Collection	Provide payment at toll collection stop	Decreased travel times, increased capacity, lower costs
Emergency Management	Monitor vehicle location, Provide fleet management support	Decreased travel times, customer satisfaction
Highway-Rail Crossing Management	Provide remote monitoring of highway-rail intersections	Safety
Regional Multimodal Traveler Information	Provide information distribution on weather conditions	Lower costs, customer satisfaction, fuel savings/lower emissions

ITS generated data usually provide information similar to that traditionally used in transportation planning, operations, administration, and research.

Traditionally, data was collected by loop detector and pneumatic tube counter stations. Today's ITS data is voluminous in quantity and temporal coverage due to new technology. There are several ITS interlinked systems that are being used to collect data. Some of these ITS interlinked systems are:

Loop Detectors – used to collect traffic volume, lane occupancy, and average speed data. These stations are permanent, thus allowing larger quantities of information to be collected. A local controller unit (LCU) provides the storage and aggregates the loop detector information in the field. Information from the LCUs is later retrieved by the TMCs. These detectors consist of a buried wire placed below the surface of the pavement. An electric current is sent through the loop by attaching it to a power source, creating an electromagnetic field. Vehicles are detected when they move into this field. Loop detectors are located in either single-loop or double-loop configurations. Single loop detectors acquire vehicle volume and lane occupancy information. Double-loop detectors can be used to obtain additional information, such as spot speeds, using both loops and calculating the difference in arrival times between consecutive loops (Albert 1999). A dataset obtained from loop detectors would have variables such as: date, time, location, speed, volume, vehicle length, and lane occupancy.

Automatic Vehicle Identification (AVI) – used to collect real-time traffic information. This system consists of four primary components: probe vehicles with electronic transponders; roadside antennas to detect the transponders; roadside readers to collect data; and a central computer facility to collect and interpret data (Albert 1999; TranStar[®] 2002). It is most useful for direct

calculation of travel times with a reported accuracy of 99 percent (Albert 1999). Some of its disadvantages are: data collection is limited to the probe density, thus it cannot be used for volume counts; the capital, installation, and maintenance costs are very high; and the fact that each AVI transponder has a unique identification number raises concerns about the personal privacy of motorists. A dataset obtained from an AVI system would have variables like: day, station, time, identification, and travel time (Albert 1999).

Closed Circuit Television – used for traffic monitoring. These systems give an immediate, comprehensive picture of traffic conditions to users through the Internet. They are useful for planning and managing incident response if the incident is within the camera's view (Dahlgren et al. 2002).

2.3 Measuring Traffic Flow on Expressway Facilities

Transportation service is measured in terms of the highway's ability to accommodate vehicle traffic safely and with some acceptable level of performance (e.g., providing acceptable vehicle speeds). The analysis of vehicle traffic provides the basis for measuring the operating performance of highways. However, in order to do that, one needs to consider the various dimensions of traffic, such as the number of vehicles per unit time, vehicle types, vehicle speeds, and the variation in traffic volumes over time (Mannering and Kilareski 1998).

Traffic flow, speed, and density are the foundation of any traffic analysis.

Traffic flow can be obtained from Eq. 1:

$$q = \frac{n}{t} \quad \text{Eq. (1)}$$

where:

q = traffic flow,

n = number of vehicles,

t = interval of time vehicles pass on some designated highway point.

Traffic flow, density, and speed are related as shown in Eq. 2:

$$q = k * u \quad \text{Eq. (2)}$$

where:

q = traffic flow,

k = traffic density,

u = speed

The time headways are related to t in Eq. 1 by:

$$t = \sum_{i=1}^n h_i \quad \text{Eq. (3)}$$

where:

h_i = time headway of the i th vehicle.

Thus, by substituting Eq. 3 into Eq. 1:

$$q = \frac{n}{\sum_{i=1}^n h_i} \quad \text{Eq. (4)}$$

and reorganizing the terms:

$$q = \frac{1}{\bar{h}} \quad \text{Eq. (5)}$$

where:

$$\bar{h} = \text{average headway } (\Sigma h_i/n).$$

For the analysis of traffic flow, average traffic speed is defined as space-mean speed. The space-mean speed is obtained using Eq. 6:

$$u = \frac{\left(\frac{1}{n}\right) \sum_{i=1}^n l_i}{\bar{t}} \quad \text{Eq. (6)}$$

where:

u = space-mean speed,

l = known length of highway,

l_i = length of highway used for the speed measurement of vehicle i .

$$\bar{t} = \frac{1}{n} [t_1(l_1) + t_2(l_2) + \dots + t_n(l_n)] \quad \text{Eq. (7)}$$

where:

$t_n(l_n)$ = time necessary for vehicle n to traverse a section of highway of length l .

If all vehicle speeds are measured over the same length of highway ($L = l_1 = l_2 = \dots l_n$), then

$$u = \frac{1}{\left(\frac{1}{n}\right) \sum_{i=1}^n \left[\frac{1}{\left(\frac{L}{t_i}\right)} \right]} \quad \text{Eq. (8)}$$

which is the harmonic mean of speed used in traffic models.

Traffic density is the number of vehicles occupying some length of highway at some specified time. It is obtained using Eq. 9:

$$k = \frac{n}{l} \quad \text{Eq. (9)}$$

where:

k = traffic density,

n = number of vehicles,

l = known length of highway.

When modeling traffic flow, there is considerable analytic value in modeling the vehicle time headways (Mannering and Kilareski 1998). The most simplistic approach to modeling traffic flow would be to assume that all vehicles are spaced equally. However, observations show that such uniformity of traffic flow is not realistic in the majority of the cases. Therefore, a non-deterministic approach is needed for the representation of vehicle arrivals.

If we assumed that vehicles arrive randomly, we could use the Poisson distribution. Eq. 10 expresses the Poisson distribution:

$$P(n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \quad \text{Eq.(10)}$$

(Poch and Mannering 1996)

where:

t = duration of the time interval over which vehicles are counted,

P(n) = probability of having n vehicles arrive in time t ,

λ = average vehicle flow or arrival rate (in vehicle per unit time).

However, the assumption of Poisson-distributed traffic arrivals is most realistic in low volume traffic conditions only. Mannering and Kilareski (1998) suggest that other distributions be used when the traffic flow becomes heavily congested or when traffic signals cause cyclical traffic stream disturbances (Mannering and Kilareski 1998; Poch and Mannering 1996).

According to the Highway Capacity Manual (HCM), there are three performance measures that characterize a freeway segment. These are: density in terms of passenger cars per mile per lane, speed in terms of mean passenger-car speed, and volume to capacity (v/c) ratio. Each of these provides an indication of the traffic flow of a given freeway segment. Since density is directly related to speed and flow, density is the measure used to provide the six LOS A through F (HCM 2000; Khisty 1990). The HCM defines the six levels of service for freeways/expressways as (HCM 2000):

LOS A – represents free flow conditions. Vehicles are almost unaffected by the presence of others in the traffic stream. The effects of incidents do not affect the stream at this level.

LOS B – represents reasonably free flow. Vehicles are able to maneuver with only slight restriction. The effects of incidents still do not affect the stream at this level.

LOS C – represents flows at or near the free flow speeds (FFS) of the freeway (Table 2.6). Vehicles are not able to maneuver without experiencing noticeable restriction. The effects of small incidents are absorbed by this level,

however the service will experience local deterioration and queues may be expected.

LOS D – represents the reduction of speeds and the quick increase in density. The freedom to maneuver becomes more restricted and drivers experience reductions in physical and psychological comfort. Incidents can generate lengthy queues due to the higher density associated with this level. Traffic flow will be affected with minor disruption in traffic.

LOS E – represents operating conditions at or near the roadway's capacity. Even minor disruptions to the traffic stream, such as vehicles entering from a ramp or changing lanes, can cause delays as other vehicles give way to allow such maneuvers. Maneuverability in general is extremely limited and drivers experience considerable physical and psychological discomfort.

LOS F – represents a breakdown in vehicular flow. Queues form quickly behind points in the roadway where the arrival flow rate temporarily exceeds the departure rate, as determined by the roadway's capacity. Such points occur at minor incidents and on- and off-ramps where incoming traffic results in capacity being exceeded. Vehicles often proceed at reasonable speeds and then are required to stop in a cyclic fashion.

The thresholds as provided in the latest version of the HCM for each LOS are illustrated on Table 2.5.

Table 2.5 HCM levels of service thresholds for a freeway segment (Adapted from HCM 2000)

Level of Service	Density Range (passenger car equivalents per mile per lane (pc/mi/ln))
A	0-11
B	> 11-18
C	>18-26
D	> 26-35
E	> 35-45
F	> 45

The density values in Table 2.5 illustrate the quality of service drivers can expect when driving on a freeway. A freeway or expressway is defined as a divided highway facility that has two or more lanes in each direction for the exclusive use of traffic, with full control of access and outlets (Khisty 1990). In the hierarchy of highways, freeways/expressways are the only facilities that provide completely uninterrupted flow.

In HCM analysis, the ideal freeway/expressway has some basic characteristics', including 12 feet minimum lane width and 6 feet minimum lateral clearance between the edge of travel lanes and the nearest obstacle on the side of the road or median. A common design speed is 70 mph, however this number varies according to the geographical region.

Table 2.6 illustrates the relationship between speed, flow, and density for basic freeway/expressway segments with various FFS (HCM 2000). LOS F does not appear in Table 6 since it represents congestion or failure of the facility.

Table 2.6 Relationship between speed, flow, and density (Adapted from HCM 2000)

Criteria	Level Of Service				
	A	B	C	D	E
FFS = 75 mi/h					
Maximum Density (pc/mi/ln)	11	18	26	35	45
Minimum Speed (mi/h)	75.0	74.8	70.6	62.2	53.3
Maximum v/c	0.34	0.56	0.76	0.90	1.00
Maximum Service Flow Rate (pch/h/ln)	820	1350	1830	2170	2400
FFS = 70 mi/h					
Maximum Density (pc/mi/ln)	11	18	26	35	45
Minimum Speed (mi/h)	70.0	70.0	68.2	61.5	53.3
Maximum v/c	0.32	0.53	0.74	0.90	1.00
Maximum Service Flow Rate (pch/h/ln)	770	1260	1770	2150	2400
FFS = 65 mi/h					
Maximum Density (pc/mi/ln)	11	18	26	35	45
Minimum Speed (mi/h)	65.0	65.0	64.6	59.7	52.2
Maximum v/c	0.30	0.50	0.71	0.89	1.00
Maximum Service Flow Rate (pch/h/ln)	710	1170	1680	2090	2350
FFS = 60 mi/h					
Maximum Density (pc/mi/ln)	11	18	26	35	45
Minimum Speed (mi/h)	60.0	60.0	60.0	57.6	51.1
Maximum v/c	0.29	0.47	0.68	0.88	1.00
Maximum Service Flow Rate (pch/h/ln)	660	1080	1560	2020	2300
FFS = 55 mi/h					
Maximum Density (pc/mi/ln)	11	18	26	35	45
Minimum Speed (mi/h)	55.0	55.0	55.0	54.7	50.0
Maximum v/c	0.27	0.44	0.64	0.85	1.00
Maximum Service Flow Rate (pch/h/ln)	600	990	1430	1910	2250

FFS is the free-flow speed of passenger cars measured during low to moderate flows. There are two ways in which the value for FFS can be obtained:

doing field measurements or estimating the values using guidelines provided in the HCM (HCM 2000). If field studies are to be conducted, the location should be one that provides low to moderate densities, that is flow rates up to 1300 pc/h/ln. Eq. 11 is used to estimate FFS (HCM 2000).

$$FFS = BFFS - f_{LW} - f_{LC} - f_N - f_{ID} \quad \text{Eq.(11)}$$

(Poch and Mannering 1996)

where:

FFS = free-flow speed (mi/h),

BFFS = base free-flow speed, 70 mi/h (urban) or 75 mi/h (rural),

f_{LW} = adjustment for lane width from Exhibit 23-4 (mi/h) (HCM 2000),

f_{LC} = adjustment for right-shoulder lateral clearance from Exhibit 23-5 (mi/h) (HCM 2000),

f_N = adjustment for number of lanes from Exhibit 23-6 (mi/h) (HCM 2000),

f_{ID} = adjustment for interchange density from Exhibit 23-7 (mi/h) (HCM 2000).

2.4 Selecting a Case Study Site

A case study site was selected based on the following criteria:

- ✓ Territorial size of the metropolitan area in comparison with its population (high density preferred);
- ✓ In place ITS system;
- ✓ Availability of archived data;

- ✓ Indication from the transportation agency that the data could be used;
- ✓ Cooperation from the transportation agency during the research work;
- ✓ Need for a study of this nature in the selected city/metropolitan area; and
- ✓ Recurrent congestion.

Table 2.7 illustrates the types of ITS generated data archived by the 78 largest metropolitan areas in the US and US territories. The size of each metropolitan area was determined by its population. The smallest metropolitan area on the list has a population of about 500,000. Even though in theory there are nine metropolitan ITS infrastructure systems (Table 2.4), in the survey conducted by the USDOT in 2000 these metropolitan areas only archived ITS generated data for transit and freeway management systems (Table 2.7). However, it is important to mention that many of the metropolitan areas did collect data for some of the other ITS infrastructure systems (USDOT 2002; Quiñones 2003).

After examining the population for the 78 metropolitan areas, it was noticed that the largest population belonged to the area of New York-Northern New Jersey-Southwest Connecticut (18,323,382) and the smallest population belonged to the area of Harrisburg-Lebanon-Carlise in Pennsylvania (509,081) (US Census 2008).

Table 2.7 Type of ITS generated data archived by the 78 largest metropolitan areas in the US

Metropolitan Area	State	Archived Data	
		Transit Management	Freeway Management
Birmingham	AL	None	None
Little Rock-North Little Rock	AR	Trip itinerary, passenger count	None
Phoenix-Mesa	AZ	Passenger count, vehicle time and location	None
Tucson	AZ	Vehicle monitoring status, vehicle time and location	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, road condition, weather conditions, incidents, route designation, current work zones, scheduled work zones, highway operations, emergency evacuation routes
Sacramento-Yolo	CA	Passenger information, passenger count, vehicle time and location	Traffic volumes, traffic speeds, lane occupancy, ramp queues, metering rates, weather conditions, incidents
San Francisco-Oakland-San Jose	CA	Incidents, passenger count, trip itinerary planning records, vehicle time and location, passenger information	Traffic volumes, vehicle classification, route destinations
Fresno	CA	Vehicle monitoring, vehicle time and location	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, metering rate, road conditions, weather conditions, incidents, current work zones, scheduled work zones, emergency evacuation routes, highway operations
Bakersfield	CA	None	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, metering rate, road conditions, weather conditions, incidents, current work zones, vehicle occupancy, scheduled work zones, emergency evacuation routes, highway operations
Los Angeles-Riverside-Orange County	CA	Passenger information, passenger count, trip itinerary planning records, incidents, trip operations	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, metering rate, incidents, scheduled work zones, emergency evacuation routes, road conditions, weather conditions
San Diego	CA	None	None
Denver-Boulder-Greeley	CO	None	None
Hartford	CT	Passenger count, passenger information, trip itinerary	Traffic speeds, weather conditions, incidents
New Haven	CT	Passenger count	Traffic speeds, weather conditions, incidents
Washington	DC	Road conditions, passenger information, vehicle time and location, incidents, trip itinerary, transit operations	None
Jacksonville	FL	Weather conditions, passenger info., passenger count, vehicle time and location	None
Orlando	FL	None	

Metropolitan Area	State	Archived Data	
		Transit Management	Freeway Management
			None
Tampa-St. Petersburg	FL	Incidents, passenger info., passenger count, transit operations, emergency evacuation routes	None
Miami-Ft. Lauderdale	FL	Incidents, vehicle monitoring status, vehicle time and location	Traffic volumes and scheduled work zones
West Palm Beach-Boca Raton	FL	None	None
Sarasota-Bradenton	FL	Passenger info., trip itinerary planning records, highway operations, emergency evacuation routes	State traffic counts
Atlanta	GA	Passenger info., passenger count	traffic volumes, traffic speeds, lane occupancy, vehicle classification, incidents, current work zones, scheduled work zones
Honolulu	HI	Passenger info., passenger count, vehicle time and location	None
Chicago-Gary-Kenosha	IL	Incidents	None
Indianapolis	IN	Incidents, passenger info., trip itinerary, passenger count, vehicle time and location	None
Wichita	KS	Incidents, weather conditions	None
Louisville	KY	None	Traffic volumes, traffic speeds, vehicle classification, incidents
Baton Rouge	LA	None	None
New Orleans	LA	Weather conditions, passenger count, vehicle time and location, emergency evacuation routes, current road work zones, route destination, passenger information, trip itinerary, transit operations	None
Springfield	MA	None	None
Boston-Worcester-Lawrence	MA	Incidents, passenger info., trip itinerary, passenger count, vehicle time and location, transit operations, scheduled road work zones	Scheduled work zones, current work zones
Baltimore	MD	Transit vehicle signal priority events, vehicle monitoring, passenger info., trip itinerary, passenger count, vehicle time and location, transit operations	None
Grand Rapids-Muskegon-Holland	MI	Passenger count, vehicle time and location	None
Detroit-Ann Harbor-Flint	MI	Vehicle monitoring status, passenger info., passenger count, vehicle time and location	None
Minneapolis-St. Paul	MN	None	Traffic volumes, lane occupancy, metering rate, incidents, highway operations, violation rate for HOV lanes, vehicle occupancy

Metropolitan Area	State	Archived Data	
		Transit Management	Freeway Management
Kansas City	MO	None	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, road conditions, weather conditions, incidents, current work zones, scheduled work zones, highway operations, route designation, emergency evacuation routes
St. Louis	MO	None	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, weather conditions, route designation, current work zones, scheduled work zones, emergency evacuation routes
Charlotte-Gastonia-Rock Hill	NC	None	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, incidents, current work zones, scheduled work zones
Greensboro-Winston Salem-High Point	NC	Trip itinerary, vehicle time and location	None
Raleigh-Durham-Chapel Hill	NC	Weather conditions, passenger info., passenger count, vehicle time and location, route designation	None
Omaha	NE	None	None
Philadelphia-Wilmington-Atlantic City	NJ	None	None
Albuquerque	NM	None	None
Las Vegas	NV	None	None
Buffalo-Niagara Falls	NY	None	Road conditions, weather conditions, incidents, current work zones, scheduled work zones, emergency evacuation routes, highway operations, traffic volumes, traffic speeds, lane occupancy, vehicle classification
Rochester	NY	None	Traffic volumes, vehicle classification, road conditions, weather conditions
Syracuse	NY	None	Vehicle classification, traffic speeds, incidents, traffic volumes, lane occupancy
Albany-Schenectady-Troy	NY	Incidents, vehicle monitoring, passenger info, passenger count	Traffic volumes, traffic speeds, incidents, current work zones, scheduled work zones, highway operations
New York-Northern New Jersey-Southwestern Conn	NY	Passenger count, transit operations, vehicle monitoring, passenger info, vehicle time and location, highway operations, emergency evacuation routes, scheduled road work zones, current road work zones	Traffic speeds, weather conditions, incidents, traffic volumes, route designations, current work zones, scheduled work zones, emergency evacuation routes, highway operations, road conditions, lane occupancy, vehicle classification, intermodal connections, vehicle occupancy
Cincinnati-Hamilton	OH	None	Traffic volumes, traffic speeds, lane occupancy, road conditions, weather conditions, incidents, current work zones, scheduled work zones
Dayton-Springfield	OH	Incidents	None
Columbus	OH	None	None

Metropolitan Area	State	Archived Data	
		Transit Management	Freeway Management
Cleveland-Akron	OH	Weather conditions, road conditions, passenger info, trip itinerary, passenger count	Traffic volumes, vehicle classification, weather conditions, violation rate for HOV lanes
Toledo	OH	None	None
Youngstown-Warren	OH	None	Traffic volumes, traffic speeds, vehicle location, road conditions, weather conditions, current work zones, scheduled work zones, intermodal connections
Oklahoma City	OK	None	Traffic volumes, traffic speeds, vehicle classification, incidents, current work zones
Tulsa	OK	None	Traffic volumes, traffic speeds, vehicle classification, incidents, current work zones
Portland-Salem	OR	None	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, vehicle location, metering rate, road conditions, weather conditions, incidents, current work zones, scheduled work zones, emergency evacuation routes, highway operations
Pittsburgh	PA	Passenger info., passenger count, vehicle time and location, weather conditions, road conditions, vehicle monitoring, highway operations, emergency evacuation routes, intermodal connections, scheduled road work zones, current road work zones, route designation	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, incidents, current work zones, scheduled work zones
Allenton-Bethlehem-Easton	PA	None	None
Harrisburg-Lebanon-Carlise	PA	Incidents, passenger count, emergency evacuation routes	None
Scranton-Wilkes Barre-Hazleton	PA	Incidents, road conditions	Incidents
San Juan	PR	None	Traffic counts, work zones, accidents
Providence-Fall River-Warwick	RI	None	Traffic volumes, route designation
Greenville-Spartanburg-Anderson	SC	None	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, road conditions, weather conditions, incidents, current work zones, scheduled work zones, emergency evacuation routes
Charleston-North Charleston	SC	None	Traffic volumes, traffic speeds, lane occupancy, road conditions, vehicle classification, weather conditions, incidents, current work zones, scheduled work zones, emergency evacuation routes
Nashville	TN	None	Traffic volumes, vehicle classification
Memphis	TN	None	Traffic volumes, vehicle classification, road conditions, weather conditions
Knoxville	TN	None	Traffic volumes, vehicle classification

Metropolitan Area	State	Archived Data	
		Transit Management	Freeway Management
El Paso	TX	None	None
Austin-San Marcos	TX	None	Traffic volumes, traffic speeds, lane occupancy, vehicle classification, road conditions, weather conditions, incidents, highway operations
San Antonio	TX	Passenger count, vehicle time and location, scheduled road work zones, current road work zones	None
Houston-Galveston-Brazoria	TX	None	Traffic volumes, traffic speeds, lane occupancy, incidents, current work zones, scheduled work zones
Dallas-Ft. Worth	TX	Passenger info., transit operations, highway operations, emergency evacuation routes, intermodel connections, scheduled road work zones, current road work zones, trip itinerary, passenger count, vehicle time and location	Incidents, current work zones, scheduled work zones
Salt Lake City-Ogden	UT	Incidents, weather conditions, passenger info., passenger count, vehicle time and location	Weather conditions, incidents, traffic volumes, traffic speeds, lane occupancy, vehicle classification, road conditions
Richmond-Petersburg	VA	None	Traffic volumes, lane occupancy, vehicle classification, emergency evacuation routes, highway operations
Hampton Roads	VA	None	Traffic volumes, traffic speeds, lane occupancy, road conditions, incidents, emergency evacuation routes
Seattle-Tacoma-Bremerton	WA	Trip itinerary, passenger count, highway operations, emergency evacuation routes, route destination, passenger info, vehicle time and location, transit operations, weather conditions	None
Milwaukee-Racine	WI	Incidents, weather conditions, vehicle time and location, scheduled road work zones, current road work zones, passenger info.	Traffic volumes, traffic speeds, lane occupancy, ramp queues, incidents, current work zones, scheduled work zones

Table 2.8 illustrates the three (3) metropolitan areas considered as case studies. It was important for the research to examine the availability of data from metropolitan areas that had high population densities to be used as case studies. The latter would allow the use of datasets that would provide unusual events, for example longer peak periods, more work zone operations, and worst LOS. From the information about the territorial area versus population, it is evident that

Phoenix-Mesa, San Juan, and Dallas-Ft. Worth have high densities with 3,617, 3,500, and 2,936 people per square mile, respectively.

The Phoenix-Mesa metropolitan area presented a disadvantage for our case study because the only ITS generated data archived is for transit management, and for this research the use of freeway/expressway management data was the primary interest. ITS generated data for the Dallas-Ft. Worth metropolitan area has been used extensively by several researchers, thus using it as a case study could probably generate duplicate information. On the other hand, even though the types of archived ITS generated data are limited, the San Juan metropolitan area provided a great opportunity for new much needed research. No research had been conducted using these types of data from the SJMR. Based upon the criteria developed for site selection, San Juan provided the best combination of desired characteristics.

Table 2.8 Metropolitan areas considered for case study

Population	Area (mi²)	Pop/mi²	Metro Area	State
3,251,876	899	3,617	Phoenix-Mesa	AZ
1,400,000	400	3,500	San Juan	PR
5,161,544	1,758	2,936	Dallas-Ft. Worth	TX

The Texas Department of Transportation (TxDOT) was contacted earlier in the project to find out the possibilities of obtaining data from their databases to be used in this research project. However, even though in email and telephone conversations the TxDOT seemed more than glad to provide the information for

the research, a letter was later received indicating that due to liability issues data from the TxDOT could not be used for our project.

The PRHTA was contacted various times to obtain information and traffic data about the most important expressways in the San Juan Metropolitan Region. Several personal visits to the agency were conducted to various departments with the intent of gathering as much information possible. In every visit to the agency, the people were glad to help and provided the information needed for the study. PR-18 was chosen above other expressways in the SJMR due to its length, 3.78 miles, and peculiar characteristics.

The PRHTA, like many state agencies, is currently developing the ITS infrastructure needed to sustain the amount of data being collected on a daily basis on some of the most important highways in the island.

Chapter 3 provides a detailed description of the case study, Las Américas Expressway (PR-18).

3. CASE STUDY – LAS AMÉRICAS EXPRESSWAY (PR-18) SAN JUAN METROPOLITAN REGION

Las Américas Expressway (PR-18) was built over 30 years ago to serve two functions: to provide a corridor between the Old San Juan and the suburbs and to serve as an expressway connector between the SJMR and PR-52. PR-52 is one of the most important expressways in Puerto Rico because it provides access to the south part of the island. PR-18 also connects with PR-22, which is another important expressway, being the only designated Interstate highway in Puerto Rico. PR-22 provides access to and from the SJMR to the west side of the island (Figure 1.1).

At the time of construction and during the construction of PR-18, Puerto Ricans relied on four lane highways (two lanes in each direction) to travel to and from the south and north, and north and west parts of the island. A one-way trip of 35 miles took about half a day to complete and, depending on the hour of the day, it could have taken even longer than that. These highways included many at grade signalized intersections and the crossing of every small downtown area along the way. Considering the size of Puerto Rico, which is 100 by 35 miles, and the fact that the majority of the work opportunities were and still are in the SJMR, the need for expressways was imperative.

Since its construction, PR-18 has carried high volumes of traffic at various hours of the day on every day of the week. The PRHTA has conducted a series

of improvements that have increased the capacity of PR-18. Given the rapid residential and commercial development in the SJMR, there is no room for expansion of this highway. The current traffic conditions reflect an average vehicle speed during the day (from 6:30AM to 6:30PM) of 23 mph. There are forced flow conditions during most of the day on PR-18.

Additional improvements are being made to the transportation system of SJMR, such as the “Tren Urbano”, which is a huge project consisting of a heavy metro rail that will connect the six major sectors (Bayamón, Guaynabo, Río Piedras, Carolina, Old San Juan, Hato Rey, and Caguas) within the SJMR (Pesquera and González 1996). With the construction and effective operation of “Tren Urbano,” the congestion in SJMR is expected to be reduced (Pesquera and González 1996). However it is well known that more people are expected to be driving vehicles in the years to come, thus new measures are needed to describe and improve the quality of flow.

Sections 3.1 and 3.2 describe the characteristics of this facility in greater detail and some of the problems or challenges presented by the PRHTA. A description of the data that were available for this facility at the time the project began is provided in Section 3.3.

3.1 Characteristics of Facility

PR-18 consists of 3.78 mi (6.09 km) of mostly concrete pavement. In this stretch of highway, there are 10 through lanes (2 of which are reversible in the

morning and afternoon peak hours on weekdays). Its functional classification is principal arterial with a maximum design speed of 65 mph (105 kph).

Table 3.1 describes some specific characteristics of the facility per highway segment of road (Felix 2003). These segments of road were determined using the main highways that PR-18 crosses as underpasses or overpasses. Figures 3.1 through 3.3 illustrate each of these underpasses or overpasses (Guía Urbana del Área Metropolitana 2000).

Table 3.1 Facility characteristics per highway segment for 2001

Highway Segments for PR-18					
Item	Mi 0.00 (Km 0.00) (PR-1) to Mi 0.81 (Km 1.31) (PR-21)	Mi 0.81 (Km 1.31) (PR-21) to Mi 1.53 (Km 2.46) (Ave. A. Miranda)	Mi 1.53 (Km 2.46) (Ave. A. Miranda) to Mi 2.09 (Km 3.36) (PR-17)	Mi 2.09 (Km 3.36) (PR-17) to Mi 2.88 (Km 4.64) (PR-23)	Mi 2.88 (Km 4.64) (PR-23) to Mi 3.78 (Km 6.09) (PR-22)
Median Type	Concrete Barrier	Concrete Barrier	Concrete Barrier	Concrete Barrier	Concrete Barrier
Median Width	20 ft (6.1 m)	24 ft (7.3 m)	24 ft (7.3 m)	20 ft (6.1 m)	21 ft (6.4 m)
Shoulder Type	Asphalt	Asphalt	Asphalt	Asphalt	Asphalt
Shoulder Width Right	11.8 ft (3.6 m)	10 ft (3.0 m)	11 ft (3.3 m)	9 ft (2.7 m)	9 ft (2.7 m)
Shoulder Width Left	3.94 ft (1.2 m)	3.94 ft (1.2 m)	0.0	6.89 ft (2.1 m)	0.0
Lane Width	13.12 ft (4 m)	13.12 ft (4 m)	13.12 ft (4 m)	13.12 ft (4 m)	13.12 ft (4 m)
Widening Feasibility	NO	NO	NO	NO	NO
Segment Length	0.81 mi (1.31 Km)	0.72 mi (1.15 Km)	0.56 mi (0.90 Km)	0.79 mi (1.28 Km)	0.90 mi (1.45 Km)
AADT	175,600	184,300	226,700	229,600	205,200
IRI	2.99	1.75	2.05	2.11	2.41
Peak Capacity	10,213	12,310	12,250	12,310	10,208

The PRHTA has designated PR-18 as an expressway with no widening feasibility due to the existing conditions of the highway itself and the urban development surrounding this facility (Table 3.1).

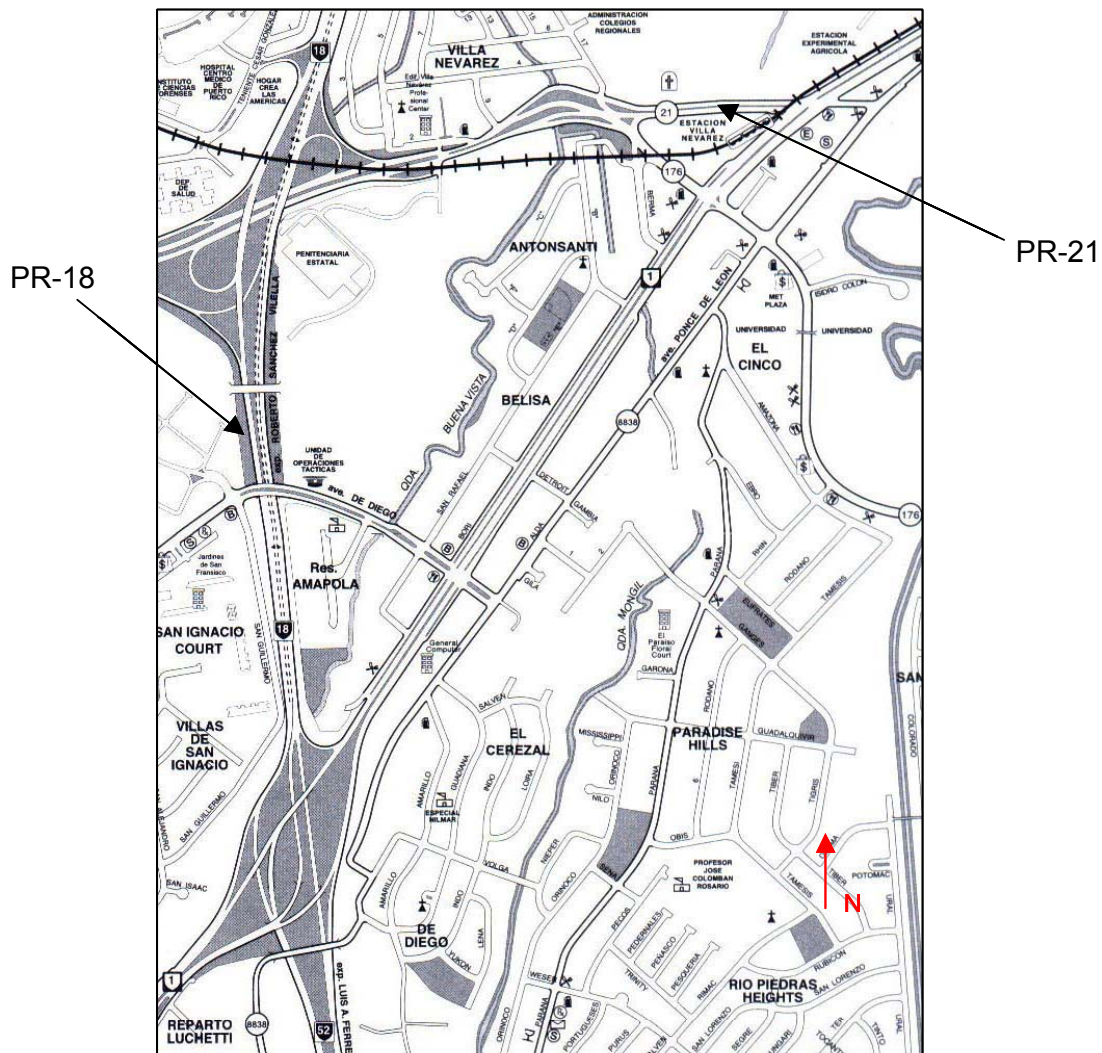


Figure 3.1 PR-18 overpassing PR-21

(Source: Guía Urbana del Área Metropolitana,
2000 San Juan Metropolitan Area Metro Data. CD-ROM)

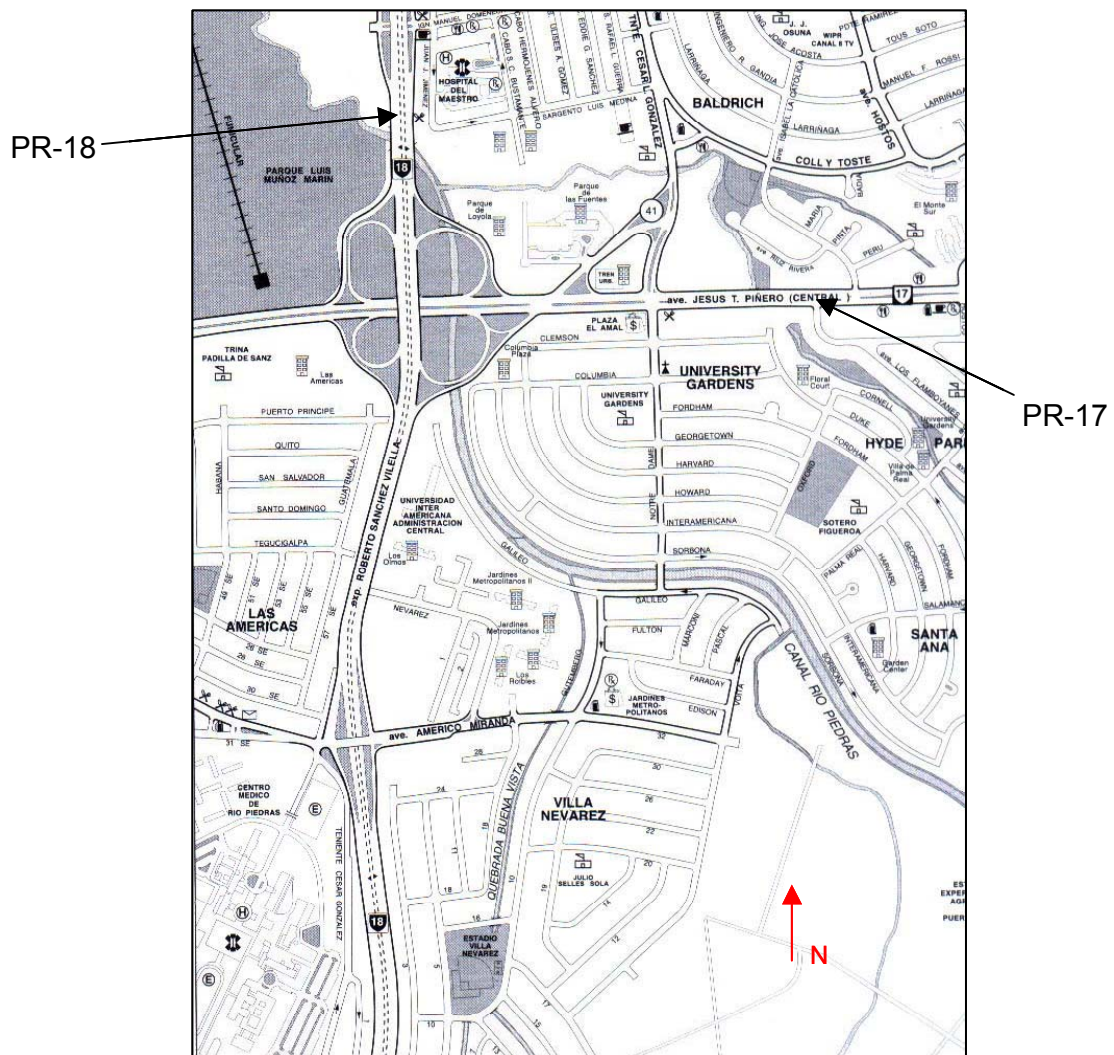


Figure 3.2 PR-18 underpassing ave. américo miranda and overpassing PR-17

(Source: Guía Urbana del Área Metropolitana,
2000 San Juan Metropolitan Area Metro Data. CD-ROM)

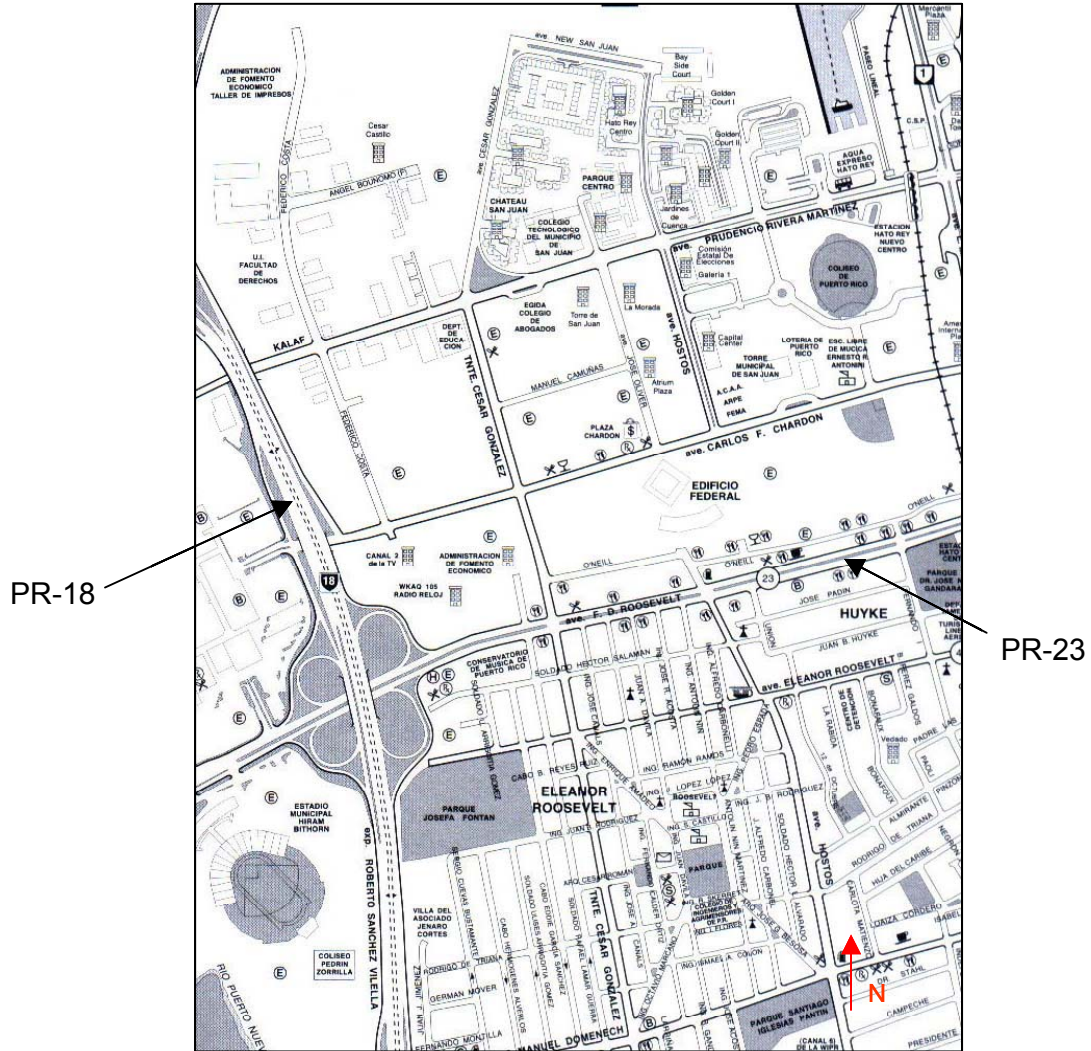


Figure 3.3 PR-18 overpassing PR-23
 (Source: Guía Urbana del Área Metropolitana,
 2000 San Juan Metropolitan Area Metro Data. CD-ROM)

3.2 Problems with Facility

There are several special conditions or problems with PR-18. Some of the most important ones are:

- ✓ There is no possibility for expanding the facility,

- ✓ The superelevation varies inconsistently throughout the facility and among lanes,
- ✓ During the redistribution of lanes on PR-18 between PR-17 and PR-23, several old manholes and drains were not relocated and are currently in odd places (in the middle of lanes and in-between lanes) making it difficult for drivers to maneuver among these,
- ✓ There is no HOV lane that could be used to improve the capacity,
- ✓ Adjacent collectors present huge problems of traffic congestion as well; studies conducted by J.M. Morales show that there is no route in the SJMR that could be used in case of an emergency evacuation (Morales 2003),
- ✓ PR-18 goes from four lanes to two lanes when entering PR-52 (Figure 3.1),
- ✓ During the morning and afternoon peak hours, right shoulders are “legally” used as additional lanes, and
- ✓ LOS F is overwhelmingly represented during most of the day (refer to Figure 3.4).

Figures 3.4 through 3.6 illustrate the capacity problems of this facility.

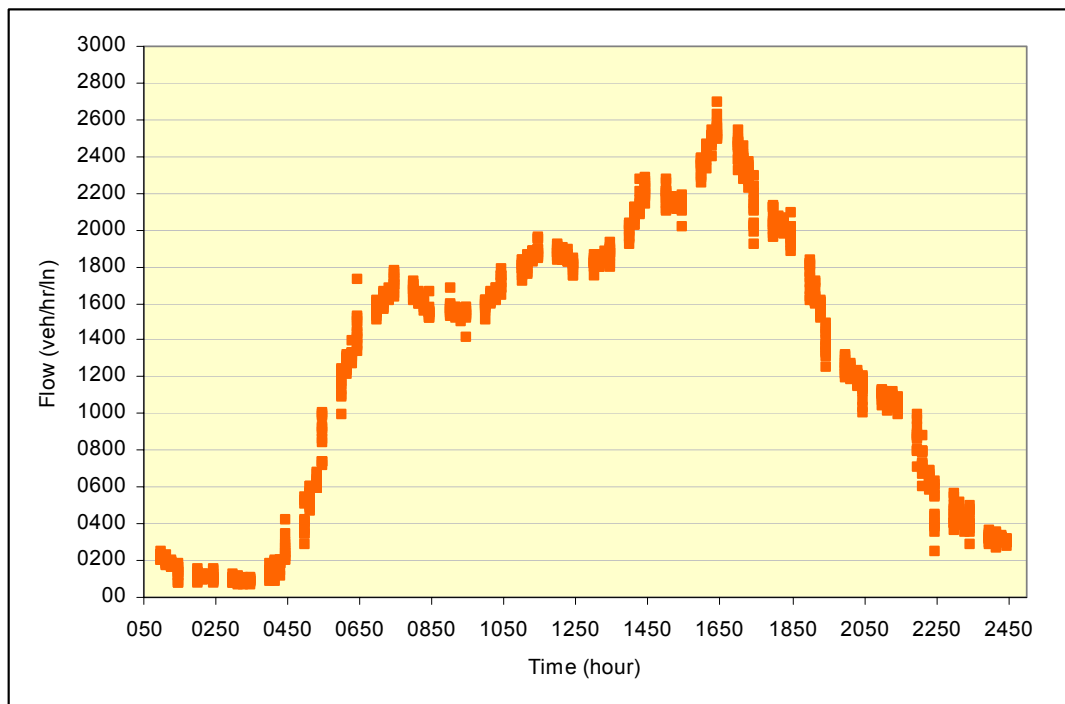
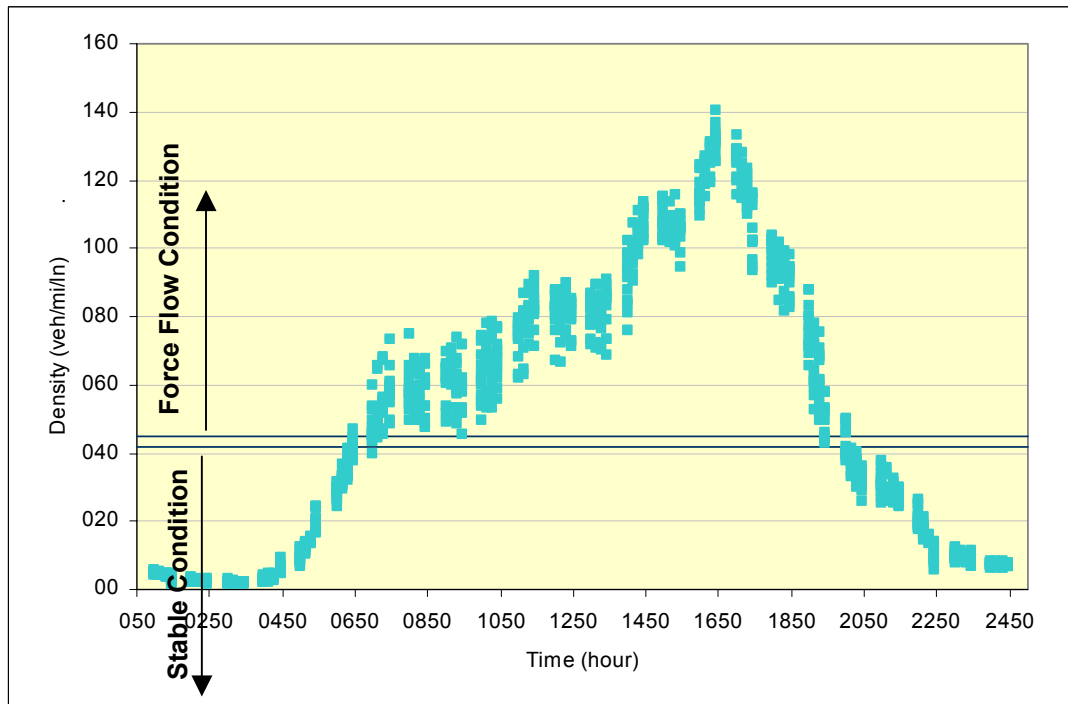


Figure 3.5 Flow versus time for a typical monday southbound

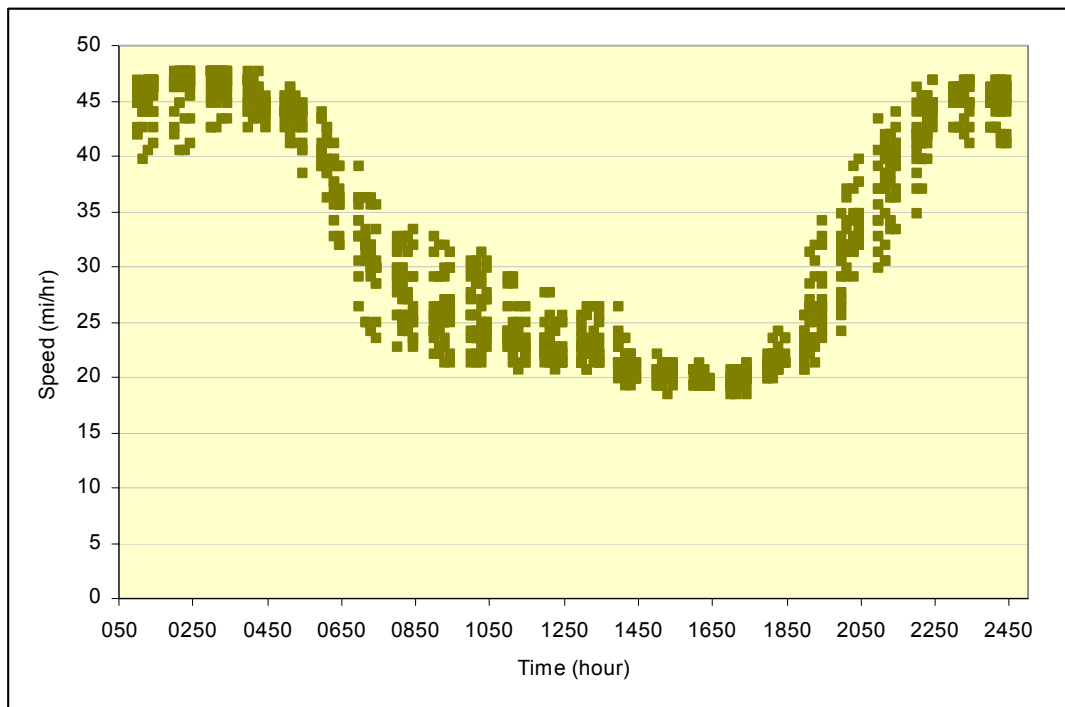


Figure 3.6 Speed versus time for a typical Monday southbound

The peak period begins around 5:45AM and ends around 6:45PM (Figures 3.4 through 3.6). High densities and high flows in Figures 3.4 and 3.5 correspond to lower speeds in Figure 3.6. The speed versus time graph indicates that the operational conditions were at forced flow conditions, LOS F. The approximate boundary between stable and forced flow conditions was indicated in Figure 3.4 to put emphasis on the large period of the day for which there was LOS F. The high peaks in the afternoon on both density and flow graphs seemed to be controlled by a downstream bottleneck.

3.3 Type of Data Available

Like many state agencies, the PRHTA is currently developing an ITS for the San Juan Metropolitan Region. Proposals for complete ITS are under

development for PR-22 and PR-18 (Pérez 2004). It has been a great challenge primarily because of the huge monetary investment required. However, several experimental projects have been undertaken and are succeeding in various sectors of the SJMR. One project is the coordination of the public transportation authority to provide continuous transportation (by means of ferries and buses) to and from provisional parking lots outside of the island of San Juan. A second project is lanes designated for emergency vehicles only going towards the area of Condado and Hato Rey, which are the closest areas that have hospitals, police departments, and fire departments. A third example is a temporary ITS management center that was prepared to give support to all of the emergency departments and the public transportation. The final example is the development of the first ez-pass program (called AVI) undertaken in Puerto Rico. The program was developed for the Teodoro Moscoso Bridge overpassing the San José Lagoon in Carolina, PR. The program is viewed by the PRHTA as a success and many people have used it since the beginning. Currently, other ez-pass programs have began to work on various toll plazas on PR-22 (an expressway facility that connects the East and West side of Puerto Rico on the North). The latter are still on its earliest stages, but are being well accepted by the public. Even though the small scale ITS programs that have been implemented so far have been a success, the process for the development of a large scale ITS for the SJMR has been a slow one. The major existing drawback for the development of an ITS of this size are the costs involved.

The data obtained for PR-18 for this project was traffic counts, work zones, accidents, and speed (Quiñones 2003; García 2003; Acevedo 2003). The data have been obtained from different divisions within the PRHTA. Most of the data have been obtained in hardcopy and through correspondence. However, the data for work zones was copied by hand during a visit to the PRHTA because of difficulties with the copy machine. It is worth mentioning that several traffic-consulting firms in Puerto Rico and Washington D.C. were contacted during the research to seek information about specific data for PR-18 (Cordova 2003; Morales 2003; Villalba 2004). No data was received from the consulting firms, although knowledge gained about studies conducted around this site was highly useful.

Chapter 4.0 provides the details of the research approach used during the project.

4. RESEARCH APPROACH AND METHODS

The approach used to develop the models consisted of the seven-step KDD process shown in Figure 4.1. The steps were: building the data mining database (refer to Section 4.1), examining and preparing the data (refer to Section 4.2), evaluating the data mining application (refer to Section 4.3), building the model (refer to Section 4.4), evaluating the model (refer to Chapter 5), understanding the information provided in the results (refer to Chapter 6), and providing conclusions and recommendations (Amado 2001). The main benefit of this process is that it encourages the user to follow each step and review prior steps if or when needed. The documentation of every step of the process is strongly recommended as it will help trace the path taken in case of the diversity of situations and possibilities and, also, will provide the documentation of the entire project.

The evaluation and interpretation of results of six models allowed the measurement of the quality of traffic flow in the stretch of highway that was analyzed. During the preparation of the data, various statistical tests were used to examine the data and to identify the amount of data that was going to be used. The statistical test results are found in Appendix B. The six models created consisted of various examples, which provided specific information about the patterns of traffic in the expressway facility studied.

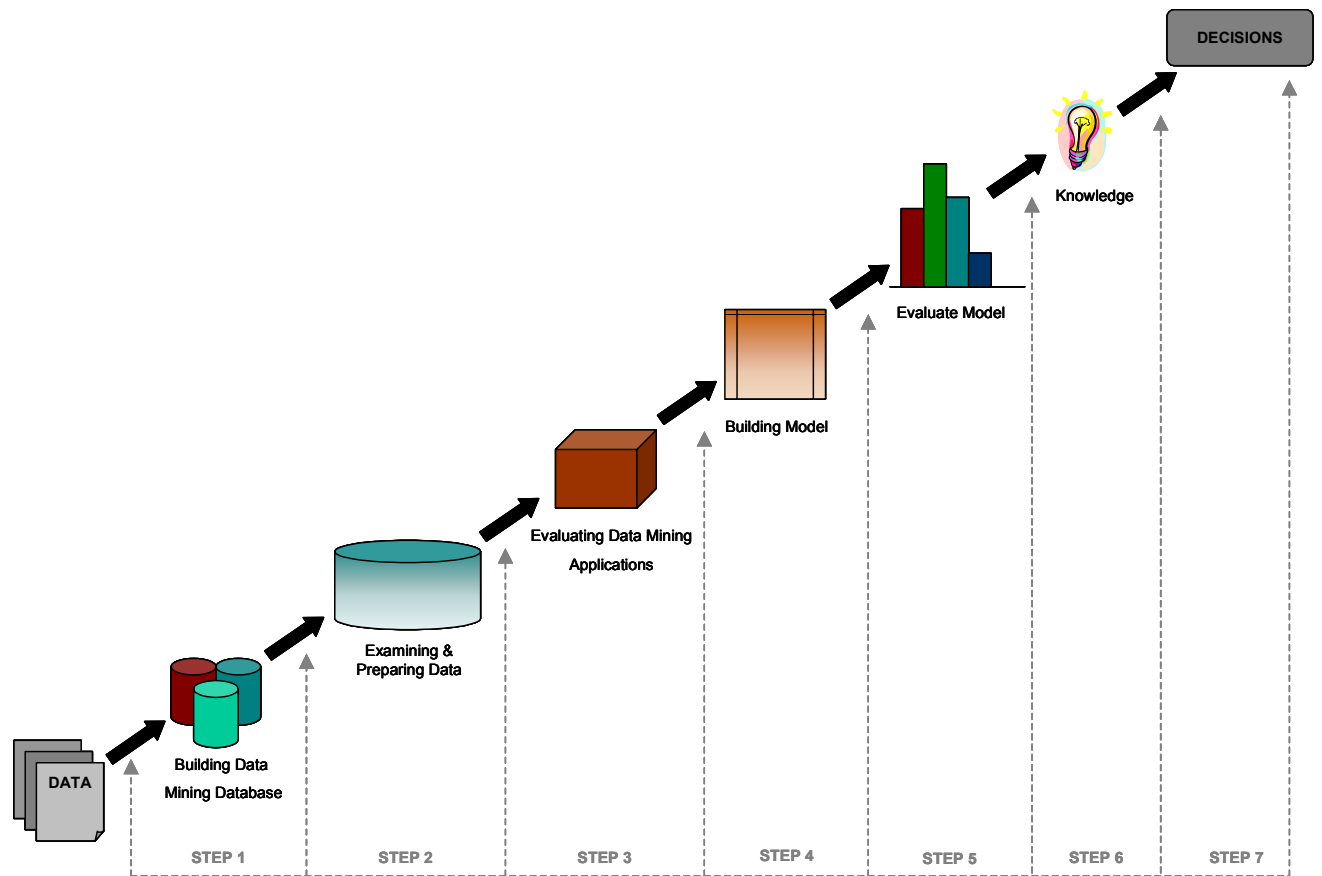


Figure 4.1 KDD seven-step process

Figure 4.2 illustrate the basic data processing procedure using the KDD process. Ideally, the process should start with ASCII-Text files. Preprocessing nonstandard data requires extra steps and time prior to processing the archived data. Automatically collected data are more likely to be stored in ASCII-text and spreadsheet files, yet that was not the case for most of the data used in this research. Given that the data obtained came from different sources and consisted of different formats, these could not be transformed into open database connectivity (ODBC) for a direct export to the IBM Intelligent Miner for Data.

during the analysis of results when a list in the form of conclusions is created. The latter will be the contributions made by the research project.

The following sections describe the details of steps 1 through 4 as applied to this research. The details involving step 5 are found in Chapter 5 entitled “Evaluation of the Model(s)” and the knowledge gained from the models or step 6 are found in Chapter 6, which are the conclusions and recommendations of the research. Given the nature of this document, being an educational research, there are no decisions (step 7) to be made.

4.1 Building the Data Mining Database

Building the data mining database was the first step in the KDD process which consisted on obtaining archived ITS generated data from the case study facility. Section 4.1.1 describes the details of the data that was used.

4.1.1 Archived Data

Historical data for this research were obtained from various divisions within the PRHTA and the National Climatic Data Center (NCDC). In Puerto Rico there is no TMC, thus the ITS generated data are collected, archived, and analyzed by the Office of Data Acquisition and Analysis of Transit within PRHTA. The data collected by the latter division is used in the HPMS study as required by the FHWA and the Transportation of Metropolitan Areas study, which is an internal study used in PRHTA (Burgos 2005). The data and information are also used for the design of projects coordinated by the Office of Environmental Studies and Pavement Management (Pérez 2004).

Data were collected at point locations using two types of automatic traffic counters, single and double inductance loops, and accumulated in roadside controllers. These field measurements were collected for each direction in PR-18. Quality control checks on data were performed once the data was taken to PRHTA offices.

Weather data are collected each hour in the San Juan Luis Muñoz Marín International Airport (SJU) station using Automated Surface Observing Systems (ASOS) equipment for measurements (McCown 2005). Hourly precipitation totals for the last 50 years is available with free access for “.edu” domain in either text version or ASCII comma-delimited files within the NCDC website (NCDC 2005).

4.1.1.1 File Formats

The format of each dataset was summarized as:

- ✓ Archived traffic data was obtained both in text forms as hard copies through US mail and in spreadsheets (.xls) via electronic communication.
- ✓ Historical weather data from the San Juan Metropolitan Area was obtained in delimited ASCII-text files through electronic communication.
- ✓ Historical work zones data was submitted in text forms as hard copies during a visit to the agency.
- ✓ Accident data was obtained in text forms as hard copies through US mail.

The inconsistencies in the data format created a huge disadvantage for the time frame of the first step of the KDD process (building the data mining database) for this research. Approximately, thirty percent (30%) of the effort

employed in conducting this research was used for the creation of the data mining database. The ideal situation would be to have all electronic files, in ASCII-text files preferably, and transform the datasets using an ODBC to be read by the IBM Intelligent Miner for Data. The latter would allow the user to invest about five percent (5%) of his or her time in this step of the process.

4.1.1.2 Data Elements

The level of detail of the archived traffic data was 15-minute time periods and by direction of traffic. Table 4.1 provides the detail of each dataset.

Table 4.1 Data Elements

Type of Data	Data Elements
Traffic	Time (hh:mm in 24-hour clock), date (mm/dd/yyyy), machine number, vehicle traffic volume count, average speed, station, station number, direction, route number, and municipality
Accident	Date, event, location, time, surface condition (wet, dry, or muddy), and type of accident (fatal, injured, or damage)
Work Zones	Year, project number, project description, and location
Weather	Date, hour, precipitation (in inches), station description, year, and month

4.1.1.3 Counter Location Information

Location information for the traffic counters in PR-18 was provided in text form as hard copies and includes:

- ✓ Station number
- ✓ Counter activation date
- ✓ Machine number
- ✓ Station description
- ✓ Number of through travel lanes
- ✓ Roadway name and designation

- ✓ Roadway milepost
- ✓ Roadway Direction

Figure 4.3 illustrates the location of the counters that collected the data used in the research. The counters were placed across the four (4) lanes on both north- and southbound directions, and across the two (2) reversible lanes.

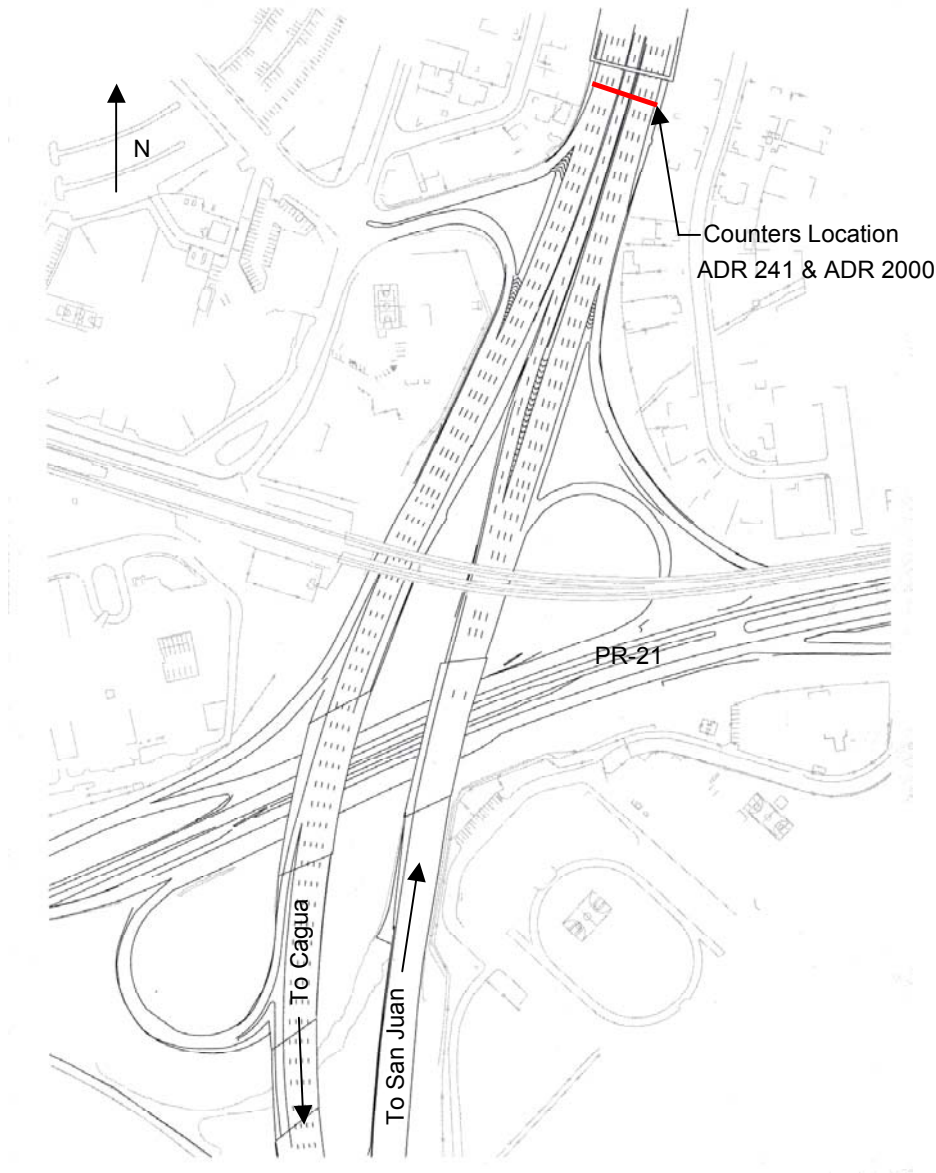


Figure 4.3 Counters location

4.1.1.4 Data Collection Technology

PRHTA currently uses automatic methods involving surface detectors, such as pneumatic road tubes to collect historical traffic data. The equipment used to collect the historical data used in this research was ADR 241 and ADR 2000 (Pérez 2005). These were used with pneumatic tubes. In the stations where there was vehicle classification, two pneumatic tubes were used spaced at 8 or 10 feet. If the purpose of the station was to collect volume data only, then one pneumatic tube was used. The counters are portable multilane automatic vehicular traffic counters with a rate of 200 counts per second per input and have a capacity of approximately 3,280 days of volume data (Pérez 2005).

4.1.1.5 Quality Control Checks

In the Office of Data Acquisition and Analysis of Transit of the PRHTA the process of quality control on raw data is conducted by two statisticians and one statistics supervisor. The data is compared to historical datasets to determine if it has a logical progression. In the instances where there are differences, a further study is conducted to determine the causes. Examples in which differences will occur are the opening of new routes or construction work conducted on adjacent routes that affect the expected traffic. For the permanent stations, the monthly average daily traffic (ADTs) are compared. If the ADT of a given month is not comparable to others and there is no explanation for it, then the entire month is eliminated. Malfunctioning of the equipment is usually the explanation for the latter cases. In the stations that classify vehicles, a quick

quality control check is performed by examining the percentage of vehicles in category 15 (this category indicates the vehicles that the equipment was unable to classify). A percentage of 5 or more (in category 15) requires a reinstallation of the station. In one of many personal communications, Mr. Pérez (sub-director of The Office of Data Acquisition and Analysis of Transit of the PRHTA) indicated that they were in the process of using the software capabilities of VITRIS to perform future quality control checks (Pérez 2004).

4.2 Examining and Preparing the Data

The second step of the KDD process involved examining and preparing the data to be mined. The examination of the gathered data was performed to the data mining database (or working database).

4.2.1 Data Analysis

The analysis of the data was one of the most extensive tasks conducted on the research. The archived ITS generated data from PR-18 consisted of over 133,000 data records. The total processing time was measured in days. Erroneous values and inconsistencies were the most important characteristics to be identified during the data analysis. Each dataset, that is work zones, traffic counts, speed, and accidents, was examined separately. The datasets were also examined as a whole to gain knowledge of the working database.

During this stage of the research, the work yielded a set of questions that helped identify remaining data needs, and the ranges of flow conditions

applicable to this project. It served as the root of the knowledge needed prior to the application of the data mining algorithms.

4.2.2 Processing the Data

The greatest challenge during the processing of the data was, perhaps, standardizing the archived datasets from the different sources. The lack of metadata, which is the information describing the data, made the interpretation and analysis more complicated. Additional telephone and email communications were conducted to obtain descriptive information about most of the archived data. For example, the spreadsheets with archived traffic data contained headings that were short of information for the understanding of the outside user. Descriptive information was requested. Weather data from NCDC presented the most complete descriptive information (NCDC 2005).

The standardized datasets were produced from the baseline data processing. Converting the submitted data into standardized data involved the longest amount of processing time. Once every dataset was standardized, the necessary statistical tests (i.e., maximum, minimum, average, standard deviation) and queries were easily completed. All the primary statistical tests were performed in MS Excel. The database capabilities of MS Access were used to carry out a few basic queries to initiate the learning phase of the KDD process.

This step of the process involved about twelve percent (12%) of the effort employed in the research, approximately. The major reason was the fact that the data did not include metadata. However, this was not a peculiarity of the data

used in this research, much of the datasets provided by USDOT's for research do not include metadata but a simple list of acronyms and description. Thus, the amount of effort required by users using datasets similar to the latter will not be too different.

The process of preparing the data is frequently confused with the process of examining the data. Even though one process is needed to obtain the other, these are two very different processes (Pyle 1999). The preparation of data involves the selection of variables with the purpose of manipulating and transforming raw data into data that are more easily accessible. Therefore, variables (fields) and rows (records) were selected carefully to minimize the time it took to build the model and to optimize the output obtained from the model. Examining the data consists of cleaning the data from any noise that it might have. After each dataset was transformed into a MS Excel file, the working database was created. The working database, for the purpose of this project, is the table that embraces all the necessary fields and records that were mined.

The data mining applications of IBM were chosen as an appropriate software package to achieve the objectives of this research. As part of the IBM scholars program (for teachers and graduate students), we were able to download all the requisites for the DB2 Intelligent Miner (IBM 2004). IBM's association rules allowed the identification of trends in people's driving behavior when combining factors such as work zones, accidents, and weather conditions.

The numerous IBM visualizer features facilitated the interpretation and analysis of results. Likewise, it produced excellent graphics for high quality presentations. Graphics developed by means of MS Excel were also used.

4.2.2.1 Data Quality and Validity Checks

Basic data quality checks were performed using the analysis tool packs of MS Excel 2000. The statistical test results from the north- and southbound traffic data are found in Appendix B.

4.3 Evaluating the Data Mining Application

The application of data mining algorithms in the transportation field is still new, yet the constant advances in the ITS technology, which have made the storage of multiple and very large databases possible, provide the basis for such algorithms to be used in the transportation domain. Technologies such as those described in Chapter 2 of this document are some examples of the resources that have made possible the proliferation of countless databases of research and government nature.

It is not feasible for people to analyze great amounts of data without assistance of appropriate computational tools. Traditional statistical analyses are used to provide descriptive information about the numerical portion of data in the databases. These usually start with a hypothesis, and statisticians develop their own equations to match such hypothesis. Data mining algorithms, on the other hand, do not require a hypothesis; the tool automatically develops the equations, and different types of data (e.g., categorical, numerical, continuous) can be used.

Manila (2000) suggests that the difference between statistical and data mining approaches does not have to do with the volume of the data itself, but with the number of variables or attributes which often have a much profound impact on the applicable analysis method. Therefore, the benefit of data mining over traditional statistical analysis is the ability to deal more effectively with complex interactions among variables rather than from the ability to process massive volumes of instances (Carey et al. 2003). The use of different types of data allows the combination of internal data (data collected by the agency) with external data (e.g., regulations, geographic, and demographic data from other agencies) to be used by the data mining applications in the extraction of hidden patterns. However, the interpretation of results is not easy for which statistical concepts must still be applied to the data during the data mining process (Carvalho 2007; Moss 2007). It should be acknowledged, that data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods.

There are four general data mining techniques: classification, clustering, numeric prediction, and association rules. The main difference between the four techniques is the way each extracts the information (i.e., the algorithms and methods used) and the way the results (knowledge discovery/rules) are expressed (Nassar 2007). Classification analysis is the most popular method. It focuses on identifying the characteristics that indicate the group or class to which each record in the database belongs. Clustering is probably the most complex method. It is used to group items that seem to fall naturally together. Numeric

prediction is similar to the classification method except that the result is a numeric value instead of a category. And, the association method consists of finding sets of items that occur simultaneously and frequently in a database.

There was minimum effort (approximately five percent (5%)) involved in choosing the data mining technique to utilize. The objectives of the research were listed early in the research and, thus, these required the association of itemsets from the working database. In addition, the remarkable difference of the data mining techniques facilitated the selection.

As the collection and archiving of data increases and computers get faster and cheaper, the interest of researchers and data analysts of discovering new patterns in the data increase as well. While this research applies the association method to ITS generated data from an expressway in the SJMR in Puerto Rico, there are many more applications available within the civil engineering domain.

Knowledge discovery in databases or KDD is considered to be the whole process of extraction of knowledge from data (Nassar 2007). Carvalho (2007) presents the KDD methodology as a four-stage process consisting of: system analysis, pre-processing of information, data mining, and post-processing of information. Fayyad et al. (1996) described the KDD process as having many stages. During the research described in this document, the seven-step KDD process was used (refer to Figure 4.1).

4.3.1 Association Method

The association method was introduced by Agrawal et al. in 1993 and basically consists of finding sets of items that occur simultaneously and

frequently in a database. This data mining technique basically makes repeated passes over the database to determine the commonly occurring itemsets. The discovery of association rules has been mostly used for in consumer-oriented industries such as grocery stores, phone companies, credit card companies, and banking applications. It is aimed at finding relationships between different attributes, often in large databases (Evans 2003). Association rules can predict any attribute and can predict more than one attribute's value at a time. The most common example of association mining would be whether customer who buys item A is also likely to buy item B, or customer who buys items A and B is also likely to buy item C. Another example, if a person opens a checking account, what is the probability that he/she will take out a loan? The idea is to determine the presence of some set of items given the presence of other items in a transaction. Each association rule extracted is usually provided with a confidence level and a support value; the confidence level is the statistical value presenting the probability of a certain rule, and the support value is the number of cases/projects in which the rule is found in the database (Nassar 2007).

The association mining application of the IBM Intelligent Miner for Data uses the Apriori algorithm to learn from the association rules found within the dataset. This algorithm is presented Figure 4.4. The top portion of Figure 4.4 lists the notation used by the Apriori algorithm (shown in the bottom portion). The first pass of the algorithm counts the item occurrences to determine the frequent 1-itemsets. The following pass, pass k , consists of two phases. In the first phase, the frequent itemsets L_{k-1} found in the $(k-1)^{\text{th}}$ pass are used to generate the

candidate itemsets C_k , using the Apriori candidate generation procedure. The second phase consists of scanning the database where the support of candidates in C_k is counted. A hash-tree data structure (data structure containing a tree of summary information about a larger piece of data) is used in order to efficiently determine the candidates in C_k contained in a given transaction t (Agrawal and Schafer 1996).

The candidate generation, as described by Agrawal and Shafer (1996), consists of taking L_{k-1} , described as the set of all frequent $(k-1)$ -itemsets, then generating a superset of the set of all frequent k -itemsets. The intuition behind the Apriori candidate generation procedure is that if an itemset X has minimum support, so do all subsets of X . Assuming that the items in each itemset are in chronological order, the candidate generation takes two steps. In the first step or the join step, L_{k-1} and L_{k-1} are joined:

```
insert into  $C_k$ 
select p.item1, p.item2, ..., p.itemk-1, q.itemk-1
from  $L_{k-1}$  p,  $L_{k-1}$  q
where p.item1 = q.item1, ..., p.itemk-2 = q.itemk-2, p.itemk-1 < q.itemk-1;
```

In the second step or the prune step, all the itemsets $c \in C_k$ are deleted in such a way that some $(k-1)$ -subset of c is not in L_{k-1} . For example, let L_3 be $\{\{1\ 2\ 3\}, \{1\ 2\ 4\}, \{1\ 3\ 4\}, \{1\ 3\ 5\}, \{2\ 3\ 4\}\}$. After the join step, C_4 will be $\{\{1\ 2\ 3\ 4\}, \{1\ 3\ 4\ 5\}\}$. The prune step will delete the itemset $\{1\ 3\ 4\ 5\}$ because the itemset $\{1\ 4\ 5\}$ is not in L_3 . Then we will be left with $\{1\ 2\ 3\ 4\}$ in C_4 .

k-itemset	An itemset having k items
L^k	Set of frequent k -itemsets (those with minimum support) Each member of this set has two fields: i) itemset and ii) support count
C^k	Set of candidate k -itemsets (potentially frequent itemsets) Each member of this set has two fields: i) itemset and ii) support count
P^i	Processor with id i
D^i	The dataset local to the processor P^i
DR^i	The dataset local to the processor P^i after repartitioning
C_k^i	The candidate set maintained with the Processor P^i during the k th pass (there are k items in each candidate)

```

 $L_1 := \{\text{frequent 1-itemsets}\};$ 
 $k := 2;$  //  $k$  represents the pass number
while ( $L_{k-1} <> 0$ ) do
  begin
     $C_k :=$  New candidates of size  $k$  generated from  $L_{k-1};$ 
    for all transactions  $t \in D$  do
      Increment the count of all candidates in  $C_k$  that are contained in  $t;$ 
     $L_k :=$  All candidates in  $C_k$  with minimum support;
     $k := k + 1;$ 
  end
Answer :=  $\bigcup_k L_k;$ 

```

Figure 4.4 Apriori Algorithm (Adapted from Agrawal and Shafer 1996)

Knowing the basic theory of association mining permits the application of the concept to ITS generated data. If instead of transactions one considers the presence of events, such as the events that occur in a highway at certain time of the day given the presence of other events, one might be able to find the relationship between events. For example, if it rains, what is the probability that drivers will reduce the speed? How often does rain affect traffic conditions? Other information that could be found after mining the data is how much will drivers reduce the speed? Likewise, if there is a work zone on the side of the road, what is the probability that the capacity of the highway be affected? Questions of this nature are important to highway engineers to broaden the knowledge of driver

behavior under specific external conditions, whether these are weather related or caused by human factor. These are events that may affect traffic flow conditions affecting, as a consequence, the capacity.

Association mining brings the unique ability to find the combinations of events that are not evident by simply scanning the data. Furthermore, it offers a different means of studying ITS generated data.

4.3.1.1 Sample Size

Given that data mining is viewed as a tool that can perform undirected and semi-directed analysis, the full length and width of very large datasets can be analyzed at much lower costs. Still, applying data mining to entire datasets of very large databases can involve high data management costs (in dollar and time required) for the agency or owner of the data. Carey et al. (2003); Zaki et al. (1996); Toivonen (1996) suggest that data mining models based on a sample of available data may be appropriate in many instances. Kotsiamtis and Kanellopoulos (2006) suggest that sampling the database may increase the efficiency of association rule algorithms. Nonetheless, the relationship between sample size and model accuracy is an important issue for data mining. A study conducted by Zaki et al. (1996), concluded that sampling does result in performance savings (computation time and reduced I/O costs) and good accuracy in practice. The same study also concluded that very small samples may generate many false rules, and degrade the performance. Oates and Jansen (1998) indicated that “increasing the amount of data used to build a model often results in linear increase in model size, even when that additional

complexity results in no significant increase in model accuracy.” On a similar note, Musick et al. (1993) suggest that the “economically rational decision is to use only a subset of the available data.”

An organization, such as a state agency, could find it equally costly to re-analyze a model on the basis of a sample of 20,000 records once a month or to re-analyze the model based on all available data once a year. In this case, the most effective strategy will be to model based on a sample if the event being modeled is relatively dynamic because the accuracy of the modeled sample will be representative of the model based on all available data. However, the most remarkable constrain of modeling all available data is the amount of rules that would be extracted by the data mining tool. The number of rules extracted is related to the number of itemsets within the model; thus, a model containing all available data would generate an enormous amount of insignificant rules. Rules that could be large in size as well, making the interpretation of new information an extremely tedious task.

Transportation generated data are generally dynamic datasets; thus different data types would require sampling models for convenience in achieving the objectives of the analysis with a cost effective approach. For example, it would not be cost effective to apply association mining to learn about the hidden patterns of work zones’ data and include all available data from the agency in the model. There are thousands of miles of highway in each state’s highway system from which a few hundred miles could have some type of work zone activities at the same time. Zaki et al. (1996) examined the relationship between itemset size

and the number of large itemsets and found out that the accuracy of sample models increased, being closer to the model of all available data, as the minimum support percentage decreased. The association mining tool from the IBM Intelligent Miner for Data allows the user to adjust the minimum support of every model being analyzed (refer to Figure C2.5 in Appendix C). Thus, the user is able to control the accuracy of the sample model being analyzed. A study conducted by Carey et al. (2003) concluded that accuracy increased at a modest rate when sampling models of 16,000 records or more, and accuracy increased at a decreasing rate when sampling models of 8,000 records or more.

In addition to sampling, the selection of relevant data for the working dataset should be conducted. For example when studying work zones, a type of data that should be excluded from the working dataset is data related to bridges, unless the work zone involves a particular bridge. The mining tool could extract a huge amount of insignificant rules that would overwhelm the user if exclusions of this sort were not performed prior to running the analysis. Therefore, it would be wiser to perform sampling models and also conduct a good selection of data to build the data mining database when dealing with transportation generated data.

4.4 Building the Model

According to the HCM there are three performance measures to characterize a freeway segment: density in terms of passenger cars equivalents per mile per lane, speed in terms of mean passenger car speed, and the ratio of volume to

capacity (HCM 2000). These measures capture the performance of a freeway segment, but how are these related to the quality of a trip from a drivers' point of view? The common expression of drivers concerning the quality of a trip is from the perspective of safety or comfort and travel time. These measures are the subject of this research.

A list of primary factors influencing drivers' sense of safety or comfort within the expressway include the following:

1. Traffic volume;
2. Average headway;
3. Effect of vehicle speed;
4. Effect of percentage of trucks;
5. Effect of the presence of work zones
 - ✓ On roadway, and
 - ✓ Off roadway;
6. Characteristics of freeway
 - ✓ Lane width,
 - ✓ Shoulder width (left and right shoulders),
 - ✓ Number of lanes,
 - ✓ Surface type (asphalt and/or concrete),
 - ✓ Separation of interchanges, and
 - ✓ Measured pavement roughness, international roughness index (IRI);
7. Weather conditions;
8. Time of day;

9. Number of accidents;
10. Speed of other vehicles;
11. Presence and width of shoulders;
12. Volume of traffic;
13. Presence of work zones;
14. Condition of pavement; and
15. Number of accidents.

These factors are the independent variables.

The model(s) was (were) built using the variables presented in Table 4.2. These variables were selected for the working database that was used in the mining tool. The table includes the number of variables used, variable name, description, position within the IBM data mining tool (for the purpose of the data mining application), and the data type. Notice that every variable was converted into categorical type variable as a requirement of the association mining tool. The query applications of the mining tool were used to select the appropriate variables to be studied in each of the models created.

This step of the KDD process involved approximately five percent (5%) of the effort employed in the research. The association mining applications of the IBM Intelligent Miner for Data provides user friendly windows that allow the creation of models fairly easy.

Table 4.2 Variables, description, position, and category of data as used in the mining tool

No.	Variable	Description	Position (IBM-DM Application)	Data Type
1	MONTH	Month	1-3	Categorical
2	TIME	Time	7-10	Categorical
3	OBS	Observations	14-17	Categorical
4	DIR	Direction	21-22	Categorical
5	SU	Sunday Traffic	26-29	Categorical
6	SUS	Sunday Speed	33-34	Categorical
7	SUD	Sunday Density	38-40	Categorical
8	SULOS	Sunday LOS	44-44	Categorical
9	W_SU	Weather Sunday	48-50	Categorical
10	M	Monday Traffic	54-57	Categorical
11	MS	Monday Speed	61-62	Categorical
12	MD	Monday Density	66-68	Categorical
13	MLOS	Monday LOS	72-72	Categorical
14	W_M	Weather Monday	76-78	Categorical
15	T	Tuesday Traffic	82-85	Categorical
16	TS	Tuesday Speed	89-90	Categorical
17	TD	Tuesday Density	94-96	Categorical
18	TLOS	Tuesday LOS	100-100	Categorical
19	W_T	Weather Tuesday	104-106	Categorical
20	W	Wednesday Traffic	110-113	Categorical
21	WS	Wednesday Speed	117-118	Categorical
22	WD	Wednesday Density	122-124	Categorical
23	WLOS	Wednesday LOS	128-128	Categorical
24	W_W	Weather Wednesday	132-134	Categorical
25	TH	Thursday Traffic	138-141	Categorical
26	THS	Thursday Speed	145-146	Categorical
27	THD	Thursday Density	150-152	Categorical
28	THLOS	Thursday LOS	156-156	Categorical
29	W_TH	Weather Thursday	160-162	Categorical
30	F	Friday Traffic	166-169	Categorical
31	FS	Friday Speed	173-174	Categorical

No.	Variable	Description	Position (IBM-DM Application)	Data Type
32	FD	Friday Density	178-180	Categorical
33	FLOS	Friday LOS	184-184	Categorical
34	W_F	Weather Friday	188-190	Categorical
35	S	Saturday Traffic	194-197	Categorical
36	SS	Saturday Speed	201-202	Categorical
37	SD	Saturday Density	206-208	Categorical
38	SLOS	Saturday LOS	212-212	Categorical
39	W_S	Weather Saturday	216-218	Categorical
40	WZ	Work Zones	222-227	Categorical
41	ADY	Accident Day	231-232	Categorical
42	ALO	Accident Location	236-239	Categorical
43	ADU	Accident Duration	244-248	Categorical
44	ATY	Accident Type	252-256	Categorical

Three of the variables (work zone, accident type, and weather) consisted of a list of categories that are worth mentioning. Table 4.3 presents the categories within each of these variables. There were three types of work zones, twelve types of accidents, and two types of weather for the time period the data was obtained for. Numerical weather values were transformed to categorical type data by using the simplest categories: WET or DRY. These types of data were mainly included to study the associations of weather and accidents, and how these affect the traffic flow and speed. Given the complexity of quantifying the length of time that takes the pavement to dry, the values used to describe the weather discard any accumulation of water.

Table 4.3 Work zone, accident, and weather variables used in the study

Work Zones Categories	Description
RHPAVE	Rehabilitation of Pavement
GSFIMP	Geometric and Traffic Safety Improvement
CNSBRR	Construction of Noise Barrier
111111	No Work Zone
Accident Categories	Description
W_BRD	With Bridge
PHOLE	Pot Hole
BARRI	Temporary Concrete Barrier
2_VEH	2 Vehicles
3+VEH	3+ Vehicles
DRUMS	Drums
H_&_R	Hit & Run
PKVEH	Parked Vehicle
FIXOB	Fixed Object
WTREE	With Tree
PEDST	Pedestrian
DEBRI	Debris
11111	No Accident
Weather Categories	Description
WET	0.01+ rainfall reported in SJU
DRY	Zero "0" rainfall reported in SJU

Six (6) specific studies in which association mining was used to extract hidden patterns from the large database are presented in Chapter 5.0.

5. EVALUATION OF THE MODEL(S)

Six (6) specific studies were conducted using the working dataset described in Chapter 4.0 to evaluate the use of data mining for the analysis of ITS generated data. The first five (5) studies consisted of the application of association mining to analyze different variables while conducting a series of exclusions in the dataset, and the sixth study consists of the comparison between the traditional method (i.e., the analysis of traffic data in terms of density versus time, flow rate versus time, and mean speed versus time) and the data mining approach for the analysis of ITS generated data. Table 5.1 presents a summary of the different studies conducted.

The evaluation of the models created throughout the research, which were far more than the ones presented in this document, consisted of the most extensive step of the research. The time and effort involved in the evaluation of the sets of rules extracted for each model consisted of approximately forty percent (40%) of the effort employed in the entire research. The amount of rules extracted by the mining tool in the early stages of the research can be overwhelming and its analysis can be time consuming. However, the more models one create and the more sets of rules one evaluates, the more information one can extract from the data itself. Thus, this would almost certainly be the most time and effort consuming step of the KDD process.

For the purpose of the analyses conducted in each study, when referred to as the relationship of “X” variable(s) versus “Y” variable, the “X” variable(s) is the variable used as the transaction field by the algorithms of the mining tool, thus it will not be shown on any of the rules. “Y” variable, on the other hand, will always be shown on the association rules (for details, refer to the association mining preparation description in Appendix C).

Table 5.1 Summary of studies

Study Section	Description of Study
STUDY I	Relationship between accident type and average vehicle speed for Friday
	Relationship between accident types and average vehicle speed for Friday during work zone activities
STUDY II	Relationship between work zone activities and density data for Mondays southbound
	Relationship between work zone activities and density data for Mondays northbound
STUDY III	Relationship between duration of accidents and accident types
	Relationship between accident types and accident day
STUDY IV	Relationship between accident types and accidents of 3 or more vehicle
	Relationship between accident types during work zone activities and accident types during no work zone activities
	Relationship between accident types during no work zone activities with a tree
STUDY V	Relationship between months versus LOS southbound and months versus LOS northbound - 5:45AM-9:30AM
	Relationship between months versus LOS southbound and months versus LOS northbound - 9:30AM-1:15PM
	Relationship between months versus LOS southbound and months versus LOS northbound - 5:30PM-7:15PM
	Relationship between the worst LOS during work zone activities
	Relationship between LOS and weekdays
	Relationship between weather, work zone activities, and weekdays LOS
STUDY VI	Comparison of traditional and data mining approach for analyzing ITS generated data

5.1 Study I – Association Mining to Analyze Average Vehicle Speeds

Association mining was applied to analyze average vehicle speeds in this study. In order to evaluate the methodology, several scenarios were examined. The first scenario examines the relationships between accident types and average vehicle speeds for Fridays by direction. Thus, the dataset was queried twice, once to exclude all records that did not contain some type of accident and all records that did not contain San Juan as the direction of travel, and once more to include all records that contained some type of accident and to exclude all records that did not have Caguas for the direction of travel. Figure 5.1 illustrates the rules extracted for the former, and Figure 5.2 illustrates the rules extracted for the latter.

Rule	▼ Support	Confidence
[23] ==> [47]	16.6667%	100.0000%
[21] ==> [20]	16.6667%	100.0000%
[49] ==> [33]	8.3333%	100.0000%
[45]+[20] ==> [21]	8.3333%	100.0000%
[45] ==> [21]	8.3333%	100.0000%
[45] ==> [20]	8.3333%	100.0000%
[40]+[47] ==> [36]	8.3333%	100.0000%
[40]+[47] ==> [32]	8.3333%	100.0000%
[40]+[36]+[47] ==> [32]	8.3333%	100.0000%
[40]+[36]+[32] ==> [47]	8.3333%	100.0000%
[40]+[36] ==> [47]	8.3333%	100.0000%
[40]+[36] ==> [32]	8.3333%	100.0000%
[40]+[32]+[47] ==> [36]	8.3333%	100.0000%
[40]+[32] ==> [47]	8.3333%	100.0000%
[40]+[32] ==> [36]	8.3333%	100.0000%
[36]+[47] ==> [40]	8.3333%	100.0000%
[36]+[47] ==> [32]	8.3333%	100.0000%
[36]+[32]+[47] ==> [40]	8.3333%	100.0000%
[36]+[32] ==> [47]	8.3333%	100.0000%
[36]+[32] ==> [40]	8.3333%	100.0000%
[36] ==> [47]	8.3333%	100.0000%
[36] ==> [40]	8.3333%	100.0000%
[36] ==> [32]	8.3333%	100.0000%
[34] ==> [40]	8.3333%	100.0000%
[33] ==> [49]	8.3333%	100.0000%
[32]+[47] ==> [40]	8.3333%	100.0000%
[32]+[47] ==> [36]	8.3333%	100.0000%
[32] ==> [47]	8.3333%	100.0000%
[32] ==> [40]	8.3333%	100.0000%
[32] ==> [36]	8.3333%	100.0000%
[29] ==> [41]	8.3333%	100.0000%
[23]+[18] ==> [47]	8.3333%	100.0000%
[22]+[47] ==> [19]	8.3333%	100.0000%
[22]+[19] ==> [47]	8.3333%	100.0000%
[21]+[45] ==> [20]	8.3333%	100.0000%
[19]+[47] ==> [22]	8.3333%	100.0000%
[19] ==> [47]	8.3333%	100.0000%
[19] ==> [22]	8.3333%	100.0000%
[18]+[47] ==> [23]	8.3333%	100.0000%
[18] ==> [47]	8.3333%	100.0000%
[18] ==> [23]	8.3333%	100.0000%

Figure 5.1 Association rules - accident type versus average vehicle speeds friday - northbound

The rules extracted by the mining tool are the classes (from the variable being analyzed) found to relate given the query conditions. The rules consist of the body which is the IF- condition(s) and the head which is the THEN- condition.

▼ Rule	Support	Confidence
[45] ==> [21]	9.0909%	100.0000%
[44] ==> [39]	9.0909%	100.0000%
[43] ==> [21]	9.0909%	100.0000%
[39] ==> [44]	9.0909%	100.0000%
[26] ==> [22]	9.0909%	100.0000%
[19] ==> [25]	9.0909%	100.0000%

Figure 5.2 Association rules - accident type versus average vehicle speeds friday - southbound

A general pattern was observed, the number and size of the rules was different for each direction of traffic given the same conditions. The mining tool extracted 41 rules for the northbound model and only 6 rules for the southbound model given the same conditions. In addition, the rules extracted for the northbound model consisted of larger rules (i.e., more IF- statements in the body of the association rules) than those for the southbound model. A particular pattern was found from the rules extracted for the northbound direction (Figure 5.1): both 20 and 47 mph repeated the most and the same number of times. These speeds are a representation of the upper and lower end range of average vehicle speeds found for Friday (refer to Tables B.1 and B.2). Several average vehicle speeds repeated equally for the southbound model (Figure 5.2). On the contrary, 50 percent of the speeds extracted for the southbound model are in the lower end of the speed range for Friday (refer to Tables B.1 and B.2). This could

indicate that vehicles are traveling at slower speeds on the southbound direction during some type of accident.

To determine if the size of the rules was a real pattern, models for other days of the week were developed for which similar patterns were detected. Thus, the raw data was examined as well for some type of information that could indicate the relationship between the fewer and smaller rules for the southbound direction and the numerous and larger rules for the northbound direction. The number of accidents in the southbound direction given the conditions indicated to the mining tool was only 65% of those for the northbound direction. As a result, the mining tool found many more relationships for the northbound direction model(s) because more accidents were found as opposed to for the southbound direction model(s) in which fewer accidents were found in the raw data.

The second scenario consisted of the same conditions as the first scenario but excluded all accidents that did not occur during work zone activities. Figure 5.3 shows the rules extracted for the northbound given the conditions previously mentioned.

Rule	▼ Support	Confidence
[23] ==> [47]	28.5714%	100.0000%
[36]+[40]+[47] ==> [32]	14.2857%	100.0000%
[36]+[32]+[47] ==> [40]	14.2857%	100.0000%
[36]+[40]+[32] ==> [47]	14.2857%	100.0000%
[40]+[32]+[47] ==> [36]	14.2857%	100.0000%
[36]+[40] ==> [47]	14.2857%	100.0000%
[40]+[32] ==> [36]	14.2857%	100.0000%
[36]+[40] ==> [32]	14.2857%	100.0000%
[32]+[47] ==> [40]	14.2857%	100.0000%
[32]+[47] ==> [36]	14.2857%	100.0000%
[36]+[47] ==> [32]	14.2857%	100.0000%
[36]+[47] ==> [40]	14.2857%	100.0000%
[40]+[47] ==> [32]	14.2857%	100.0000%
[40]+[47] ==> [36]	14.2857%	100.0000%
[23]+[18] ==> [47]	14.2857%	100.0000%
[18]+[47] ==> [23]	14.2857%	100.0000%
[36]+[32] ==> [47]	14.2857%	100.0000%
[40]+[32] ==> [47]	14.2857%	100.0000%
[36]+[32] ==> [40]	14.2857%	100.0000%
[36] ==> [40]	14.2857%	100.0000%
[40] ==> [47]	14.2857%	100.0000%
[40] ==> [32]	14.2857%	100.0000%
[40] ==> [36]	14.2857%	100.0000%
[18] ==> [47]	14.2857%	100.0000%
[32] ==> [47]	14.2857%	100.0000%
[18] ==> [23]	14.2857%	100.0000%
[36] ==> [47]	14.2857%	100.0000%
[32] ==> [40]	14.2857%	100.0000%
[36] ==> [32]	14.2857%	100.0000%
[32] ==> [36]	14.2857%	100.0000%

zone activity - northbound

Fewer rules (30 rules) were extracted for the northbound model given the new conditions. Also, the average vehicle speed with the highest support (or the average vehicle speed that repeated the most in the queried dataset) was likely to be 47 mph; in the first scenario, the average speed was divided between 20 and 47 mph. The rules for the southbound model also decreased in number; the mining tool extracted only 5 rules (Figure 5.4). For these conditions, the slower vehicles were dropped out of the THEN- portion of the association rules extracted for the first scenario.

Rule	▲ Support	Confidence
[44] ==> [39]	10.0000%	100.0000%
[26] ==> [22]	10.0000%	100.0000%
[39] ==> [44]	10.0000%	100.0000%
[22] ==> [26]	10.0000%	100.0000%
[19] ==> [25]	10.0000%	100.0000%

Figure 5.4 Association rules - accident type versus average vehicle speeds during work zone activity - southbound

The application of data mining to analyze average vehicle speeds could be useful for transportation officials to obtain a general idea of the average vehicle speed(s) that would most likely to be found given the conditions established (presented in the THEN- portion of the rules). Furthermore, it could provide the direction with the most accidents in a particular segment of highway by examining the number and the size of the rules extracted. The latter could also be obtained easily by querying the database.

5.2 Study II – Association Mining to Analyze Density

Data mining was used in the analysis of density data. Several conditions were chosen to evaluate the information that could be obtained from the rules extracted by the mining tool. The relationships between work zone activities and density data for Mondays were examined. The dataset was queried to only include the records for which work zone activities and some type of accident had occurred. In addition, the data was segregated by direction of travel. Figure 5.5 shows a portion of the association rules extracted by the mining tool given the previous conditions for the southbound direction. 1,018 rules were extracted for this model. Figure 5.6 lists the rules extracted for the northbound model. Only

12 rules were extracted for the same conditions. From examining the raw dataset, it was learned that the number of accidents during work zone activities for the southbound direction was slightly higher than those for the northbound direction. However, many more different density categories were found within the excluded dataset for the southbound direction than that for the northbound direction. Thus, there were more categories to relate to the rest of the dataset in the southbound direction model.

Rule	Support	Confidence
[1 06]+[082]+[1 14]+[074]+[043]+[1 25]+[059] ==> [068]	25.0000%	100.0000%
[082]+[068]+[1 14]+[074]+[043]+[1 25]+[059] ==> [1 06]	25.0000%	100.0000%
[1 06]+[082]+[068]+[074]+[043]+[1 25]+[059] ==> [1 14]	25.0000%	100.0000%
[1 06]+[068]+[1 14]+[074]+[043]+[1 25]+[059] ==> [082]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[074]+[1 25]+[059] ==> [043]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[074]+[043]+[1 25] ==> [059]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[074]+[043]+[059] ==> [1 25]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[043]+[1 25]+[059] ==> [074]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[074]+[043] ==> [1 25]	25.0000%	100.0000%
[1 06]+[082]+[1 14]+[074]+[043]+[059] ==> [1 25]	25.0000%	100.0000%
[1 06]+[082]+[1 14]+[074]+[043]+[059] ==> [068]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[074]+[059] ==> [1 25]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[074]+[059] ==> [043]	25.0000%	100.0000%
[068]+[1 14]+[074]+[043]+[1 25]+[059] ==> [082]	25.0000%	100.0000%
[068]+[1 14]+[074]+[043]+[1 25]+[059] ==> [1 06]	25.0000%	100.0000%
[082]+[068]+[1 14]+[043]+[1 25]+[059] ==> [074]	25.0000%	100.0000%
[082]+[068]+[1 14]+[043]+[1 25]+[059] ==> [1 06]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[074]+[1 25] ==> [059]	25.0000%	100.0000%
[1 06]+[082]+[1 14]+[074]+[043]+[1 25] ==> [068]	25.0000%	100.0000%
[1 06]+[082]+[068]+[074]+[043]+[1 25] ==> [1 14]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[043]+[1 25] ==> [074]	25.0000%	100.0000%
[1 06]+[082]+[1 14]+[043]+[1 25]+[059] ==> [074]	25.0000%	100.0000%
[1 06]+[082]+[1 14]+[043]+[1 25]+[059] ==> [068]	25.0000%	100.0000%
[1 06]+[068]+[1 14]+[043]+[1 25]+[059] ==> [074]	25.0000%	100.0000%
[1 06]+[068]+[1 14]+[043]+[1 25]+[059] ==> [082]	25.0000%	100.0000%
[1 06]+[082]+[1 14]+[074]+[043]+[1 25] ==> [059]	25.0000%	100.0000%
[082]+[1 14]+[074]+[043]+[1 25]+[059] ==> [068]	25.0000%	100.0000%
[082]+[1 14]+[074]+[043]+[1 25]+[059] ==> [1 06]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[074]+[1 25] ==> [043]	25.0000%	100.0000%
[1 06]+[068]+[1 14]+[074]+[043]+[059] ==> [1 25]	25.0000%	100.0000%
[1 06]+[068]+[1 14]+[074]+[043]+[059] ==> [082]	25.0000%	100.0000%
[1 06]+[082]+[1 14]+[074]+[1 25]+[059] ==> [043]	25.0000%	100.0000%
[1 06]+[082]+[1 14]+[074]+[1 25]+[059] ==> [068]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[043]+[1 25] ==> [059]	25.0000%	100.0000%
[1 06]+[082]+[068]+[074]+[043]+[059] ==> [1 25]	25.0000%	100.0000%
[1 06]+[082]+[068]+[074]+[043]+[059] ==> [1 14]	25.0000%	100.0000%
[1 06]+[082]+[068]+[043]+[1 25]+[059] ==> [074]	25.0000%	100.0000%
[1 06]+[082]+[068]+[043]+[1 25]+[059] ==> [1 14]	25.0000%	100.0000%
[1 06]+[1 14]+[074]+[043]+[1 25]+[059] ==> [068]	25.0000%	100.0000%
[1 06]+[1 14]+[074]+[043]+[1 25]+[059] ==> [082]	25.0000%	100.0000%
[1 06]+[082]+[068]+[1 14]+[074]+[043] ==> [059]	25.0000%	100.0000%
[1 06]+[068]+[1 14]+[074]+[043]+[1 25] ==> [082]	25.0000%	100.0000%
[1 06]+[082]+[068]+[074]+[043]+[1 25] ==> [059]	25.0000%	100.0000%
[082]+[068]+[1 14]+[074]+[043]+[059] ==> [1 25]	25.0000%	100.0000%
[082]+[068]+[1 14]+[074]+[043]+[059] ==> [1 06]	25.0000%	100.0000%
[082]+[068]+[1 14]+[074]+[1 25]+[059] ==> [043]	25.0000%	100.0000%
[082]+[068]+[1 14]+[074]+[1 25]+[059] ==> [1 06]	25.0000%	100.0000%

265 item sets and 1,018 rules in the model

Figure 5.5 Association rules - work zones versus density monday - southbound

Rule	Support	Confidence
[057]+[035] ==> [049]	12.5000%	100.0000%
[035]+[049] ==> [057]	12.5000%	100.0000%
[057]+[049] ==> [035]	12.5000%	100.0000%
[057] ==> [049]	12.5000%	100.0000%
[057] ==> [035]	12.5000%	100.0000%
[120] ==> [090]	12.5000%	100.0000%
[044] ==> [048]	12.5000%	100.0000%
[048] ==> [044]	12.5000%	100.0000%
[049] ==> [035]	12.5000%	100.0000%
[049] ==> [057]	12.5000%	100.0000%
[035] ==> [049]	12.5000%	100.0000%
[090] ==> [120]	12.5000%	100.0000%
[035] ==> [057]	12.5000%	100.0000%

Figure 5.6 Association rules - work zones versus density monday - northbound

During work zone activities, it is more likely to breakdown. Another distinctive pattern observed involved the density values in each set of rules extracted. In the southbound model, there were larger density values in the THEN- portion of the rules than in the northbound model.

This application could be useful for transportation officials to determine the segment of highway with the most changes in density given particular conditions, such as work zone activities. Large changes in density values within a particular segment of highway indicate large fluctuations in the conditions of traffic. When comparing the same segment of highway in different directions given the same conditions, one would be able to determine which direction of traffic is most likely to operate near capacity.

5.3 Study III – Association Mining to Analyze Accident Data

In this study, the application of association mining to analyze accidents in a segment of PR-18 was explored. The main purpose was to explore any relationship between relevant variables that might reveal hidden knowledge about the accidents. Given that accidents do not always occur at exactly the same location, every accident within 1 mile ahead from where the traffic count data was collected was used in this study. Several queries were attempted to become familiar with the application and also to learn as much from the data as possible. Figure 5.7 lists the association rules developed for the relationships between the duration of accidents and the accident types (refer to Table 4.3 for all the accident types in the working database). The color code identifies the support or number of cases/times in which the rule is found in the database. In this example, the rule IF temporary concrete barrier is found, THEN accident types between 2 vehicles ([BARRI]==>[2_VEH]) is found more times than any other rule, thus it is shown in a different color. The rest of the rules were found the same amount of times within the database and therefore are shown in the same color. Given that the confidence indicates the statistical value presenting the probability of a certain rule, notice that every rule has a 100% confidence value.

Rule	▲ Support	Confidence
[BARRI]+[3+VEH] ==> [2_VEH]	11.1111%	100.0000%
[BARRI]+[3+VEH] ==> [FIXOB]	11.1111%	100.0000%
[3+VEH]+[2_VEH] ==> [FIXOB]	11.1111%	100.0000%
[DRUMS] ==> [W_BRD]	11.1111%	100.0000%
[3+VEH]+[FIXOB]+[2_VEH] ==> [BARRI]	11.1111%	100.0000%
[BARRI]+[3+VEH]+[FIXOB] ==> [2_VEH]	11.1111%	100.0000%
[FIXOB]+[2_VEH] ==> [3+VEH]	11.1111%	100.0000%
[3+VEH]+[2_VEH] ==> [BARRI]	11.1111%	100.0000%
[BARRI]+[FIXOB] ==> [2_VEH]	11.1111%	100.0000%
[FIXOB]+[2_VEH] ==> [BARRI]	11.1111%	100.0000%
[W_BRD] ==> [DRUMS]	11.1111%	100.0000%
[BARRI]+[FIXOB] ==> [3+VEH]	11.1111%	100.0000%
[BARRI]+[3+VEH]+[2_VEH] ==> [FIXOB]	11.1111%	100.0000%
[FIXOB] ==> [2_VEH]	11.1111%	100.0000%
[FIXOB] ==> [3+VEH]	11.1111%	100.0000%
[3+VEH]+[FIXOB] ==> [2_VEH]	11.1111%	100.0000%
[FIXOB] ==> [BARRI]	11.1111%	100.0000%
[3+VEH]+[FIXOB] ==> [BARRI]	11.1111%	100.0000%
[BARRI]+[FIXOB]+[2_VEH] ==> [3+VEH]	11.1111%	100.0000%
[BARRI] ==> [2_VEH]	33.3333%	100.0000%

Figure 5.7 Association rules - accident duration vs accident type

The actual accidents that occurred during weekdays on the 1-mile stretch of highway in the northbound direction are presented in Table 5.2. The information of Table 5.2 was obtained by manipulating the raw data in an Excel spreadsheet. The association tool was able to relate six of the nine accident types with the accident duration.

Table 5.2 Accidents northbound

ACCIDENT TYPE	ACCIDENT DAY	NO. OF ACCIDENTS
FIXOB*	M	1
BARRI*	M, W, W, F	4
3+VEH*	M, M, W	3
2_VEH*	T, TH, F, F, F, F	6
H_&_R	M	1
W_BRD*	F	1
PHOLE	T	1
DRUMS*	TH, TH	2
PKVEH	TH	1

* Accidents used by mining tool

Accident types BARRI (temporary concrete barrier) and 2_VEH (2 vehicles) occurred the most often, according to the data in Table 5.2. The mining tool was able to pick that information from the data, as it presented the rule [BARRI]==>[2_VEH] with the highest support percentage. In the example shown in Figure 5.7, both the IF- and THEN- side of the rules were accident types. The most significant rule that was found states that where accidents with temporary barriers are occurring, so do accidents between 2 vehicles (with a 100% confidence and supported by 33.33% of the times). The confidence being the statistical value presenting the probability of that rule and the support is the number of times in which the rule was found.

The day of the week in which the accidents occurred was also obtained from the mining tool. This study was performed by changing the input fields; now the accident types were related to the accident day. Figure 5.8 shows the support of

each day of the week in which the accidents occurred. The highest number of accidents appear to have occurred on both Monday (M) and Friday (F). This information can also be found in Table 5.1, although a few more steps were required to obtain the same information.

Item Set	▲ Support
[F_]+[W_]	6.6667%
[T_]	13.3333%
[TH]	20.0000%
[W_]	20.0000%
[F_]	26.6667%
[M_]	26.6667%

Figure 5.8 Support of days in which accidents occurred

It is interesting that most of the accidents occurred both on the busiest (Friday) and the slowest (Monday) day of the week. Thus, the mining tool allowed the identification of the weekday in which more or fewer accidents are likely to occur.

The goal is not just to find rules that are similar to one another, but also to extract a number of similar rules that would resemble a trend. For example, there are three patterns in the association rules presented in Figure 5.7. The rules have been sorted for easier reading in Figure 5.9. Pattern 1 indicates that accidents between 2 vehicles occurred the most; Pattern 2 indicates that accidents with temporary barriers occurred second to most; and Pattern 3 indicates that accidents between 3 or more vehicles occurred third to most if we queried the database in the 1-mile segment of highway that was analyzed.

However, the rules indicate that similar events occurred that led to that head accident in the rule to occur. For example, from pattern 3 in Figure 5.9 it is observed that the mining tool extracted accident type [FIXOB] (or fixed object) on every rule. The number of times that accidents with a fixed object occurred is not relevant, but the events that led to that type of accident. Thus, the mining tool extracted a relationship between the events that led to accidents with fixed objects with those that led to accidents between 3 or more vehicles (depicted as [3+VEH]).

Body of Rule

Head of Rule

Pattern 1

Rule	Support	Confidence
[BARRI]+[3+VEH] ==> [2_VEH]	11.1111%	100.0000%
[BARRI]+[3+VEH]+[FIXOB] ==> [2_VEH]	11.1111%	100.0000%
[BARRI]+[FIXOB] ==> [2_VEH]	11.1111%	100.0000%
[FIXOB] ==> [2_VEH]	11.1111%	100.0000%
[3+VEH]+[FIXOB] ==> [2_VEH]	11.1111%	100.0000%
[BARRI] ==> [2_VEH]	33.3333%	100.0000%
[FIXOB]+[2_VEH] ==> [3+VEH]	11.1111%	100.0000%
[BARRI]+[FIXOB] ==> [3+VEH]	11.1111%	100.0000%
[FIXOB] ==> [3+VEH]	11.1111%	100.0000%
[BARRI]+[FIXOB]+[2_VEH] ==> [3+VEH]	11.1111%	100.0000%
[3+VEH]+[FIXOB]+[2_VEH] ==> [BARRI]	11.1111%	100.0000%
[3+VEH]+[2_VEH] ==> [BARRI]	11.1111%	100.0000%
[FIXOB]+[2_VEH] ==> [BARRI]	11.1111%	100.0000%
[FIXOB] ==> [BARRI]	11.1111%	100.0000%
[3+VEH]+[FIXOB] ==> [BARRI]	11.1111%	100.0000%
[W_BRD] ==> [DRUMS]	11.1111%	100.0000%
[BARRI]+[3+VEH] ==> [FIXOB]	11.1111%	100.0000%
[3+VEH]+[2_VEH] ==> [FIXOB]	11.1111%	100.0000%
[BARRI]+[3+VEH]+[2_VEH] ==> [FIXOB]	11.1111%	100.0000%
[DRUMS] ==> [W_BRD]	11.1111%	100.0000%

Pattern 2

Pattern 3

Figure 5.9 Patterns within the association rules

This type of analysis also allowed for the determination of accident types that are most common as a group, for example accidents with temporary barriers, between 2 vehicles, and between 3 or more vehicles. These results could allow transportation researchers to examine particular cases, such as the ones mentioned, that could lead to further investigations. The approach presented provides an additional tool for transportation officials to use in the analysis of data for their region, given that similar concerns are common in different cities/metropolitan regions. Some of the questions that were answered using this approach were:

- ✓ What type of accident is occurring more often?
- ✓ In what day of the week are these accidents occurring?
- ✓ What group of accidents is occurring more often?

Other questions that could be answered with data mining or database query are:

- ✓ What are the average vehicle speeds for the most common accident types?
- ✓ In what segment of road are more accidents occurring?

On the other hand, some questions could be raised from this analysis. These are:

- ✓ Why are more accidents occurring on certain days of the week?
- ✓ Why are certain accident types related to certain average vehicle speeds?
- ✓ Which accident types could be managed in a more than reasonable time frame?

- ✓ Are there new/old traffic control measures that could be used to assess certain accident types?
- ✓ What events lead to a particular accident type?

More utilization and experimentation with the archived data will improve the performance of the agency or TMC for which the research is being conducted. Researchers will become more familiar with the data and the new information will lead to further investigations.

5.4 Study IV – Association Mining to Analyze Work Zones Data

Association mining was applied to analyze work zones data in this study. Two principal parameters were established prior to the analysis; only records that contained weekday data and records that contained work zones were used. This was done to avoid having the mining tool generate any relationships with the type of work zone depicted as [111111] (“no work zone”) or work zones during weekends. From the 176 rules extracted by the mining tool, certain patterns were identified. Figure 5.10 lists the 30 rules that relate different accidents with accident type 3 or more vehicles.

Rule	Support	Confidence
[BARRI]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[BARRI] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[W_BRD]+[BARRI] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[BARRI]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[W_BRD]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[W_BRD]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[DRUMS]+[W_BRD]+[BARRI]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[DRUMS]+[BARRI]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[DRUMS]+[W_BRD]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[W_BRD]+[BARRI]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[W_BRD]+[BARRI] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[W_BRD] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[W_BRD] ==> [3+VEH]	25.0000%	100.0000%
[DRUMS]+[BARRI] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[BARRI]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[DRUMS] ==> [3+VEH]	25.0000%	100.0000%
[DRUMS]+[W_BRD] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[W_BRD]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[BARRI] ==> [3+VEH]	25.0000%	100.0000%
[DRUMS]+[W_BRD]+[BARRI] ==> [3+VEH]	25.0000%	100.0000%
[W_BRD]+[BARRI]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[W_BRD] ==> [3+VEH]	25.0000%	100.0000%
[DRUMS]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[W_BRD]+[BARRI]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[W_BRD]+[BARRI] ==> [3+VEH]	25.0000%	100.0000%

Figure 5.10 Relationship of accidents during work zones with accidents between 3+ vehicles

A pattern that stands out from the rules extracted is the manifestation of accidents with a parked vehicle (depicted as [PKVEH]). Considering that a query was conducted prior to the analysis to only include those records for which a work zone was found, the fact that this particular accident was related to others is interesting. It could imply that the parked vehicle was (were) one of the construction vehicles. The events that led to this type of accident led to 53.33%

(i.e., 16 rules in which [PKVEH] appear / 30 rules in which [3+VEH] is the head) of the accidents that led to accidents between 3 or more vehicles.

Figure 5.11 shows the longest rules extracted; these contain five (5) rules in the IF- side of the equation. The interesting fact about these rules is that accident type between 2 vehicles (depicted as [2_VEH]) was not extracted in any of them, although it was found in the item set. Every other accident type was extracted, indicating that there were similar trends in the events that led to each. The first rule in Figure 5.11 would read as follows:

IF work zone AND weekday AND accident with a parked vehicle AND accident with a drum AND accident with temporary concrete barrier AND accident with pothole, THEN accident between 3 or more vehicles (with a 100% confidence and 25% support).

The previous rule indicates that the events that led to accidents with parked vehicles, drums, bridge, temporary concrete barrier, and pothole were the events that led to accidents between 3 or more vehicles.

Rule	Support	Confidence
[PKVEH]+[DRUMS]+[W_BRD]+[BARRI]+[PHOLE] ==> [3+VEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[W_BRD]+[PHOLE]+[3+VEH] ==> [BARRI]	25.0000%	100.0000%
[PKVEH]+[W_BRD]+[BARRI]+[PHOLE]+[3+VEH] ==> [DRUMS]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[W_BRD]+[BARRI]+[3+VEH] ==> [PHOLE]	25.0000%	100.0000%
[DRUMS]+[W_BRD]+[BARRI]+[PHOLE]+[3+VEH] ==> [PKVEH]	25.0000%	100.0000%
[PKVEH]+[DRUMS]+[BARRI]+[PHOLE]+[3+VEH] ==> [W_BRD]	25.0000%	100.0000%

Figure 5.11 Relationships of five accidents

The mining tool was also able to extract, from the working database and considering the initial queries, the accident types that occurred more often during

a work zone activity. Figure 5.12 lists the support of accidents between 3 or more vehicles ([3+VEH]) and accidents with a temporary concrete barrier ([BARRI]) as the mining tool presented. From every accident type that was extracted by the mining tool, accidents involving 3 or more vehicles and temporary concrete barriers occurred the most and the same number of times. Given that [BARRI] or temporary concrete barrier is a temporary traffic control device regularly used in work zones, it would be interesting for transportation officials to learn which device is being impacted more often by moving vehicles.

Item Set	Support	I
[3+VEH]	50.0000%	
[BARRI]	50.0000%	
[3+VEH]+[BARRI]	25.0000%	

Figure 5.12 Support of accidents that occurred during a work zone

The type of information presented in Figure 5.12 could also be obtained from conducting a query in the spreadsheet containing the working database. However, the relationship between either of these accident types ([3+VEH] or [BARRI]) with other accident types that have occurred on the segment of highway with work zone activity could not be obtained by querying the data in a spreadsheet. The patterns extracted among the different accident types would help transportation officials conduct further investigations that could lead to decreasing the accident rate on work zones.

The approach allowed answering a series of questions, such as:

- ✓ What accident types are occurring the most during work zone activities on weekdays?
- ✓ What particular vehicle accident types could be prevented, on the operational portion of the highway, if better safety measures are taken on the work zones?
- ✓ What temporary traffic control device is being impacted most often by vehicles, on the operational portion of the highway, during work zone activities?

Still, the mining tool could also be used to compare the accidents during work zone activities with accidents during no work zone activity. Figure 5.13 shows such a comparison. Six (6) accident types were extracted in each analysis, only two (2) were common on both scenarios: 3 or more vehicles ([3+VEH]) and with temporary concrete barrier ([BARRI]). The fact that accident type [BARRI] or temporary concrete barrier was extracted in the no work zone activities scenario, could indicate that such traffic control device was not removed after the construction work was completed. To transportation officials, this simple analysis could provide enough information for decision-making purposes. Furthermore, the simple removal of a temporary concrete barrier that is no longer in use would improve the accident rate and the operational conditions in the segment of highway that was analyzed.

Work Zone Activities		NO Work Zone Activities	
Item Set	Support	Item Set	Support
[3+VEH]	50.0000%	[2_VEH]	100.0000%
[BARRI]	50.0000%	[H_&_R]	100.0000%
[PHOLE]	25.0000%	[WTREE]	100.0000%
[W_BRD]	25.0000%	[FIXOB]	100.0000%
[DRUMS]	25.0000%	[BARRI]	100.0000%
[PKVEH]	25.0000%	[3+VEH]	100.0000%

Figure 5.13 Accident types during work zone and no work zone activity

Further information on accidents with a tree (depicted as [WTREE]) should be conducted to discard that the tree(s) that these drivers are impacting has(have) the clearance from the side of the road required by the latest ASSHTO regulations. Figure 5.14 lists the rules extracted for these relationships. It appears as though there is a pattern between the events that led to accidents with a fixed object, (depicted as [FIXOB]), and the events that led to accidents with a tree. Given that the data is from the same site, it would be easy to conduct field investigations that could lead to the reduction of these two types of accidents.

Rule	Support	Confidence
[BARRI]+[FIXOB]+[2_VEH]+[H_&_R]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[FIXOB]+[2_VEH]+[H_&_R] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[FIXOB]+[2_VEH]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[FIXOB]+[H_&_R]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[FIXOB]+[2_VEH]+[H_&_R]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[2_VEH]+[H_&_R]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[FIXOB]+[2_VEH]+[H_&_R] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[FIXOB]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[FIXOB]+[H_&_R]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[2_VEH]+[H_&_R]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[2_VEH]+[H_&_R] ==> [WTREE]	100.0000%	100.0000%
[FIXOB]+[2_VEH]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[FIXOB]+[H_&_R] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[2_VEH]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[H_&_R]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[FIXOB]+[2_VEH] ==> [WTREE]	100.0000%	100.0000%
[2_VEH]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[2_VEH] ==> [WTREE]	100.0000%	100.0000%
[H_&_R]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[FIXOB]+[H_&_R] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[FIXOB] ==> [WTREE]	100.0000%	100.0000%
[FIXOB]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[H_&_R] ==> [WTREE]	100.0000%	100.0000%
[FIXOB]+[2_VEH] ==> [WTREE]	100.0000%	100.0000%
[BARRI]+[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[2_VEH]+[H_&_R] ==> [WTREE]	100.0000%	100.0000%
[BARRI] ==> [WTREE]	100.0000%	100.0000%
[3+VEH] ==> [WTREE]	100.0000%	100.0000%
[FIXOB] ==> [WTREE]	100.0000%	100.0000%
[2_VEH] ==> [WTREE]	100.0000%	100.0000%
[H_&_R] ==> [WTREE]	100.0000%	100.0000%

Figure 5.14 Relationship of accidents during no work zone activity with a tree

These examples could be translated into a larger scale project, for a typical transportation agency, in which work zone and accident data from different site locations could be analyzed by means of data mining to make a priority list of the accidents that would be managed first depending on the locations with the worst accident rates.

5.5 Study V – Association Mining to Analyze Weekday LOS

In this study, association mining was used to analyze the LOS on weekdays. Given that the LOS' constantly vary depending on the average vehicle speeds and the flow rates, different scenarios were examined to evaluate the information that could be obtained from mining the data. The analyses considered the relationships between LOS and months because it is known that traffic patterns are seasonal.

The first scenario evaluated the relationship between months and LOS on different time periods in each direction of travel (north- and southbound). The data was queried to only include the records that were consistent to the particular direction (north- or southbound) and the records that were within a particular time period (for example, 5:45AM to 9:30AM). Figure 5.15 lists a portion of the rules extracted for the first time period.

Southbound			Northbound		
Rule	▼ Support	Confidence	Rule	▼ Support	Confidence
[B] ==> [E]	82.7586%	100.0000%	[F]+[B]+[C]+[E] ==> [A]	85.7143%	100.0000%
[B] ==> [C]	82.7586%	100.0000%	[F]+[B]+[C]+[A] ==> [E]	85.7143%	100.0000%
[B] ==> [A]	82.7586%	100.0000%	[F]+[C]+[E]+[A] ==> [B]	85.7143%	100.0000%
[B] ==> [F]	82.7586%	100.0000%	[F]+[B]+[E]+[A] ==> [C]	85.7143%	100.0000%
[C] ==> [E]	82.7586%	100.0000%	[B]+[C]+[E]+[A] ==> [F]	85.7143%	100.0000%
[C] ==> [B]	82.7586%	100.0000%	[F]+[C]+[A] ==> [E]	85.7143%	100.0000%
[C] ==> [A]	82.7586%	100.0000%	[F]+[C]+[A] ==> [B]	85.7143%	100.0000%
[C] ==> [F]	82.7586%	100.0000%	[F]+[B]+[E] ==> [A]	85.7143%	100.0000%
[A] ==> [E]	82.7586%	100.0000%	[F]+[B]+[E] ==> [C]	85.7143%	100.0000%
[A] ==> [C]	82.7586%	100.0000%	[F]+[B]+[A] ==> [E]	85.7143%	100.0000%
[A] ==> [B]	82.7586%	100.0000%	[F]+[B]+[A] ==> [C]	85.7143%	100.0000%
[A] ==> [F]	82.7586%	100.0000%	[B]+[C]+[A] ==> [E]	85.7143%	100.0000%
[E] ==> [C]	82.7586%	100.0000%	[B]+[C]+[A] ==> [F]	85.7143%	100.0000%
[E] ==> [B]	82.7586%	100.0000%	[C]+[E]+[A] ==> [B]	85.7143%	100.0000%
[E] ==> [A]	82.7586%	100.0000%	[C]+[E]+[A] ==> [F]	85.7143%	100.0000%

Figure 5.15 Rules comparison south- and northbound 5:45AM - 9:30AM

For example purposes, these rules were sorted to start with the highest confidence and support percentages. Each LOS (i.e., A, B, C, E, and F) extracted was present in the IF- and THEN- statement indicating that the presence of either LOS will lead to the presence of the other, thus this was not a pattern. A particular pattern was observed; the southbound direction rules only have one IF- statement, while the northbound direction rules have multiple IF- statements. A person not familiar with this highway could actually tell that the northbound direction has the highest flow rates during this time period. The rules extracted for the northbound contain more IF- statements because there are more flow rates during this time period in this direction and many more relationships are extracted.

Figure 5.16 shows the rules extracted for the time period between 9:30AM and 1:15PM. During this time period, a person not familiar with this highway would not be able to tell the direction of the most flow rates from observing the rules extracted. The patterns imply the beginning of the shift in the peak period (from northbound AM to southbound PM).

Southbound			Northbound		
Rule	▼ Support	Confidence	Rule	▼ Support	Confidence
[F]+[E]+[A] ==> [B]	85.1852%	100.0000%	[F]+[B]+[E] ==> [A]	91.6667%	100.0000%
[E]+[B]+[A] ==> [F]	85.1852%	100.0000%	[F]+[A]+[E] ==> [B]	91.6667%	100.0000%
[F]+[E]+[B] ==> [A]	85.1852%	100.0000%	[F]+[B]+[A] ==> [E]	91.6667%	100.0000%
[F]+[B]+[A] ==> [E]	85.1852%	100.0000%	[B]+[A]+[E] ==> [F]	91.6667%	100.0000%
[F]+[E] ==> [A]	85.1852%	100.0000%	[A]+[E] ==> [B]	91.6667%	100.0000%
[F]+[E] ==> [B]	85.1852%	100.0000%	[A]+[E] ==> [F]	91.6667%	100.0000%
[E]+[A] ==> [B]	85.1852%	100.0000%	[F]+[A] ==> [E]	91.6667%	100.0000%
[E]+[A] ==> [F]	85.1852%	100.0000%	[F]+[A] ==> [B]	91.6667%	100.0000%
[F]+[A] ==> [B]	85.1852%	100.0000%	[B]+[E] ==> [A]	91.6667%	100.0000%
[F]+[A] ==> [E]	85.1852%	100.0000%	[B]+[E] ==> [F]	91.6667%	100.0000%
[E]+[B] ==> [A]	85.1852%	100.0000%	[B]+[A] ==> [E]	91.6667%	100.0000%
[E]+[B] ==> [F]	85.1852%	100.0000%	[B]+[A] ==> [F]	91.6667%	100.0000%
[F]+[B] ==> [A]	85.1852%	100.0000%	[F]+[B] ==> [E]	91.6667%	100.0000%
[F]+[B] ==> [E]	85.1852%	100.0000%	[F]+[B] ==> [A]	91.6667%	100.0000%
[B]+[A] ==> [E]	85.1852%	100.0000%	[F]+[E] ==> [A]	91.6667%	100.0000%

Figure 5.16 Rules comparison south- and northbound 9:30AM - 1:15PM

In Figure 5.17 the shift in peak periods is observed. The rules extracted for the time period between 5:30PM and 7:15PM indicate that traffic is mostly moving southbound on this highway. This is represented by the larger body of rules for the southbound as opposed to the smaller body of rules for the northbound.

Southbound			Northbound		
Rule	▼ Support	Confidence	Rule	▼ Support	Confidence
[F]+[A]+[B]+[C] ==> [E]	82.7586%	100.0000%	[F]+[C]+[A] ==> [E]	92.3077%	100.0000%
[A]+[B]+[C]+[E] ==> [F]	82.7586%	100.0000%	[F]+[C]+[E] ==> [A]	92.3077%	100.0000%
[F]+[B]+[C]+[E] ==> [A]	82.7586%	100.0000%	[F]+[E]+[A] ==> [C]	92.3077%	100.0000%
[F]+[A]+[C]+[E] ==> [B]	82.7586%	100.0000%	[C]+[E]+[A] ==> [F]	92.3077%	100.0000%
[F]+[A]+[B]+[E] ==> [C]	82.7586%	100.0000%	[F]+[A] ==> [E]	92.3077%	100.0000%
[F]+[A]+[C] ==> [E]	82.7586%	100.0000%	[C]+[E] ==> [A]	92.3077%	100.0000%
[F]+[A]+[C] ==> [B]	82.7586%	100.0000%	[F]+[A] ==> [C]	92.3077%	100.0000%
[F]+[A]+[B] ==> [E]	82.7586%	100.0000%	[C]+[E] ==> [F]	92.3077%	100.0000%
[F]+[A]+[B] ==> [C]	82.7586%	100.0000%	[F]+[E] ==> [A]	92.3077%	100.0000%
[A]+[B]+[C] ==> [E]	82.7586%	100.0000%	[F]+[E] ==> [C]	92.3077%	100.0000%
[A]+[B]+[C] ==> [F]	82.7586%	100.0000%	[E]+[A] ==> [C]	92.3077%	100.0000%
[F]+[B]+[E] ==> [C]	82.7586%	100.0000%	[E]+[A] ==> [F]	92.3077%	100.0000%
[F]+[B]+[E] ==> [A]	82.7586%	100.0000%	[F]+[C] ==> [A]	92.3077%	100.0000%
[A]+[B]+[E] ==> [C]	82.7586%	100.0000%	[F]+[C] ==> [E]	92.3077%	100.0000%
[A]+[B]+[E] ==> [F]	82.7586%	100.0000%	[C]+[A] ==> [E]	92.3077%	100.0000%

Figure 5.17 Rules comparison south- and northbound 5:30PM - 7:15PM

The patterns extracted allowed the most flow rates to be matched to the direction of travel on PR-18 on early and late hours of the day. In addition, from the rules extracted for midday hours, the shift in peak period was observed. Thus, the use of association mining to evaluate the LOS per time period on both direction of travel was achieved.

The use of association mining could also answer questions, such as:

- ✓ What relationships can be identified from the worst LOS of each weekday during work zone activities?
- ✓ What day of the week presents the largest change in LOS?
- ✓ What are the most common LOS per day of the week?
- ✓ What LOS are observed the most during work zone activities?

The following scenarios describe the application of association mining to answer these questions.

Figure 5.18 lists the rules extracted for the worst LOS during work zone activities. The most obvious pattern extracted was Tuesday presenting the worst LOS during work zone activities in the THEN- portion of the rules (with LOS F). For every other day of the week, the worst LOS in the THEN- portion of the rules was LOS E. The mining tool extracted the relationship of events that had similar characteristics as that of LOS F for Tuesday and did not find those relationships within the dataset for any other day of the week. Thus, this pattern stands out from the rest.

Rule	Support	Confidence
[D]+[A]+[B] ==> [F]	5.6818%	100.0000%
[D]+[A]+[C]+[E] ==> [F]	5.6818%	100.0000%
[D]+[A] ==> [F]	5.6818%	100.0000%
[D]+[A]+[B]+[C]+[E] ==> [F]	5.6818%	100.0000%
[D]+[A]+[B]+[C] ==> [F]	5.6818%	100.0000%
[A]+[E] ==> [F]	5.6818%	100.0000%
[D]+[B]+[C]+[E] ==> [F]	5.6818%	100.0000%
[B]+[C]+[E] ==> [F]	5.6818%	100.0000%
[A]+[B]+[C]+[E] ==> [F]	5.6818%	100.0000%
[A]+[C]+[E] ==> [F]	5.6818%	100.0000%
[D]+[B]+[C] ==> [F]	5.6818%	100.0000%
[D]+[A]+[B]+[E] ==> [F]	5.6818%	100.0000%
[D]+[B] ==> [F]	5.6818%	100.0000%
[D]+[A]+[C] ==> [F]	5.6818%	100.0000%
[B]+[E] ==> [F]	5.6818%	100.0000%
[D]+[B]+[E] ==> [F]	5.6818%	100.0000%
[A]+[B]+[E] ==> [F]	5.6818%	100.0000%
[D]+[A]+[E] ==> [F]	5.6818%	100.0000%
[D]+[C]+[E] ==> [F]	13.6364%	100.0000%
[C]+[E] ==> [F]	13.6364%	100.0000%
[D]+[C] ==> [F]	13.6364%	100.0000%
[D]+[E] ==> [F]	17.0455%	100.0000%
[D] ==> [F]	19.3182%	100.0000%
[E] ==> [F]	19.3182%	100.0000%

Figure 5.18 Worst level of service during work zone activities tuesday

The application of association mining can also be used to learn about the relationships of LOS and weekdays without considering whether there were work zone activities or accidents in the segment of highway being analyzed. The most important benefit data mining offers is the use of large amounts of data, the more data the better. Data mining application will discover more patterns from each analysis as the dataset increases. Thus, in the next scenario the relationships between months and LOS were examined. Figure 5.19 shows the LOS, as head rules, which were extracted from this analysis. Head rules are the THEN- portion of the equations.

Monday		Tuesday		Wednesday		Thursday		Friday	
Item ...	Support	Item ...	Support	Item ...	Support	Item ...	Support	Item ...	Support
[A]	80.6452%	[A]	80.0000%	[D]	77.4194%	[A]	83.3333%	[F]	86.2069%
[F]	77.4194%	[B]	80.0000%	[C]	77.4194%	[F]	83.3333%	[E]	82.7586%
[E]	77.4194%	[C]	80.0000%	[B]	77.4194%	[E]	80.0000%	[C]	82.7586%
[C]	77.4194%	[D]	80.0000%	[A]	77.4194%	[C]	80.0000%	[B]	82.7586%
[B]	77.4194%	[F]	80.0000%	[F]	77.4194%	[B]	80.0000%	[A]	82.7586%

Figure 5.19 LOS as head rules

A pattern was observed from the LOS, as head rules, for Tuesday and Wednesday and another for Monday, Thursday, and Friday. The mining tool did not extract LOS E as a head rule for Tuesday or Wednesday, and it did not extract LOS D as a head rule for Monday, Thursday, and Friday. This does not indicate that these LOS were not found in the dataset in each of these days, but rather that there were no events extracted by the mining tool indicating LOS E and D as previously mentioned. In other words, the mining tool was not able to extract the relationship of events that would indicate LOS E in the analysis of Tuesday or Wednesday, even though 55 records indicating LOS E for Tuesday

were found from querying the raw data in a spreadsheet. This type of information would be useful for transportation officials to have a rough idea of the overall operational conditions of a particular segment of highway. This was not the case with this working dataset, but other datasets could indicate that the events in a particular segment of highway do not indicate LOS A. Such a pattern would indicate that there are very few instances during the 24-hour day period in which there are low flow rates and, thus, it would be difficult for the mining tool to discover those little events. In that case, transportation officials could conduct further investigations.

Association mining was used to study the relationship of work zone activities and LOS during rainy weather. Figure 5.20 shows the most common rules extracted for each weekday model. Rule sets (a) through (e), on Figure 5.20, show the rules extracted during the peak hours and rule sets (f) through (j) show the ones extracted during the off peak hours. The patterns found were identified to provide simplicity to the description. It was found that during rainy weather Mondays and Fridays provided better operational conditions (LOS B) for drivers than Tuesdays, Wednesdays, and Thursdays (LOS F) during peak hours. The events occurring during work zone activities on rainy weather created the worst operational conditions for drivers on Tuesdays, Wednesdays, and Thursdays from 6:00AM to 6:00PM.

During the off peak hours, the events occurring during work zone activities in rainy weather created the worst operational conditions for drivers on Tuesdays and Fridays (LOS D and LOS E, respectively). The events occurring during rainy

weather did not affect the conditions of traffic on Mondays, Wednesdays, and Thursdays (LOS B) during off peak hours.

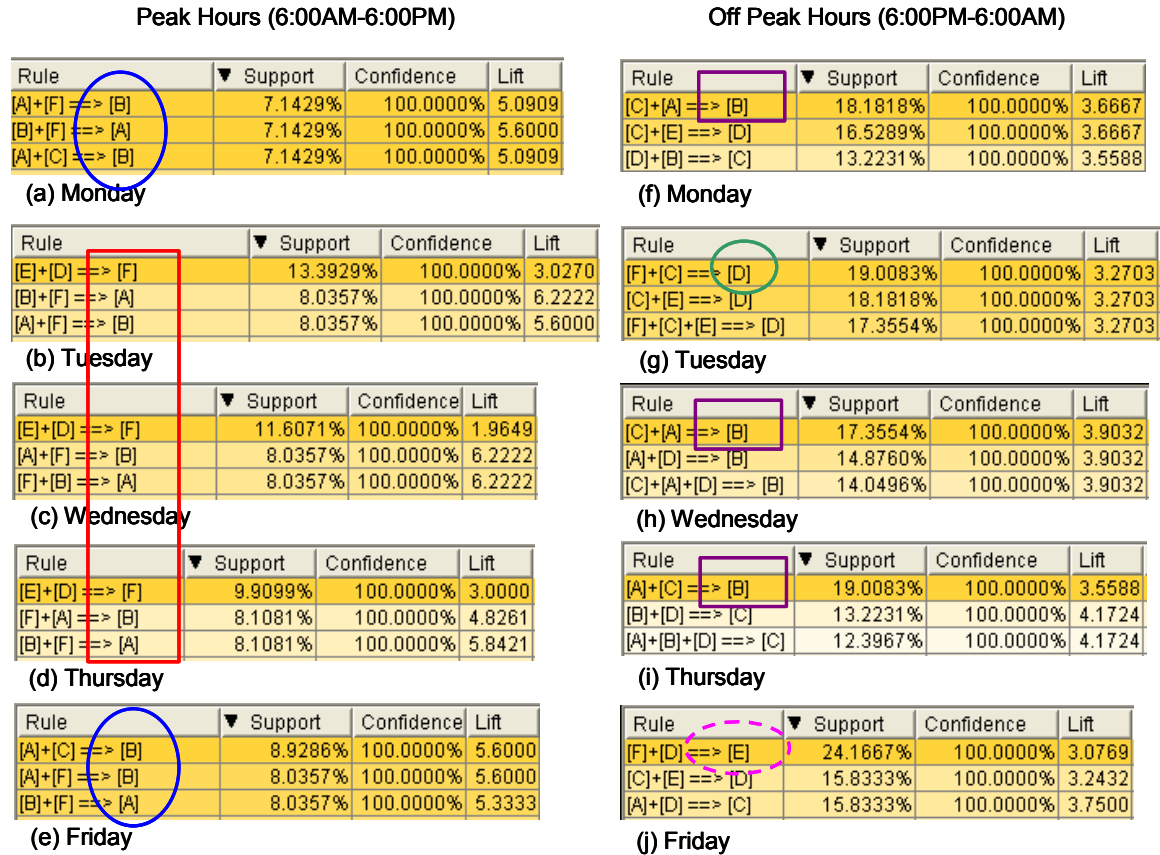


Figure 5.20 Relationship between work zone activities and LOS during rainy weather

Transportation officials could benefit from this type of information because it would allow them to study the operational conditions of traffic in a macroscopic approach. Thus, considering all the events (variables from multiple levels of information) in the working database during rainy weather and during work zone activities. Furthermore, transportation officials could use the same approach on several locations of the same highway system with the goal of improving the

conditions of traffic during these particular activities while improving the decision-making process.

5.6 Study VI – Comparison of Traditional and Data Mining Approach for Analyzing ITS Generated Data

Whether one is to use conventional methods or data mining applications to analyze ITS generated data, there are no differences in the challenges one could confront when trying to create a working database from different sources. However, once the working database is fully developed and cleaned, the advantages of using either method (conventional or data mining) depend primarily on the objectives of the project. The major benefit learned from applying data mining to ITS generated data was that it allowed the analysis of multiple variables, and, if done correctly, the analysis is completed in a matter of minutes.

In order to evaluate the benefits of using data mining, especially association mining, to analyze ITS generated data, the information obtained from the traditional methods was compared to the information obtained from the mining tool. The data for Monday were used throughout this Section in order to compare the same data on both methodologies. Figure 5.21 shows the statistical results extracted by the mining tool for the relationships between months and LOS during a 24-hour period. Notice that the number of visible rules extracted for the northbound model was 83% of those extracted for the southbound model. From examining the raw data it was found that the northbound direction

consisted of slightly more records within the majority of the categories of LOS (categories being A, B, C, D, E, F). However, the category/LOS with the fewest records (LOS D with 57 records) was not extracted in any of the rules for the northbound model. Figure 5.22 shows the LOS extracted in the THEN- portion of the rules. The mining tool does not reject a category for having too few records, such as category LOS D in the northbound direction; it extracts the patterns within the data searching for events that are similar to each other that explain other events. Thus, a category could have even fewer records, such as LOS C in the southbound direction that consisted of 33 records, and still be extracted because other events had similar characteristics to LOS C (refer to Figure 5.22).

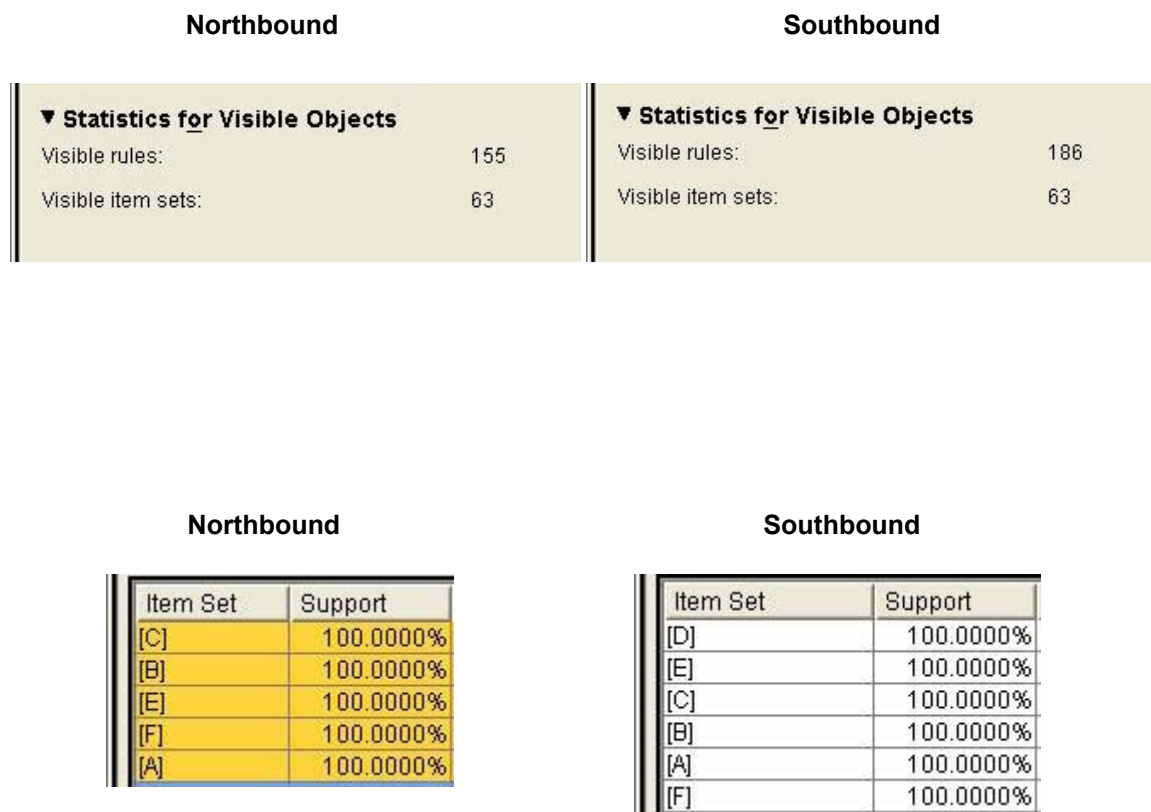


Figure 5.22 Single itemsets per model (north- and southbound)

It would be an arduous task to try to track the entire algorithm that the mining tool used in relating the records within this dataset. Still, some of the information was described to explain the discovery process used by the tool. The basic information provided in Figure 5.22 indicates that it was not likely that any event(s) could cause LOS D in the 24-hour period in the northbound direction on Mondays. However, in the southbound direction there were events that indicated the presence of every LOS at some point of the 24-hour period. This type of information could be useful to provide transportation officials with a general idea of the expected conditions of traffic on a particular highway on any day of the week.

Figures 5.23 and 5.24 illustrate the density, flow rates, and speed versus time graphs for the north- and southbound direction, respectively. The information obtained from these traditional graphs provides the conditions of traffic in terms of density, flow rates, and average speed of vehicles per hour of the day. For example, the density could be examined according to the limits established by HCM (2000) to determine a LOS (refer to Table 2.5). From the graph of the relationship of flow rates and time, the morning and afternoon peak periods of the traffic would be known. This type of information cannot be obtained by the application of data mining as it does not provide definite answers but a set of association rules that can be interpreted to find patterns that would lead to new information from the dataset.

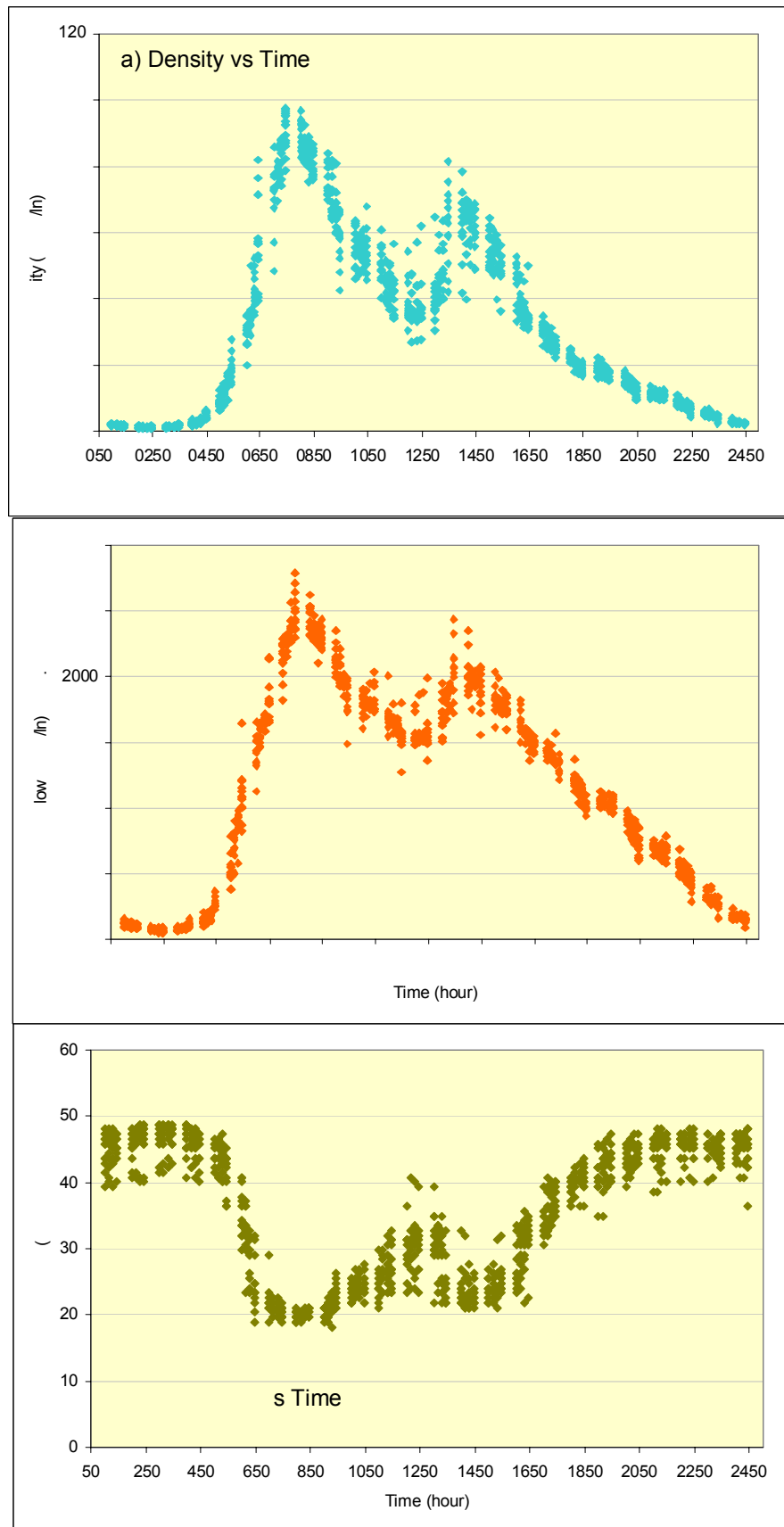


Figure 5.23 Monday northbound density, flow, and speed graphs

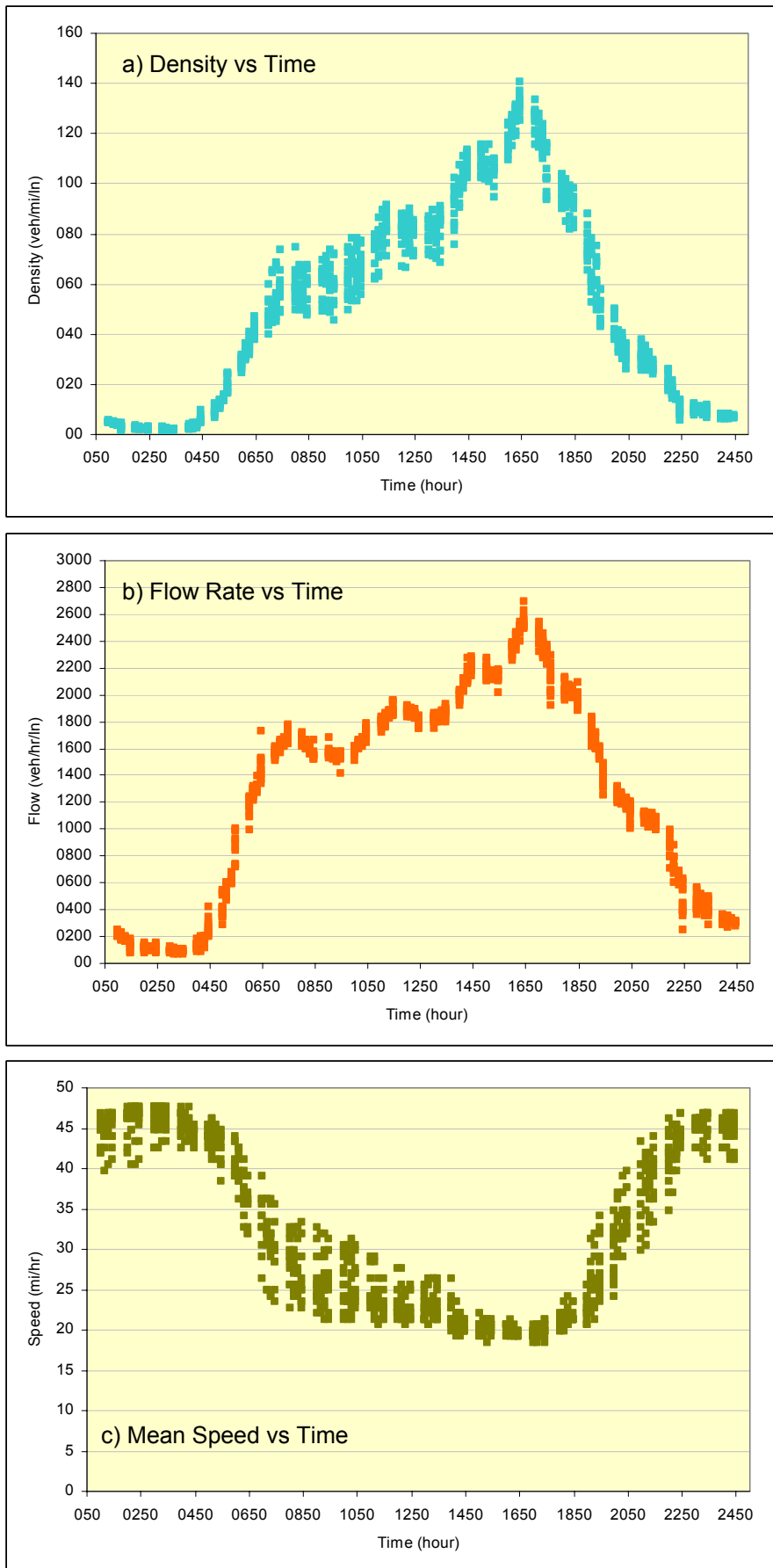


Figure 5.24 Monday southbound density, flow, and speed graphs

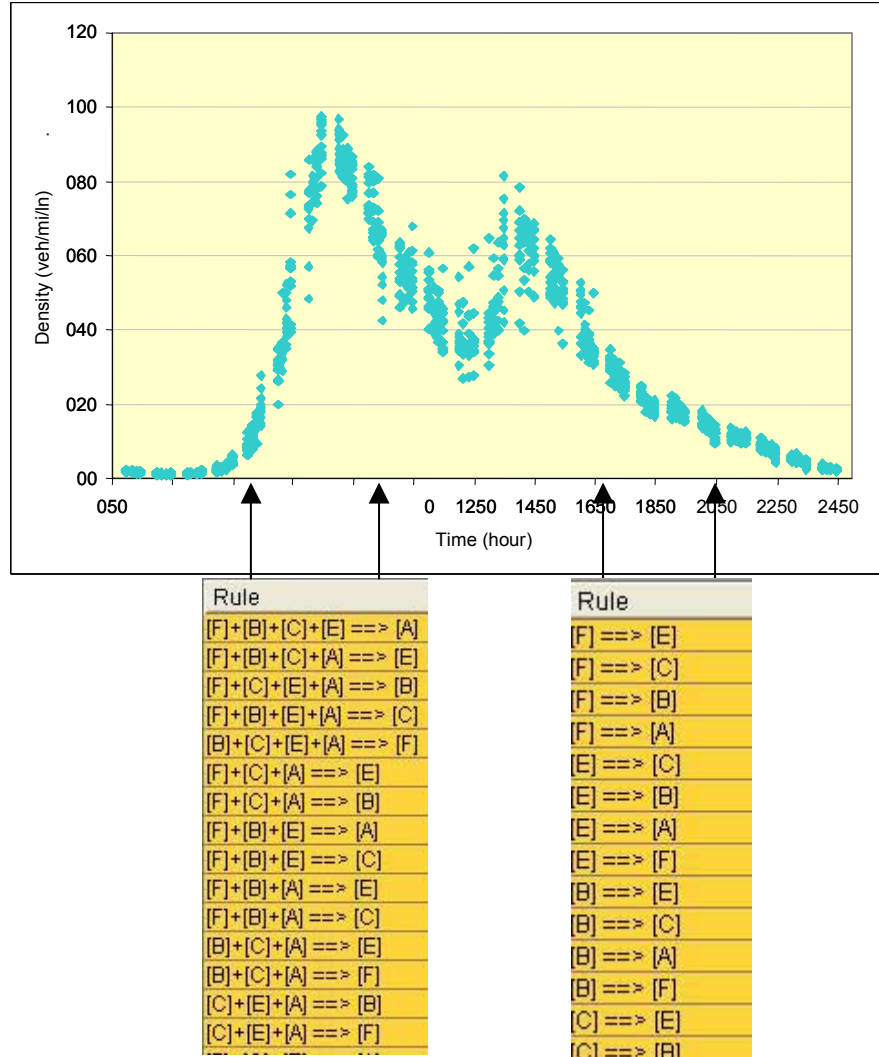


Figure 5.25 Combination of methodologies – northbound direction

Figure 5.25 illustrates the combination of the methodologies. The upper portion of the figure presents the density-time relationship and the lower portion presents a sample of the rules extracted for the time periods indicated. The pattern observed from the rules extracted indicates that the body of the rules is, in a way, related to the performance condition of the highway. For example, during the morning peak period where more vehicles were traveling northbound

more rules were extracted as opposed to the period in the afternoon (past the peak period as indicated in Figure 5.25) in which fewer rules were extracted. There were 186 and 151 rules extracted for the morning and afternoon periods, respectively.

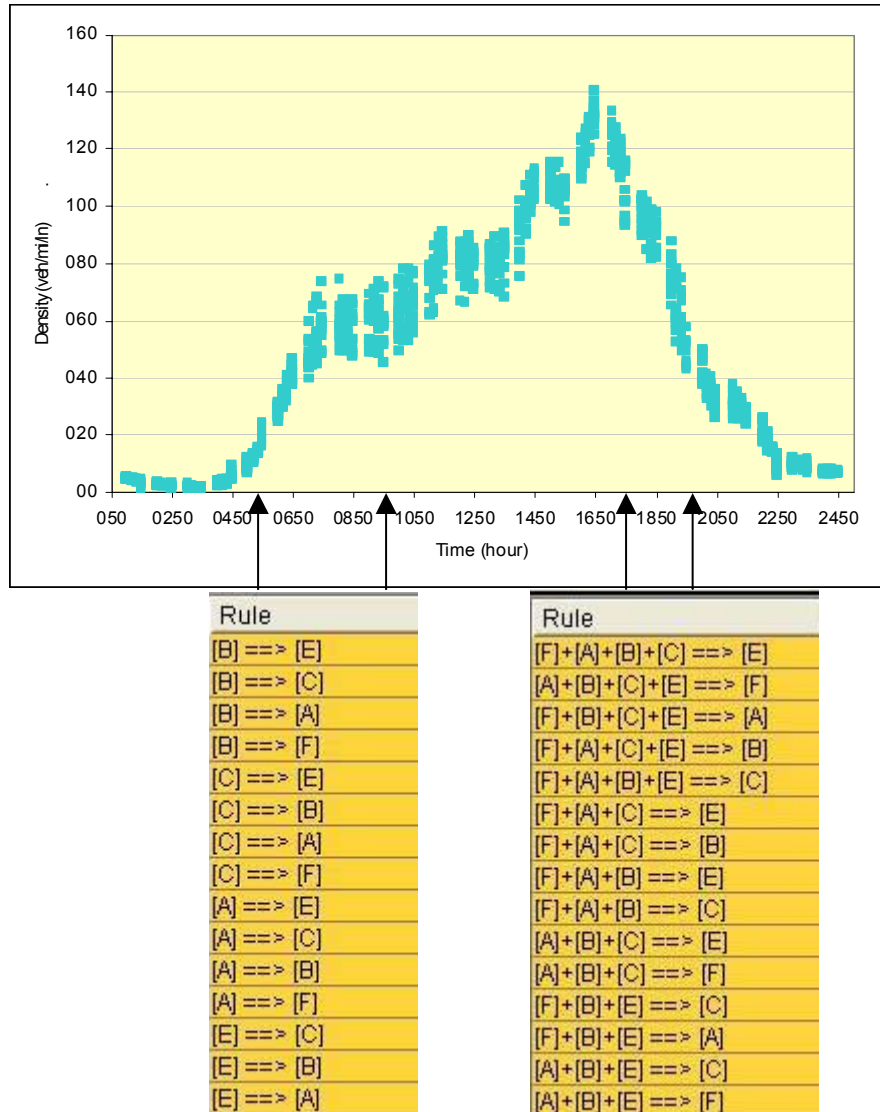


Figure 5.26 illustrates the combination of methodologies for the southbound direction. Again, a portion of the rules extracted for each time period was shown to demonstrate the patterns. There were smaller rules for the time period in which fewer vehicles were traveling southbound and larger rules for the time period in which more vehicles were traveling southbound.

This approach provides transportation officials with a different tool to examine the ITS generated data. The mining tool would extract the most recurrent LOS due to specific events, providing a general idea of the expected conditions of traffic for that particular time period. Association mining could not replace the traditional methods for analyzing ITS generated data, rather it allows the user to learn some of the hidden patterns that the data holds. In the same way, it benefits the data analysis process as more information is being learned from the data itself.

6. CONCLUSIONS AND RECOMMENDATIONS

The application of association mining to gain information from the archived ITS generated data was presented in this document. Several specific studies were conducted to test the applicability of data mining, especially association mining, as an analysis method. A working database was constructed from various datasets received from the PRHTA and the NCDC. The case study was first chosen based on the high population density of the major metropolitan areas within each state and US territory of the US to narrow the possible state agencies that would be targeted. The chosen case study was based upon data availability.

The combined data received from the PRHTA and the NCDC consisted of 133,000 records of different formats, from which approximately 90% were received as hard copies. Most of the data did not include metadata, thus in order to have a complete understanding of each dataset a number of personal- and phone-meetings were held with personnel from the PRHTA to discuss the meanings, data ranges, special characteristics, local rules of thumb, equipment used for the collection of data, etc. of each dataset provided. Some datasets required more effort to obtain than others. The work zones dataset, for example, had to be written down by hand during several visits to the agency due to the oversize books in which these are stored. The size of these books made its reproduction impossible. Nonetheless, even though the study was limited to 1-mile stretch of highway (i.e., data that was available), the process of preparing

each dataset and creating the robust working database was a crucial part of the research.

This research has shown that data mining can provide information on a dataset or database that would not likely be found through the use of general statistical analysis only and potentially provide a source of valuable information that could not have been detected otherwise. For example, the identification of “red flags” during work zone operations; the similar patterns in LOS between Tuesdays and Wednesdays and similar patterns in LOS between Mondays, Thursdays, and Fridays; and the analysis of LOS over time. Furthermore, the results imply that the rules derived from the original dataset could be applied to support decision-making. The method can extract information faster than any other analysis method and can reveal the relation of variables that are not evident to the naked eye. The major benefit learned from applying data mining to ITS generated data was that it allowed the analysis of multiple variables, and, if done right, the analysis is completed in a matter of minutes. The use of the KDD process facilitated the replication of previous steps; it was used as a guide throughout the research.

The quality of traffic flow was measured through the specific analyses conducted. However, the method gave better results when variables with fewer categories (such as accidents, work zones, and LOS) were used. Another general observation was that more information was obtained in the form of rules extracted when fewer initial conditions were used.

Other general lessons learned include:

- ✓ In order to have a well-balanced schedule for the research, in which most of the efforts are allocated in the evaluation of the models, the data should have an ASCII-text format for the user to use an ODBC to transform the data into the IBM Intelligent Miner for Data.
- ✓ The use of hard copied data should be avoided because it hinders the benefits of using data mining and the KDD process as an analysis method.
- ✓ The use of smaller minimum support percentages allowed the mining tool to extract more significant rules. On the contrary, the use of higher minimum support percentages caused the mining tool to extract large amounts of insignificant rules.
- ✓ A single case study can contribute to local and national knowledge. In an ideal world, every state or local agency would have complete sets of historical data for each highway within their highway network. However, the high costs for data collection and management prevent agencies from collecting more data.
- ✓ The initial applications of the data mining tool can yield the needs of additional data (many time the data is not available but can be obtained).
- ✓ Greater use of the data collected will improve the quality of highway performance. Some state agencies are reluctant to collect data about vital highways within their network because, aside from the operating costs involved, they do not know what to do with the data other than storing it.

However, during the data mining process it is learned quickly what other types of data are needed and in what format the data should be collected and stored. The value of “good” data is learned as soon as the amount of missing data surpasses the amount of complete or clean data. There will always be missing data and erroneous data given that machines malfunction occasionally and/or people misread numbers or values, but the local and state agencies use data, the more complete sets of data they will eventually have in their databases.

- ✓ Spreadsheets can be used easily for manipulating the data and performing quality control checks on the data. However, data should be stored in ASCII-text files for the more advanced analyses to be performed, such as data mining applications.
- ✓ One of the benefits of using data mining is the extraction of patterns from the data. These patterns can be displayed and understood by the average person.
- ✓ Data mining and the KDD process allow the inter-relationship of variables from multiple levels of information. The working database used in a data mining analysis could contain data from various databases as long as it is maintained as a relational database.
- ✓ If the working database consisted of numerical data only, one could attempt to use traditional statistical analysis and compare the findings to those from the application of data mining. The drawbacks from using traditional statistical analysis would be that the user would have to know

exactly what fields and from what datasets these fields should be used in the analysis. The data mining algorithms, on the other hand, examine every record in every field within the working database searching for patterns and extract as many patterns as it can find. Thus, a greater effort would be needed if traditional statistical analysis were to be used for the same objectives.

- ✓ Data mining allows one to work with the data in its rawest form and present it in its most fancy form. In fact, it is during the mining process itself when information is gathered and the new knowledge is reached when the process is completed. The multiple information including statistics, mining tools, and patterns found in tables of data and graphical views provide a great deal of benefit for the data mining user. The more information that is obtained about a certain highway the greater understanding one has about the operational performance of that facility.

For this case, the association mining was able to extract hundreds of rules that led to new information about the dataset used in this research. Some of the information that was learned through the application of association mining included:

- ✓ During work zone activities, the motorists traveling southbound were most likely to experience worse conditions in traffic than the motorists traveling northbound given the same conditions. This information was interpreted from the number and size of the rules extracted for the analyses of density data for the south- and northbound directions.

- ✓ The mining tool was able to identify the most common types of accidents in the northbound direction of the 1-mile segment of highway examined. It was learned that accidents with temporary concrete barriers and accidents between 2 vehicles were the most common. The raw data was queried to validated the information learned.
- ✓ The mining tool was able to provide the most common types of accidents in the northbound direction during work zone activities; these were accidents between 3 or more vehicles and accidents with temporary concrete barriers.
- ✓ The day(s) of the week on which more accidents occurred was also obtained through the association rules. The weekdays with the most accidents were Monday and Friday; this information was validated with the raw data using a spreadsheet.
- ✓ The application was found useful for the identification of “red flags.” For example, accidents with parked vehicles appear to be related to other types of accidents (e.g., drums, temporary concrete barrier, 3 or more vehicles, and pothole) during work zone activities. The number of parked vehicle accidents in the raw data was not a crucial factor for the data mining application as its goal is to extract patterns within the events that led to this type of accident. However, if traditional statistical analysis were to be used, this type of data (categorical data, such as accident types) would have to be numerically coded, and it would require lots of accidents in order for the analysis to provide valuable information. Traditional

statistical analysis does not conduct the intra-relationship of variables that the data mining application performs.

- ✓ The application was also useful in the analysis of LOS. The hundreds of rules developed for the LOS of each weekday pointed to a particular pattern. A portion of the rules developed for the Tuesday model contained LOS F in the THEN- portion of the rules. Such information indicates that Tuesday was the weekday with the worst LOS for the 1-mile of highway that was examined. This type of information would have been difficult, if not impossible, to obtain through the application of traditional statistical methods. The application of traditional statistical analysis would have provided the number of times in which LOS F appeared on each weekday model, but not the patterns relating the LOS to the rest of the variables in the dataset.
- ✓ After analyzing the LOS for each weekday (Monday, Tuesday, Wednesday, Thursday, and Friday) it was learned that Tuesdays and Wednesdays presented similar patterns and Monday, Thursday, and Friday presented similar patterns. The relationship of LOS to other variables in the dataset would have been extremely difficult to identify through the application of statistical analysis alone.
- ✓ The application allowed the analysis of associations between work zone activities and LOS during rainy weather. It was learned that during peak hours (6:00AM-6:00PM) Mondays and Fridays presented the best operational conditions for drivers, while the rest of the weekdays

presented the worst conditions. During off peak hours (6:00PM-6:00AM), Tuesdays and Fridays presented the worst operational conditions. This type of analysis, in which categorical data is being analyzed in combination to numerical data, is best performed using data mining algorithms. These algorithms relate variables from different levels of information that would be extremely difficult to perform with statistical analysis alone.

- ✓ The use of the application allowed the analysis of LOS over time. The information obtained from these analyses allowed the identification of the shift in the peak period. In addition, the direction of traffic was identified from the association rules extracted for the analyses of early morning and late afternoon. This type of information was seen to complement the traditional methods for analyzing ITS generated data (refer to Section 5.6).

Given that the application of association mining was able to identify red flags during work zone activities, temporary control devices being impacted by vehicles, most common accidents, and the day of the week with the worst LOS, the approach can be used to improve safety on work zones. The information learned could be used by the agency to improve safety on the segment of highway that was analyzed through the use of variable message boards, additional warning/regulatory signs, and maintenance and management decisions. New regulations could also arise from the information learned that could be implemented for work zone operations. Thus, the agency could

improve work zone operations for the benefit of the safety of drivers and the construction workers.

The approach could also be used by the agency to improve the efficiency of managing accidents during work zone operations and providing the necessary warning signs and/or variable message boards during work zone operations and expected bad weather conditions.

The information obtained through the application of data mining can assist traffic engineers in understanding the relationship between variables. The tool has been demonstrated to be useful for the analysis of very large databases, such as the databases contained in most local and state agencies. The visualization tool allowed the identification of hidden relationships between two specific variables. It was an essential part of the KDD process, making the process much faster by allowing a quick view of the results. The tool performed well on both personal and on network operational computers, providing the benefit of allowing the analysis of data on remote locations when needed.

Although the application of data mining or the extraction of patterns from large datasets and the KDD process, which encompasses the whole process of extraction of knowledge, were successfully used in this research, there are some drawbacks.

- ✓ It will require considerable staff time. Due to the numerous details involving the collection of data, preparation and processing of the data, creating the work database, running the algorithms, examining the rules extracted for meaningfulness, repeating any intermediate step, discussing

the results with peers, and making the conclusions, this would require a full-time commitment. Thus, it would require state agencies to hire or train skilled personnel for these tasks alone.

- ✓ It requires an extensive learning process. Due to the complexity of the algorithms, it would require state agencies to provide the proper training and time before the technical personnel could actually conduct a full study.
- ✓ It does not develop particular documentation for each project. The KDD is a step by step process that allows users to follow and track their steps for replication purposes, but it is up to the user to document each step and the assumptions made in order to be able to replicate the exact project or perform future projects.
- ✓ It returns far too many rules. Depending on the type of analysis being conducted, the application returns a huge number of rules that could be overwhelming for a person to understand. There were many analyses conducted throughout this research from which the most reasonable rules were for those fields (variables) that contained the fewest number of classes (itemsets).
- ✓ It does not indicate whether the dataset is large enough for these types of applications. These types of algorithms (data mining) provide the best results when very large datasets are used, however what constitutes a “very large dataset”? The user must be aware of the size of the dataset being used in the research prior to selecting any of the data mining applications. There has been some research work conducted in which the

relationship of the accuracy of sampling models and the accuracy of models using all available data have been studied, but these have been with data from consumer oriented databases (Carey et al. 2003; Zaki et al. 1996; Kotsiantis and Kanellopoulos 2006; Toivonen 1996).

- ✓ It does not provide one definite answer. Given that the application used in this research extracts rules from the working dataset, each time the application is run for a new set of queries; these rules will have different numbers and will have different sizes. Thus, it is up to the user to interpret each set of rules extracted by the application for meaningfulness that will lead to the new knowledge.

The research attempted to understand the quality of traffic flow on expressway facilities considering work zone operations and accidents in combination with the more traditional ITS generated data. It will be likely that other more advanced methods will be developed based upon this one. Knowing that substantial research remains to be done in the area of knowledge discovery in the traffic/transportation domain, possibilities for building on the work described in this document include:

- ✓ Researching the area of data needs for acquiring high accuracy from sampling models and models using all available data with transportation generated data. The findings from this type of work could be of great benefit for researchers using the approach described in this document and researchers attempting to apply any other data mining technique (for

example clustering and classification mining) to transportation generated data.

- ✓ Increasing the size of the dataset by increasing the number of traffic count stations. Integrating data from a series of traffic count stations on the same highway system could yield specific trends in the traffic conditions along that highway. This information could allow traffic engineers to compare the rules extracted per count station and possibly find patterns from the rules for some of the count stations.
- ✓ Incorporating photo images. One of the great benefits of using data mining is the ability to extract patterns from different levels of information. It would be interesting to have the extraction of association rules and the photo images representing those patterns.
- ✓ Conducting a before and after study from a highway that is pending a complete reconstruction (e.g., the addition of one or two lanes, addition of an intersection/interchange, addition of an exit or entrance ramp). This would be an extremely extensive study because the reconstruction would have to be finished and fully operational before the after data can be collected. Nonetheless, it would be an opportunity to apply data mining to extract the patterns from the traffic conditions before and after the reconstruction. The information learned from the rules extracted could incorporate the traditional traffic study that would be developed as part of the reconstruction project.

- ✓ Incorporating vehicle classification data. The monitoring of freight is part of the efforts of state agencies. Therefore, the application of data mining to a large dataset that has vehicle classification in addition to the variables used in this research could provide the opportunity to examine the trends extracted from mining different vehicle classes (e.g., auto, trucks, SUV's, semi trailers). The rules extracted from using the vehicle classification related to the transportation of goods could provide state agencies with additional information that could lead to the improvement of the decision-making process.
- ✓ Incorporating GIS data. By incorporating geo-spatial data into the dataset and also expanding the study area in which data from various traffic count stations from several nearby highways are collected, a better assessment of the types of accidents (i.e., fatal, injury, or PDO) that are occurring within an area of the city could be performed. Geo-spatial data from work zone operations, given the previous data conditions, could also be useful for state agencies. The trends extracted from the relationship of work zones and the other variables in the dataset using geo-spatial data could allow traffic engineers identify the highways that are receiving continuous maintenance and highways that are not receiving maintenance and portray that information in a base map for management purposes. Such display of information could improve the decision-making process regarding the scheduling of maintenance activities around the city.

The use of archived data can lead to new knowledge gain by means of the methodology approach presented in this research. The use of this data will increase in transportation agencies as more information is learned from the patterns within the data. This knowledge may improve the performance of the TMC or transportation agency from which the data is being conducted.

APPENDIX A: GLOSSARY AND LIST OF ACRONYMS

As explained in Section 1.4, this research combines approaches and terminology from several paradigms, including traffic engineering, software engineering, statistics, and computer science. The glossary is intended to provide a common point of reference for the terminology used in the document. A list of acronyms is also included (OHPI 2000).

A1. GLOSSARY

association – creates rules that describe how often events have occurred together

attribute – specifies the name of each field or column

attribute reduction – removal of attributes that may not have significant contribution to the analysis

categorical data – data that fit into a small number of discrete categories

cleaning – preparing data for a data mining activity, in which “obvious data errors are detected and corrected and missing data is replaced” (Pilot Software 2000)

confidence – measure of how much likely it is that B occurs when A has occurred, also called “condition probability” (Two Crows 1999)

data – “values collected through record keeping, observing, or measuring, typically organized for analysis or decision making” (Two Crow 1999)

database – collection of data that is organized so that its contents can easily be accessed, managed, and updated

dataset – arrangement of data to be used in a mining function

evaluation – “to determine the significance, worth, or condition of usually by careful appraisal and study” (Merriam-Webster’s 2000)

field – synonymous to attribute or column

flat file – “list or table of items, file that has no hierarchical structure” (IBM 1996)

hash tree – type of data structure which contains a tree of summary information about a larger piece of data, for example a file used to verify its contents

information – knowledge obtained from investigation, study, or instruction

itemsets – list of events/classes occurred at a reference time point t

knowledge – information and principles acquire by humankind

lift – factor by which the confidence exceeds the expected confidence; a value >1 indicates that the rule may be interesting in some sense as the rule occurs more frequent

metadata – data about data; high-level data that describe low-level data; information describing the data; describe the attributes and contents of an original document or work

missing data – data that were not measured, not answered, were unknown, or were lost

model – a function of the data mining process; there are two types of models, descriptive and predictive

noisy data – data that contain errors such as missing or incorrect values

numerical fields – fields in which values can be used to formulate computations

pattern – relationship between two variables

prune – elimination of lower level splits or entire sub-tree in a decision tree.

Algorithms that adjust the topology of a neural net by removing (i.e., pruning) hidden nodes.

range – difference between the maximum value and minimum value of the data

raw data – data that have not been processed, cleaned, or filtered

record – “set of one or more related data items grouped for processing” (IBM 1996); synonymous to rows

relational database – collection of data items organized as a set of tables where a primary key comprises a single column or set of columns

sampling – creating a subset of data from the whole

speed – *computer field*: refers to the computational costs involved in generating and using the model; *transportation field*: relative rate of motion or progress

support – number of times an item appears on the model; item frequency

variables – values that can change depending on conditions or on information passed to the program

visualization – graphical display of data that facilitate better understanding of its meaning

A2. LIST OF ACRONYMS

AADT – Average Annual Daily Traffic

ADR – Automatic Data Recorder

ADT – Average Daily Traffic

ADUS – Archived Data User Service

A&M – Agricultural and Mechanics

ASCII – American Standard Code for Information Interchange

ASOS – Automated Surface Observing Systems

ATIS – Advanced Traveler Information Systems

ATMS – Advanced Traffic Management Systems

AVI – Automatic Vehicle Identification

BFFS – Base Free Flow Speed

BTS – Bureau of Transportation Statistics

CAD – Computer Aided Design

CALTRANS – California Department of Transportation

CD – Compact Disc

CD-ROM – Compact Disc Read Only Memory

CG - Caguas

CVO – Commercial Vehicle Operations

DM – Data Mining

DB2 – Database 2

DOT – Department of Transportation

HPMS – Highway Performance Monitoring System

I/O – input/output

KTC – Kentucky Transportation Cabinet

PRHTA – Puerto Rico Highway and Transportation Authority

ETC – Electronic Toll Collection

FFS – Free Flow Speed

FHWA – Federal Highway Administration

GPS – Global Positioning Systems

HCM – Highway Capacity Manual

HOV – High Occupancy Vehicles

IBM – International Business Machines

ID - Identification

IRI – International Roughness Index

ITE – Institute of Transportation Engineers

ITS – Intelligent Transportation Systems

IVI – Intelligent Vehicle Initiatives

KDD – Knowledge Discovery in Databases

LCU – Local Controller Unit

LOS – Level of Service

MAGEP – Missouri Alliance for Graduate Education and the Professoriate

MPO – Metropolitan Planning Organizations

MS – Microsoft

MTC – Missouri Transportation Consortium

NORPASS – North American Preclearance and Safety System

O/D – Origin/Destination

ODBC – Open Database Connectivity

PSR – Present Serviceability Rating

SJ – San Juan

SJMR – San Juan Metropolitan Region

SJU – San Juan Luis Muñoz Marín International Airport

TMC – Traffic Management Centers

TOC – Traffic Operations Center

TRANSCOM – Transportation Operations Coordinating Committee

TRB – Transportation Research Board

TSC – Transportation Systems Center (now NTSC)

TTI – Texas Transportation Institute

TxDOT – Texas Department of Transportation

US – United States

USDOT – United States Department of Transportation

VMS – Variable Message Signs

WIM – Weigh In Motion

APPENDIX B: STATISTICAL TEST RESULTS

The statistical test results performed on the north- and southbound traffic data (i.e., numerical data) are presented in Tables B.1 and B.2, respectively. Each table consists of the number of records, type of variable, maximum, minimum, mean, standard deviation, and the total or sum of every record within each field.

Table B.1 Statistic test results from northbound PR18 SJ1999

January 1999													
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	SS
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1652	49	2782	48	2491	48	2318	48	2668	49	2635	49	1820
Min	110	22	53	20	50	20	52	20	83	20	89	20	102
Mean	784	40	1156	35	1189	34	1207	35	1292	35	1329	33	1008
Std Dev	482	8	758	9	767	10	740	11	810	11	808	11	587
Total	75283	3806	111021	3407	114138	3269	115916	3326	124039	3326	127581	3213	96810
February 1999													
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	SS
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1602	49	2531	49	2567	49	2560	49	2523	49	2584	49	1812
Min	121	31	58	20	51	20	57	20	65	19	73	20	117
Mean	758	42	1149	36	1182	35	1240	35	1285	35	1326	34	1003
Std Dev	474	6	750	11	764	11	790	11	814	11	815	11	590
Total	72745	4053	110260	3419	113465	3381	119005	3377	123320	3363	127281	3247	96241
March 1999													
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	SS
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1574	47	2531	49	2567	48	2560	48	2523	48	2584	48	1812
Min	131	30	58	20	51	20	57	20	65	20	73	19	117
Mean	770	42	1149	36	1182	35	1240	35	1285	34	1326	34	1003
Std Dev	480	5	750	10	764	10	790	10	814	11	815	11	590
Total	73891	4038	110260	3434	113465	3360	119005	3338	123320	3229	127281	3254	96241

Legend: SU (Sunday), SUS (SU mean speed), M (Monday), MS (M mean speed), T (Tuesday), TS (T mean speed), W (Wednesday), WS (W mean speed), TH (Thursday), THS (TH mean speed), F (Friday), FS (F mean speed), S (Saturday), SS (S mean speed)

Table B.1 Statistic test results from northbound PR18 SJ1999 cont.

April 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1606	48	2531	49	2620	49	2568	48	2572	47	2580	49	1865	48	48
Min	141	31	48	20	55	20	62	19	54	19	86	17	107	23	23
Mean	772	43	1148	36	1185	36	1233	35	1295	34	1317	34	984	40	40
Std Dev	478	4	756	10	770	10	775	10	812	10	815	11	590	7	7
Total	74103	4102	110225	3480	113722	3443	118360	3335	124279	3263	126402	3285	94509	3807	3807
May 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1623	48	2335	48	2501	49	2610	49	2475	48	2618	48	1856	48	48
Min	128	25	53	20	49	20	56	19	75	20	87	19	109	23	23
Mean	776	40	1138	36	1167	34	1223	36	1271	35	1318	34	1013	40	40
Std Dev	478	7	742	10	758	10	780	10	801	11	806	10	592	7	7
Total	74459	3884	109257	3478	112041	3256	117430	3409	122043	3340	126489	3232	97267	3815	3815
June 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1638	47	2369	48	2461	48	2678	47	2527	47	2564	48	1974	48	48
Min	139	23	66	20	63	19	66	18	90	19	61	18	124	23	23
Mean	785	40	1129	37	1163	36	1229	34	1264	32	1290	34	1053	37	37
Std Dev	479	8	737	10	733	10	769	10	795	10	773	10	613	7	7
Total	75343	3867	108419	3569	111675	3437	117976	3255	121345	3079	123815	3290	101112	3523	3523

Table B.1 Statistic test results from northbound PR18 SJ1999 cont.

July 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	
Max	1671	49	2707	49	2758	49	2534	49	2669	48	2586	49	1810	47	
Min	137	25	54	19	54	20	28	18	71	19	23	19	80	25	
Mean	791	41	1150	35	1149	36	1219	35	1269	33	1311	33	1010	40	
Std Dev	470	6	743	10	743	9	779	11	810	11	803	11	593	6	
Total	75897	3932	110373	3335	110256	3465	117071	3352	121821	3174	125871	3194	96925	3872	
August 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	
Max	1643	49	2515	49	2785	49	3206	48	2645	49	2852	48	1871	49	
Min	116	29	45	18	62	18	51	17	91	19	113	19	124	23	
Mean	787	41	1174	35	1202	35	1269	34	1294	33	1348	33	1003	40	
Std Dev	477	5	760	10	782	11	840	11	821	10	830	11	591	8	
Total	75597	3963	112657	3330	115396	3376	121815	3282	124260	3150	129444	3182	96279	3813	
September 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	
Max	1594	49	2414	48	2604	50	2975	48	2680	49	2948	48	1842	48	
Min	135	25	59	19	45	20	59	18	81	19	59	18	98	21	
Mean	774	41	1160	34	1191	35	1285	32	1298	32	1351	34	1014	38	
Std Dev	472	7	756	11	772	10	839	12	819	10	853	11	592	9	
Total	74272	3954	111319	3298	114371	3403	123374	3031	124696	3085	129651	3273	97298	3628	

Table B.1 Statistic test results from northbound PR18 SJ1999 cont.

October 1999													
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	SS
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1611	48	2635	47	2571	48	2876	49	2488	49	2726	49	1859
Min	103	25	54	20	48	18	56	18	57	19	72	19	70
Mean	778	39	1162	35	1198	33	1261	34	1291	35	1334	34	1008
Std Dev	482	7	761	9	776	10	824	11	817	11	825	11	590
Total	74722	3700	111567	3383	114988	3186	121040	3267	123945	3324	128081	3253	96733
November 1999													
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	SS
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1594	48	2438	49	2451	49	2737	47	2576	49	2790	48	1830
Min	138	29	66	20	55	20	35	19	54	19	61	18	111
Mean	775	41	1164	35	1197	34	1257	34	1294	33	1341	32	1003
Std Dev	475	6	768	10	770	10	809	11	824	11	826	10	589
Total	74428	3945	111898	3358	114857	3304	120639	3265	124196	3204	128744	3051	96241
December 1999													
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	SS
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1599	48	2568	48	2742	48	3043	49	2607	49	2705	47	1861
Min	125	33	56	20	35	19	35	19	77	19	83	19	98
Mean	770	41	1159	35	1190	34	1252	35	1298	34	1336	33	1011
Std Dev	481	4	762	10	776	10	809	11	823	11	828	10	593
Total	73921	3953	111274	3376	114226	3300	120226	3325	124631	3243	128300	3163	97027

Table B.2 Statistic test results from northbound PR18 CG1999

January 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1575	47	2576	48	2372	48	2277	47	2692	48	2377	48	1964	48	48
Min	182	24	66	19	70	19	74	19	99	19	176	19	197	26	26
Mean	864	38	1285	33	1291	33	1307	33	1382	33	1364	33	1074	38	38
Std Dev	440	7	765	10	773	10	759	10	784	11	721	10	494	7	7
Total	82926	3659	123387	3216	123949	3164	125445	3142	132693	3123	130948	3211	103138	3695	3695
February 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1589	46	2538	48	2376	48	2485	48	2491	48	2366	47	1916	48	48
Min	171	30	67	19	59	19	65	18	90	19	172	19	141	23	23
Mean	858	39	1285	33	1298	33	1353	32	1373	33	1385	32	1079	38	38
Std Dev	439	5	769	10	780	10	773	11	770	10	728	10	488	7	7
Total	82370	3781	123336	3182	124593	3182	129920	3086	131852	3141	132924	3118	103588	3608	3608
March 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1544	47	2494	48	2360	48	2496	48	2595	48	2443	47	1874	47	47
Min	190	29	86	20	77	20	73	19	80	19	177	19	202	24	24
Mean	861	39	1284	34	1291	33	1348	33	1376	34	1384	33	1079	38	38
Std Dev	443	6	760	10	770	10	773	10	773	11	725	10	488	7	7
Total	82639	3751	123221	3257	123949	3178	129451	3134	132071	3229	132817	3194	103605	3663	3663

Legend: SU (Sunday), SUS (SU mean speed), M (Monday), MS (M mean speed), T (Tuesday), TS (T mean speed), W (Wednesday), WS (W mean speed), TH (Thursday), THS (TH mean speed), F (Friday), FS (F mean speed), S (Saturday), SS (S mean speed)

Table B.2 Statistic test results from northbound PR18 CG1999 cont.

April 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1568	47	2561	48	2389	48	2533	48	2616	46	2365	48	1846	48	48
Min	208	28	77	19	81	19	77	48	111	19	183	19	231	23	23
Mean	862	40	1286	32	1293	33	1349	19	1377	31	1384	32	1085	38	38
Std Dev	437	6	764	10	774	10	775	31	775	10	722	11	491	8	8
Total	82788	3805	123491	3111	124127	3161	129506	11	132191	2978	132898	3107	104132	3609	3609
May 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1549	47	2516	48	2386	48	2510	48	2592	48	2369	48	1958	48	48
Min	203	23	81	19	83	19	82	19	99	19	165	19	200	26	26
Mean	860	40	1285	33	1293	32	1350	32	1378	32	1384	33	1080	38	38
Std Dev	441	7	760	11	769	11	775	11	775	11	725	10	483	7	7
Total	82607	3836	123337	3136	124175	3063	129618	3106	132282	3071	132896	3183	103705	3648	3648
June 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1562	47	2534	48	2394	48	2594	48	2527	46	2362	47	1960	48	48
Min	212	28	64	19	74	19	67	19	102	19	164	19	163	22	22
Mean	865	40	1269	32	1288	32	1337	32	1353	32	1367	32	1069	37	37
Std Dev	443	6	756	11	774	11	765	11	763	10	721	10	479	8	8
Total	83045	3831	121791	3101	123614	3078	128393	3113	129913	3044	131248	3091	102664	3554	3554

Table B.2 Statistic test results from northbound PR18 CG1999 cont.

July 1999													
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	SS
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1594	47	2513	48	2357	48	2457	48	2491	47	2461	47	1835
Min	178	23	88	19	78	20	80	19	98	19	161	19	204
Mean	840	39	1266	32	1270	32	1327	32	1356	31	1372	32	1074
Std Dev	438	7	749	11	761	11	763	11	757	10	718	11	488
Total	80865	3784	121489	3049	121966	3063	127407	3055	130168	3010	131693	3039	103135
August 1999													
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	SS
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1578	47	2693	48	2390	48	2635	48	2639	48	2386	47	1904
Min	200	25	73	19	72	19	67	19	89	19	189	19	226
Mean	861	40	1287	32	1293	32	1352	31	1376	31	1388	31	1079
Std Dev	442	6	768	10	772	10	777	10	775	10	726	10	481
Total	82654	3798	123587	3030	124086	3035	129817	2961	132076	2973	133283	3007	103619
September 1999													
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	SS
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1566	48	2526	48	2391	48	2531	47	2611	48	2482	47	1890
Min	202	23	90	19	38	18	78	18	77	18	185	18	204
Mean	866	39	1288	32	1294	32	1353	32	1379	31	1384	33	1081
Std Dev	444	7	762	11	777	10	778	11	774	11	726	11	488
Total	83102	3789	123624	3025	124204	3044	129852	3079	132369	2977	132899	3174	103787

Table B.2 Statistic test results from northbound PR18 CG1999 cont.

October 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1555	47	2547	47	2398	48	2560	48	2519	47	2371	48	1895	48	48
Min	153	21	81	19	61	18	74	18	111	18	176	19	156	21	21
Mean	862	37	1286	31	1292	30	1349	31	1376	31	1381	32	1087	34	34
Std Dev	441	9	763	10	771	10	773	11	774	11	726	10	476	9	9
Total	82730	3532	123416	3020	124028	2870	129471	2998	132125	2995	132540	3086	104311	3282	3282
November 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1673	47	2625	48	2372	48	2551	48	2586	48	2367	45	1880	47	47
Min	188	23	77	19	79	19	84	18	99	18	164	19	185	22	22
Mean	855	38	1284	31	1290	31	1348	31	1378	32	1385	31	1079	36	36
Std Dev	445	8	763	10	781	11	775	11	773	11	724	10	487	8	8
Total	82069	3678	123275	2989	123884	2985	129412	2969	132246	3034	132051	2947	103593	3412	3412
December 1999															
	SU	SUS	M	MS	T	TS	W	WS	TH	THS	F	FS	S	SS	
# Records	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Type Var	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM	NUM
Length	4	2	4	2	4	2	4	2	4	2	4	2	4	2	2
Max	1562	47	2517	47	2390	48	2493	48	2520	48	2368	46	1924	46	46
Min	178	23	73	18	70	19	80	18	93	18	167	18	162	21	21
Mean	860	39	1276	32	1283	32	1342	32	1362	32	1374	31	1082	34	34
Std Dev	442	6	759	10	772	11	770	11	767	11	728	10	491	9	9
Total	82541	3766	122535	3052	123148	3101	128786	3038	130739	3062	131892	3015	103867	3306	3306

APPENDIX C: MINING PREPARATION

This appendix provides a step-by-step explanation of the process followed to use IBM Intelligent Miner for Data. The example used for the mining preparation presented herein is intended to be used as a mere example for users interested in using the association mining tool of the IBM Intelligent Miner for Data. Additional information can be obtained from the User's Manual provided with the software package.

C1. Data Preparation

This Chapter describes an example of the application of association mining by means of the IBM DB2 Intelligent Miner for Data. Figure C1.1 shows the main IBM DB2 Intelligent Miner data window. The first step was to create the working data, whether it is from a flat file or a database table. Flat files are notepad text (.txt) files and database table files are those structured in applications such as MS Access. In this example flat files were used.

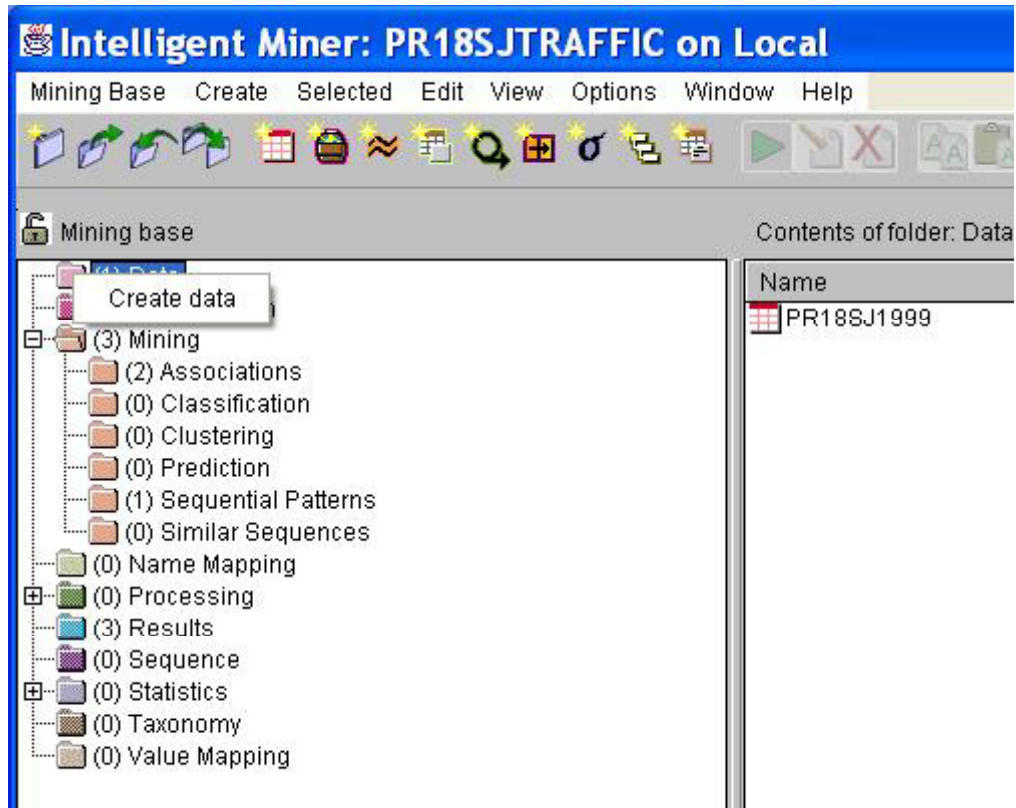


Figure C1.1 Main create data window

Figure C1.2 shows part of the data used in this example. The dataset on the left side of the figure is a spreadsheet table file structured in MS Excel and the one on the right is already converted into a flat file structured in Notepad. Cleaning the data required two parts. The first was completed using MS Excel since it allows for a faster and easier review of the data, and the second was completed using MS Word. Some of the steps that should take place while cleaning the data using MS Excel are: filling empty spaces with a known value to the user (for example, 1 or 0 depending on the objective of the application) and changing specific values such as TIME from, for example, 12:30 to 1230. The second part of cleaning the data consists on replacing every “Tab” with spaces;

this was easy to accomplish using MS Word. A list of specific characters can be found using the Help icon on the latter software package. The character used to specify “Tabs” in MS Word is “^t,” therefore, to change every “Tab” to “3 spaces” the user must use the find and replace tool from MS Words as shown on Figure C1.3.

Database Table View (MSExcel (.xls))						
	A	B	C	D	E	F
1	Month	Time	OBS	DIR	SU	SUS
2	MAY	2400	1	SJ	368	63
3	MAY	2415	2	SJ	376	62
4	MAY	2430	3	SJ	401	64
5	MAY	2445	4	SJ	351	64
6	MAY	100	5	SJ	367	65
7	MAY	115	6	SJ	341	64
8	MAY	130	7	SJ	328	63
9	MAY	145	8	SJ	338	63
10	MAY	200	9	SJ	300	62
11	MAY	215	10	SJ	284	64
12	MAY	230	11	SJ	270	64
13	MAY	245	12	SJ	251	65
14	MAY	300	13	SJ	196	65
15	MAY	315	14	SJ	174	66
16	MAY	330	15	SJ	163	65
17	MAY	345	16	SJ	187	65

Flat File View (Notepad (.txt))						
File	Edit	Format	View	Help		
MAY	2400	1	SJ	368	63	DRY
MAY	2415	2	SJ	376	62	DRY
MAY	2430	3	SJ	401	64	DRY
MAY	2445	4	SJ	351	64	DRY
MAY	100	5	SJ	367	65	DRY
MAY	115	6	SJ	341	64	DRY
MAY	130	7	SJ	328	63	DRY
MAY	145	8	SJ	338	63	DRY
MAY	200	9	SJ	300	62	DRY
MAY	215	10	SJ	284	64	DRY
MAY	230	11	SJ	270	64	DRY
MAY	245	12	SJ	251	65	DRY
MAY	300	13	SJ	196	65	DRY
MAY	315	14	SJ	174	66	DRY
MAY	330	15	SJ	163	65	DRY
MAY	345	16	SJ	187	65	DRY
MAY	400	17	SJ	141	66	DRY
MAY	415	18	SJ	128	66	DRY
MAY	430	19	SJ	139	65	DRY
MAY	445	20	SJ	167	64	DRY
MAY	500	21	SJ	182	64	DRY
MAY	515	22	SJ	178	65	DRY
MAY	530	23	SJ	174	65	DRY

Figure C1.2 Database table and flat file view



Figure C1.3 Find and replace tool in ms words

The dataset used consisted of traffic flow, speed, weather, work zone, and accident data for every 15-minutes of every day of the week for 1 week of each month of the year. There is a unique value for each row or record in the working data that was set to be OBS for observations, so there were 1152 records (calculated as $96 \times 12 = 1152$) in each dataset of each year.

Figure C1.4 shows the data format and settings for the creation of the data to be used in the mining tool. In this window the user can choose the settings name for the data. In the settings window, all the data files created are shown. Also, the user must remember to check the box at the bottom of the window in order to complete the creation of data successfully.

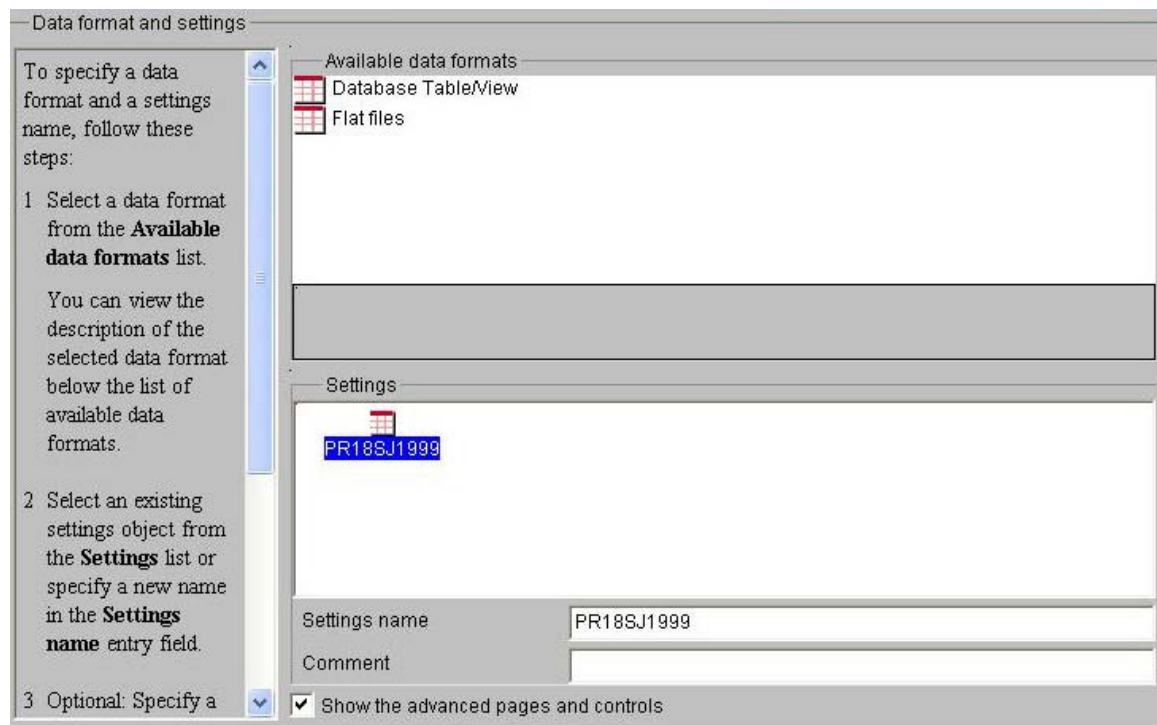


Figure C1.4 Data format and settings

The working data was then searched and added to the selected files in the flat files window (see Figure C1.5).

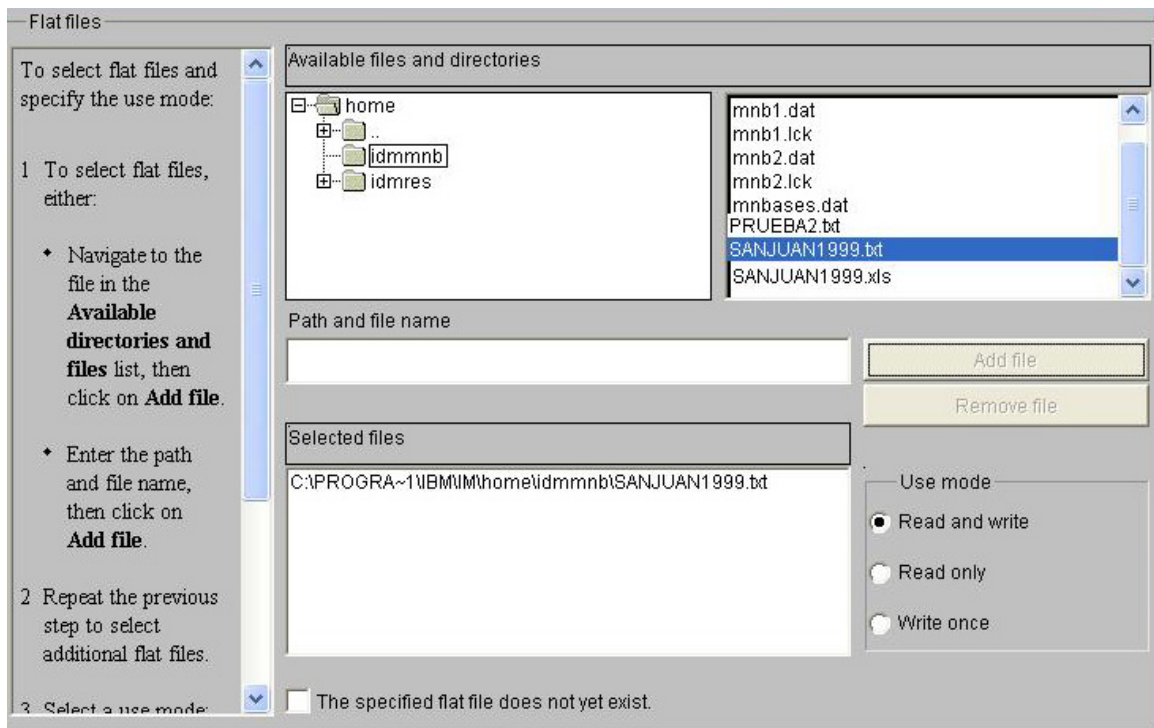


Figure C1.5 Flat files selection

In the field parameter window, as shown in Figure C1.6, the position, field name, and data type were selected. The latter information can be accessed and updated as needed prior to any data mining application. The data type display has a scroll with a few categories to chose from. The position order at the top of the flat file display window has to be shown for the data to be read and saved for further mining applications. If the position order is not shown, it is very likely that there is an error in the data.

Field parameters

To specify field parameters:

- 1 Enter the begin and end position for each field separated by the [range](#) [delimiter](#).
- 2 Enter the field name for each field.
- 3 Select the [data type](#) for each field.
- 4 Optionally, specify a [name mapping](#) for each field.

Flat file display

...	1	2	3	4	5	6	7	...
MAY	2400	1	SJ	368	63	DRY	166	65
MAY	2415	2	SJ	376	62	DRY	156	65

Selection

Begin and end position	Field name	Data type	Name Mapping
		Categorical	

Add Update Delete

Field parameters


Begin and end position	Field name	Data type	Name Mapping
1-3	MONTH	Categorical	
7-10	TIME	Categorical	

Figure C1.6 Field parameters for flat files

Figure C1.7 shows a summary of the previous steps.

Settings Flat files Field parameters Computed fields **Summary**

Summary



Summary

Parameter	Value
Data format:	Flat file
Settings	PR18SJ1999
Selected files:	C:\PROGRA~1\IBM\IM\home\idmmnb\SANJUAN1999.bc
Use mode:	Read and write
The specified flat file does not yet exist:	FALSE

Figure C1.7 Summary for flat file used

C2. Association Mining Preparation

Figure C2.1 shows the main IBM DB2 Intelligent Miner for Data applications window. In this example the association mining application was chosen. The work can, and must, be saved in a mining base for future access and modifications of the mining method.

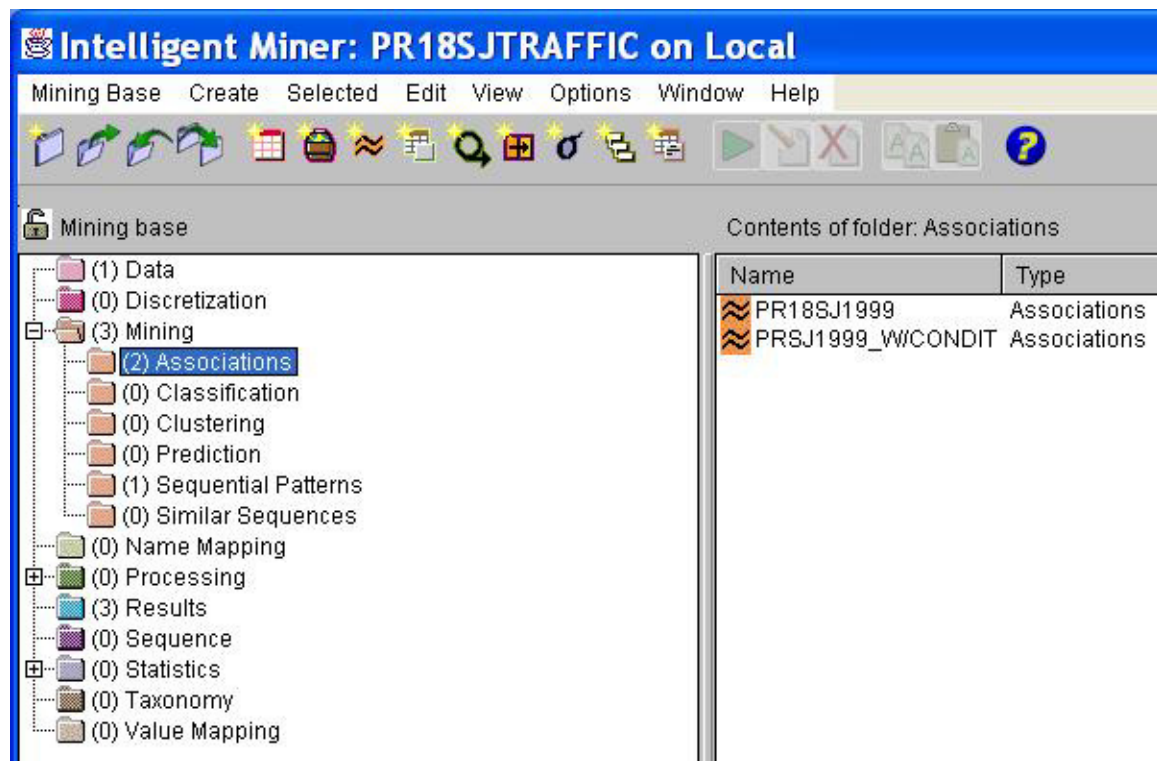


Figure C2.1 Main mining tools window

The first step of any data mining application is to select a settings name. Figure C2.2 shows the settings name chosen for this association mining. Observe that all the association mining applications created are displayed in the settings window. The box at the bottom of the window must be checked for the entire setting of the association mining to be performed.

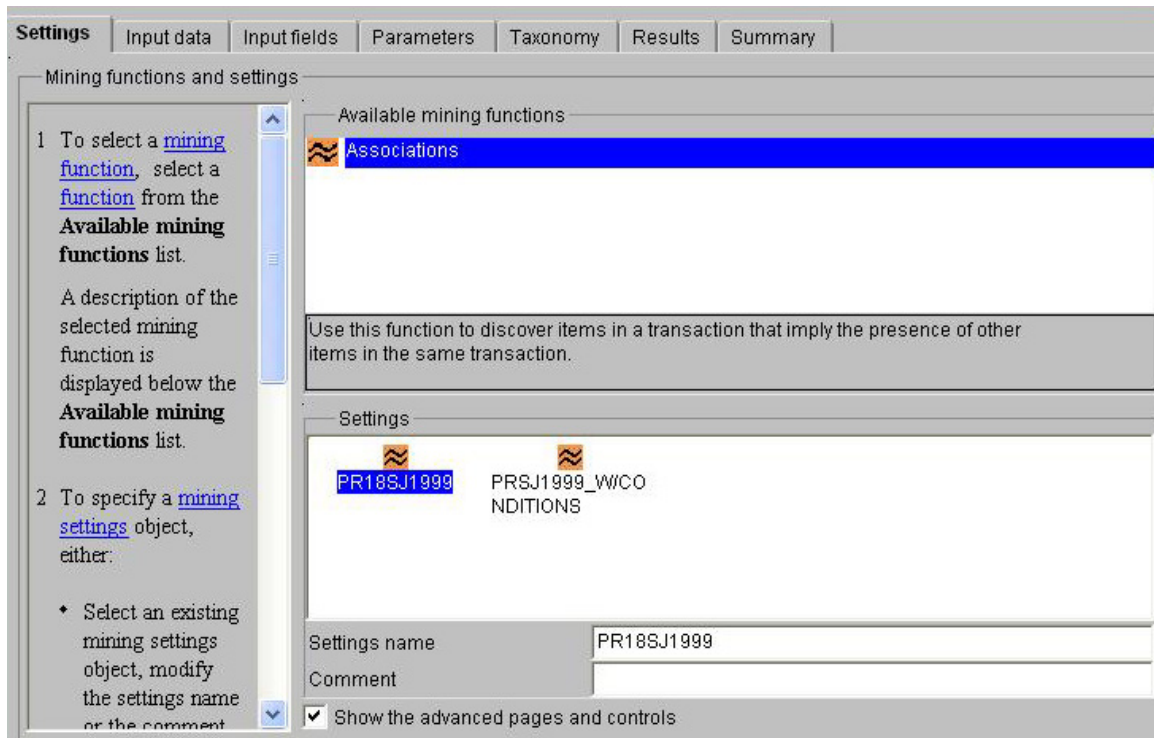


Figure C2.2 Associations settings

The second step was to select the input data, which was the same working data created earlier. All the available input data will be displayed as seen in Figure C2.3. Any association mining application can be filtered by writing a query in the filter records condition display. The latter is done in the advanced parameters option provided in the lower right corner of the window.

Figure C2.3 Associations input data

There are only two inputs needed for this type of mining application: transaction field and item field. Transaction field is the identifier that groups transactions together. In this case the transaction field chosen was M for Monday traffic flow. The transaction field must be a CATEGORICAL type field (for this particular example). Item field is a collection for things that are included under the identifier number mentioned. In this case it was MS for Monday speed (see Figure C2.4). Item fields must also be CATEGORICAL type.

By using Monday traffic flow for the transaction field and Monday speed for item field the user can analyze what speeds are related or associated to particular traffic flows. That is, which combinations of speed and flow repeat a certain amount of times on any given Monday. This application can be studied

per each weekday to find the associations of speeds used over the entire traffic flow of each week.

The checked box at the bottom was checked to ensure the transaction field or traffic flows are sorted before running the analysis. The “sort” feature is only necessary when using flat files not DB2 tables.

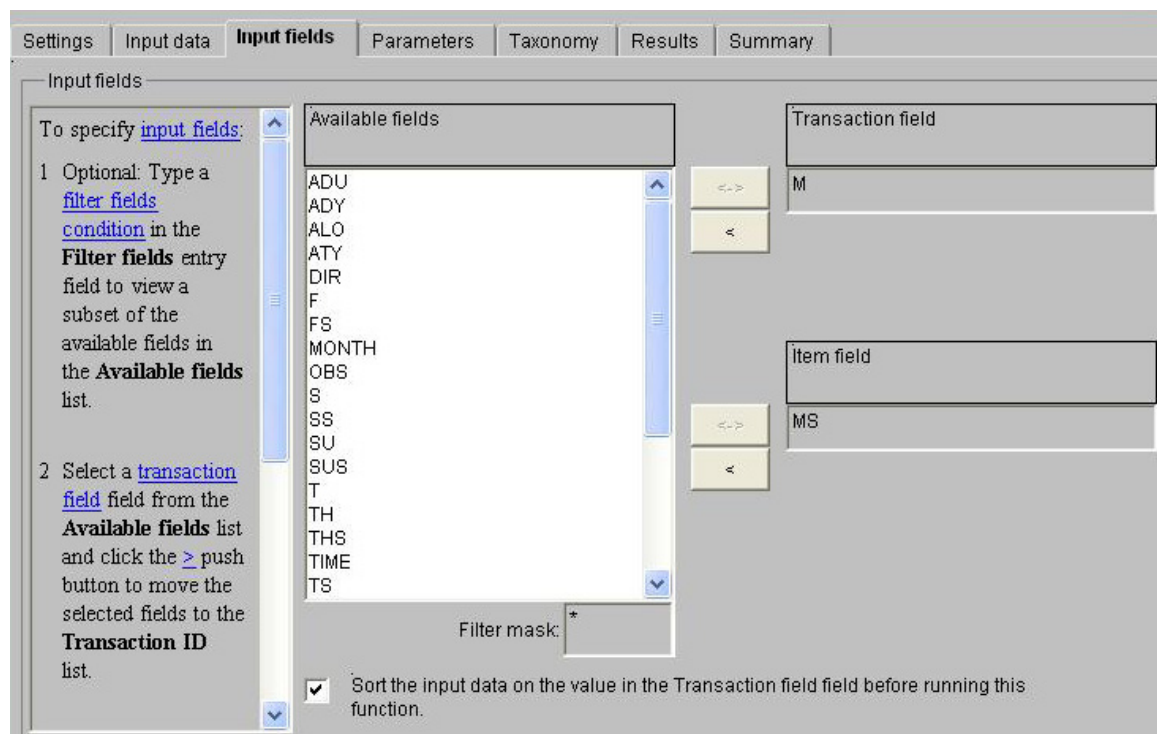


Figure C2.4 Associations input fields

On the parameters tab shown on Figure C2.5, there are four settings that can be adjusted to increase or decrease the results. The minimum support option indicates the relative occurrences of the rule within the data. It is determined by dividing the number of transactions supporting the rule by the total number of transactions. For example, in Figure C2.8 it can be seen that the combination of Monday speed 66 mph and 65 mph, [66] + [65], make up 2.45% of all the

Monday flow within this model. However, depending on the type and amount of data used the user can change the minimum support to a lower value in order for the mining tool to pick up the associations in the model. Furthermore, even when the data allows for a higher minimum support value to be used the results will be very different because the larger the value the more general the results and the lower the minimum support the more specific the results. In this example, a low minimum support value was used.

The minimum confidence indicates the relative strength or reliability of the detected rules within the input data. The confidence level is determined by dividing the number of transactions supporting the rule by the number of transactions supporting the rule body only. Referring to Figure C2.8, it can be seen that for the combination [66] + [65] the confidence level is 33.33%.

$$\text{Confidence } (65 \rightarrow 66) = \text{Support } (65 \cap 66) / \text{Support } (66)$$

$$\text{Confidence } (65 \rightarrow 66) = 2.45\% / 7.35\% = 33.33\%$$

The maximum rule length determines the number of items that are present in the association rule. In this case, a maximum rule length of 2 was chosen. Thus, in Figure C2.8 all the association rules developed were on the order of two; one item set in the rule body [66] and one item in the rule head [65], or [66] \Rightarrow [65]. Lastly, item constraints determine which rules or patterns are included or excluded from the results. For this example, the item constraints were left at the default.

The screenshot displays the 'Parameters' tab of an association mining application. The interface is divided into a sidebar on the left and a main parameter configuration area on the right. The sidebar contains a note about controlling the mining run and a list of four parameters with links: 1. [Minimum support](#), 2. [Minimum confidence](#), 3. [Maximum rule length](#), and 4. [Item constraints](#). The main area contains four parameter settings, each with a text input field, a spinner control, and a 'Use default' checkbox. The 'Minimum support' is set to 0.5, 'Minimum confidence' to 20, 'Maximum rule length' to 2, and 'Item constraints' to 'None'. The 'Use default' checkbox for 'Item constraints' is checked.

Parameter	Value	Use default
Minimum support	0.5	<input type="checkbox"/>
Minimum confidence	20	<input type="checkbox"/>
Maximum rule length	2	<input type="checkbox"/>
Item constraints	None	<input checked="" type="checkbox"/>

Figure C2.5 Associations parameters setting

Figure C2.6 shows the results developed in the association mining application. Note, that by checking the box below the comments display the user agrees to overwrite the results with the same name. This is useful when starting to use the application, since many runs are performed prior to understanding the method.

The screenshot shows a software window with several tabs: Settings, Input data, Input fields, Parameters, Taxonomy, **Results**, and Summary. The 'Results' tab is active.

Results

To specify the name for a result, either:

- Use the default name displayed in the **Results name** entry field or specify a different name for the new result. You can also associate a comment with the new result.
- Select an existing result from the **Available results** list. You can also modify the result name or the associated comment.

Available results

PR18SJ1999	PRSJ1999_W/CO NDITIONS
------------	---------------------------

Results name : PR18SJ1999

Comment

☒ If a result with this name exists, overwrite it.

Figure C2.6 Associations results

Figure C2.7 shows an overview of the inputs, functions, outputs, and parameters chosen for this example. In this case the inputs are PR18SJ1999 that were processed by the association rule PR18SJ1999 and the results were labeled PR18SJ1999 as well. Notice that the same name can be used in each part of the application without causing any problem to the applications, although different names are suggested within each individual part of the process to avoid losing information. In summary, the association run was optimized for disk space. There was a condition selected to filter the records, the default power option “-rulekind” was selected, and the transaction field used was M for Monday traffic flows. Scrolling down on the right side of the window, the user can see the additional parameters and its values; these are not visible on Figure C2.7.

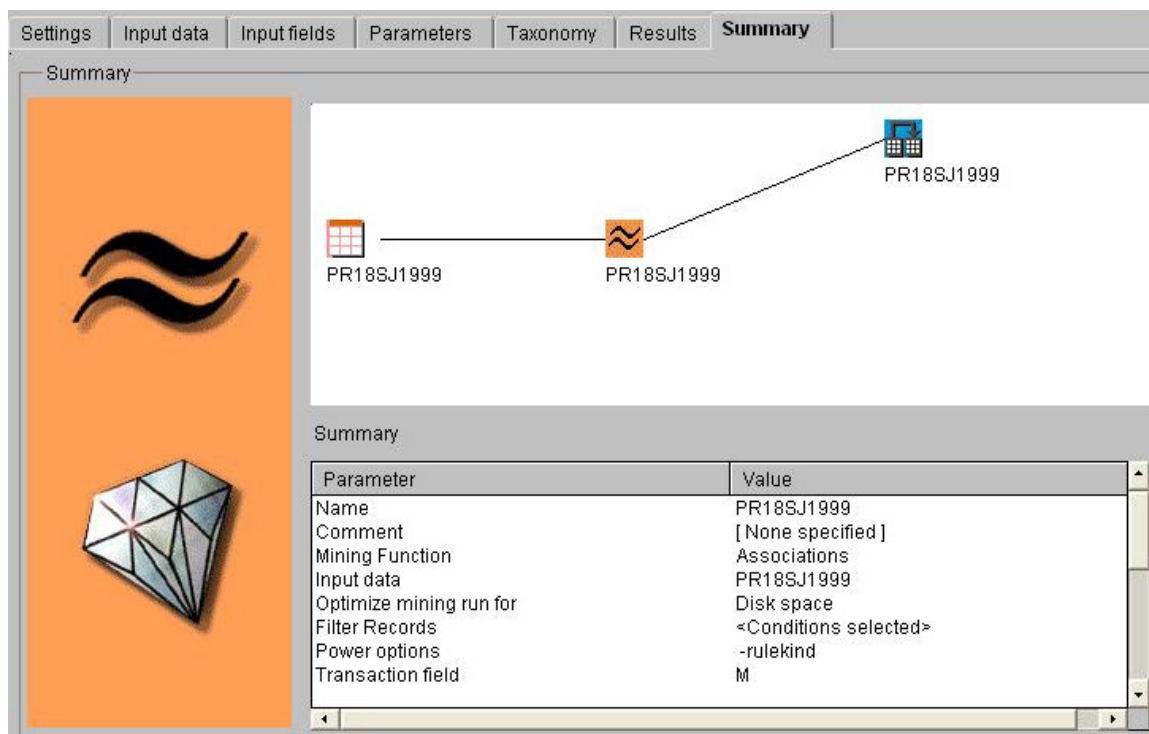


Figure C2.7 Associations summary

When the process is completed, the visualizer produces rules, itemsets, a graph, and statistics (see Figures C2.8 through C2.13). The generated rules can be seen on the first column of Figure C2.8. These are read from left to right. For example, the first rule (which is the one with the strongest confidence) says that if a speed of 66 mph is found then there is a high likelihood that a speed of 65 mph is also found. The support column gives an indication of how frequent the combination of 66 and 65 mph were found together. In this case, 2.45% of the observations include both speeds. To calculate the confidence for the 66/65 combination the user can take the support for 66 + 65, 2.45% and divide it by the support for just 66 mph alone 7.35%, which gives 33.33% (see first row of Confidence column on Figure C2.8).

Visible rules:				
Rule	Support	Confidence	Lift	AI
[66] ==> [65]	2.4510%	33.3300%	2.3446	
[61] ==> [63]	0.9804%	23.5300%	1.3715	
[67] ==> [66]	0.9804%	28.5700%	3.8855	
[67] ==> [65]	0.9804%	28.5700%	2.0097	
[61] ==> [64]	0.9804%	23.5300%	1.5484	
[67] ==> [64]	0.9804%	28.5700%	1.8801	
[67] ==> [63]	0.7353%	21.4300%	1.2491	

Figure C2.8 Example of association numerical rules

In the last column of Figure C2.8 the user can see the Lift. Lift is the factor by which the confidence exceeds the expected confidence.

$$\text{Lift}(X) = \text{Confidence}(X \rightarrow Y) / \text{Expected Confidence}(X \rightarrow Y)$$

$$\text{Lift}(X) = \text{Confidence}(X \rightarrow Y) / \text{Support}(Y)$$

The confidence of 66 + 65 is 33.33% which when divided by the support for 65, 14.21% gives 2.34 as seen on Figure C2.8. A high lift (>1) value indicates that the rule may be interesting in some sense as the right hand side of the rule occurs more often than expected.

$$\text{Lift}(X) = 33.33\% / 14.21\% = 2.34$$

Figure C2.9 illustrates the itemsets and support percentage. Support is the percentage of all the transactions (or traffic flows) that the itemsets (speeds) appear in. For example, the user can see that 17.15% of the speeds include 63 mph, 15.19% % of the speeds include 64 mph, and 14.46% of the speeds include 62 mph. The visualizer includes a color-coding that help separate support levels. Intense yellow indicate the most frequent speeds and dark blue indicate the least frequent speeds.

Visible item sets:		
Item Set	▼ Support	In Ru
[63]	17.1569%	
[64]	15.1961%	
[62]	14.4608%	
[65]	14.2157%	
[60]	13.7255%	
[55]	9.3137%	
[56]	8.8235%	
[59]	7.3529%	
[66]	7.3529%	
[58]	6.3726%	
[54]	4.4118%	
[57]	4.4118%	
[61]	4.1667%	
[67]	3.4314%	
[50]	3.1863%	
[53]	2.6961%	
[63]+[65]	2.6961%	
[65]+[66]	2.4510%	
[63]+[64]	2.4510%	
[62]+[64]	2.4510%	
[60]+[64]	1.7157%	
[62]+[65]	1.4706%	
[65]+[64]	1.4706%	
[63]+[62]	1.4706%	
[49]	1.2255%	

Figure C2.9 Example of items sets

Figure C2.10 shows the graph developed by the IBM DB2 Intelligent Miner for Data visualizer. Dots, arrows, and lines of various widths make up the graph. The dots represent the speeds (MS), the arrows represent the rules (e.g. [66] ⇒ [65]), the line colors indicate the support (remembering that yellow indicates higher support), and the line width represents the lift (where wider lines indicate higher lift values).

Figure C2.10 illustrates the speed combination [66] + [65] (strong yellow color arrow) which occur more frequently (higher support) than speed combination [67]

+ [65] (light blue color arrow). Another observation is that speed combination [67] + [66] had more lift (wider blue line) than [66] + [65] (thinner yellow line). The wider line means that occurrences of that item set are much more probable than random chance.

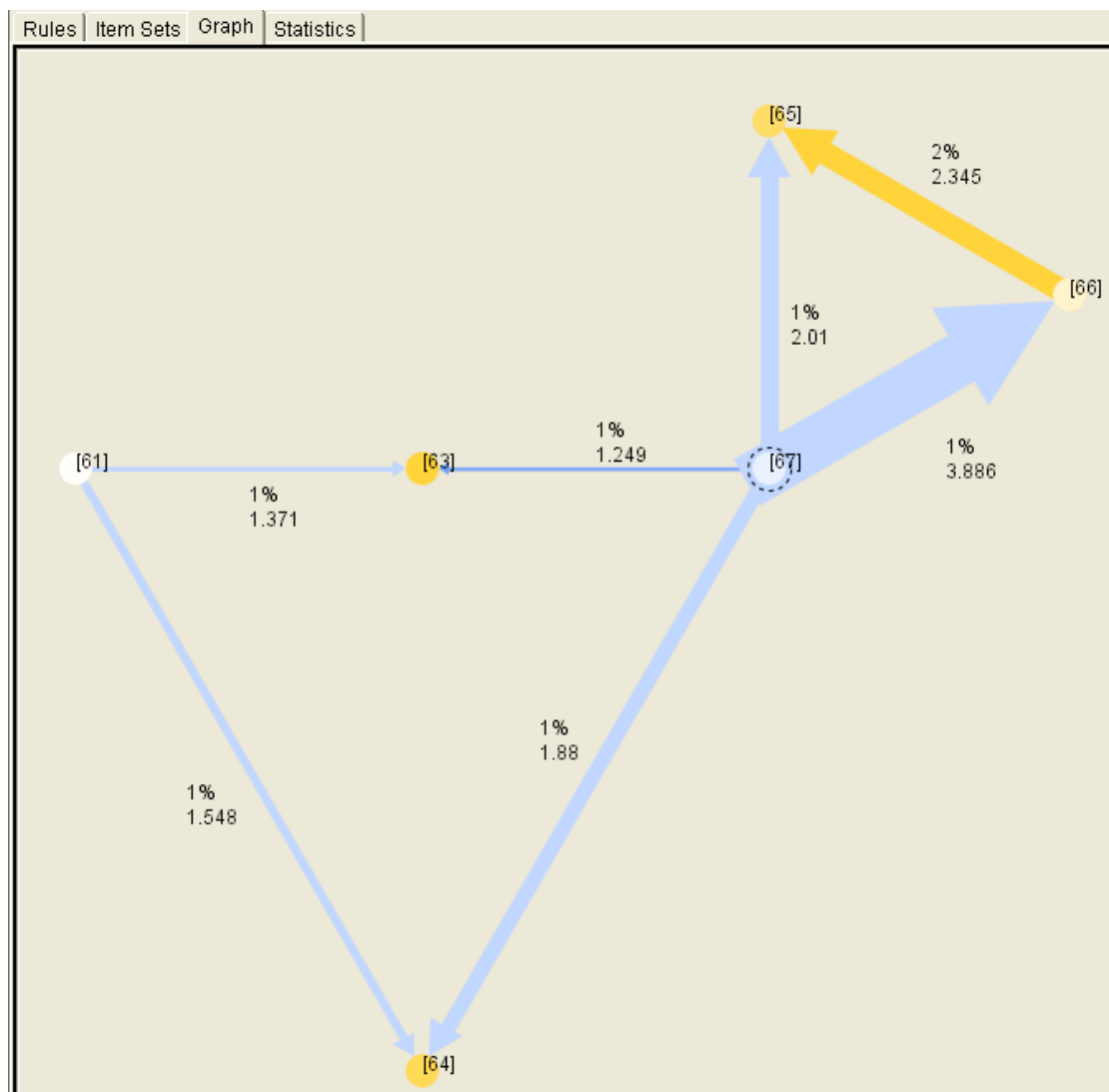


Figure C2.10 Example of graphical association rules

The color scale shown in Figure C2.11 was part of the information provided on the same tab as the graph on Figure C2.10. Again, the line color going from blue to yellow means the rule moves from less support to more support.

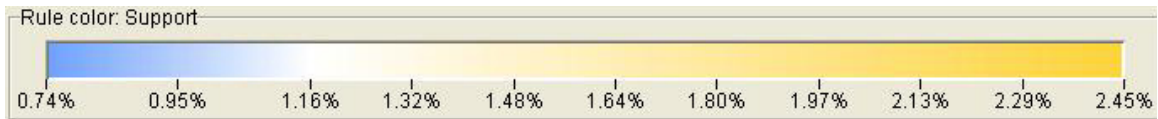


Figure C2.11 Rule color scale

Likewise, the lift is illustrated by the width of the line. As shown on Figure C2.12, the thicker the line the stronger the confidence level between the data being mined.

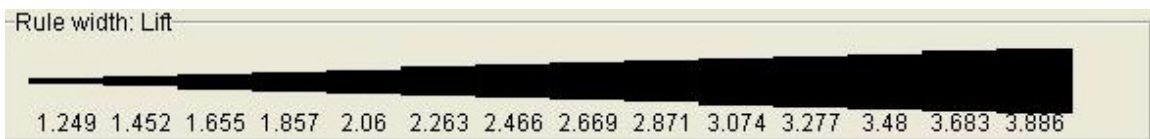


Figure C2.12 Lift legend

The global statistics, as mentioned earlier, were also part of the information provided by the IBM DB2 Intelligent Miner for Data visualizer (see Figure C2.13). These provide a summary of the data that was used. In this example, 408 traffic flows were analyzed and that on average 1.39 speeds were found in each situation. 45 itemsets or Monday speeds were found in the model of which 19 were single speeds, and 7 association rules were developed by the mining tool.

▼ Global Statistics	
Number of transactions:	408
Average number of items per transactions:	1.39
Maximum number of items per transactions:	5
Number of item sets:	45
Number of single item sets:	19
Number of item sets used in rules:	6
Minimum rule support:	0.74%
Minimum rule confidence:	21.43%
Maximum rule length:	--
▼ Statistics for Visible Objects	
Visible rules:	7
Visible item sets:	45

Figure C2.13 Associations statistics

REFERENCES

- (2000). *ITS handbook 2000: recommendations from the world road association (PIARC)*, Artech House, Boston, MA.
- Acevedo, G. (2003). Department of Transportation and Public Works of Puerto Rico (DTOPPR), written communication. May 5.
- Agrawal, R., Imielinski, T., and Swami, A. (1993). "Mining association rules between sets of items in large databases." *ACM SIGMOD Conference*, Washington, DC, May 1993.
- Agrawal, R. and Shafer, J.C. (1996). IBM Research Report. Parallel Mining of Association Rules: Design, Implementation and Experience. RJ 10004 (02/01/96). <http://www-1.ibm.com/support/docview.wss?uid=swg27008249&aid=1> (April 5, 2008).
- Albert, L. P. (1999). "An evaluation of the potential of public/private partnerships for the management of archived ITS data." Texas Transportation Institute (TTI) and Southwest Region University Transportation Center.
- Amado, V. (2001). "Expanding the use of pavement management data," MS Thesis, University of Missouri, Columbia.
- Buchheit, R., Garret, J.H., McNeil, S., and Chalkline, M.H. (2002). "Automated procedures for improving the accuracy of sensor-based monitoring data." *Proceedings of the 7th International Conference on Application of Advanced Technologies in Transportation*. Cambridge, Massachusetts, American Society of Civil Engineers.
- Bureau of Transportation Statistics (BTS). (2000). *The changing face of transportation*, Bureau of Transportation Statistics and US Department of Transportation, Washington, D.C.
- Burgos, J. (2005). Puerto Rico Highway and Transportation Authority. April 2005, electronic communication.
- California Highway Patrol. (2002). *California Highway Patrol (CHP)*, <http://www.chp.ca.gov/> (July 17, 2002).

- Carey, B., Dougherty, R., Hilchie, A., and Morgan, J. (2003). "Sample size and modeling accuracy with decision tree based data mining tools." *Academy of Information and Management Science Journal*, 2003.
- Carvalho Viglioni, G.M. (2007). "Methodology for Railway Demand Forecasting Using Data Mining." *SAS Global Forum*, Rio de Janeiro, Brazil.
- Cordova, P. (2003). Cordova & McCadney Traffic and Transportation, personal and electronic communication, April 2003.
- Dahlgren, J., García, R.C., and Turner, S. (2001). "Completing the circle: using archived operations data to better link decisions to performance." *California PATH Program, Institute of Transportation Studies, and University of California, Berkeley*, California PATH Research Report UCB-ITS-PRR-2001-23.
- Dahlgren, J., Turner, S., and Garcia, R.C. (2002). "Collecting, processing, archiving and disseminating traffic data to measure and improve traffic performance." *Transportation Research Board (TRB) 81st Annual Meeting*. Washington, D.C. Jan. 2002.
- Evans, R. (2003). "Mining your warranty data using ibm db2 intelligent miner for data association methods." *IBM Corp.* <www.106.ibm.com/developerworks/db2/library/technicalarticle/dm-0311evans/> (April 13, 2005), IBM Corp. 2005: 40.
- Fayyad, U., Shapiro, P., and Smyth, P. (1996). "From data mining to knowledge discovery: An overview." *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA: The AAAI Press / The MIT Press, 1996. 1-34.
- Felix, M. (2003). Puerto Rico Highway and Transportation Authority (PRHTA), personal communication, January 8, 2003.
- García, I.M. (2003). Department of Transportation and Public Works of Puerto Rico (DTOPPR), written communication. September, 2003.
- Gordon, S., and Trombly, J. (2000). "Tracking the deployment of the integrated metropolitan intelligent transportation systems infrastructure in the usa: fy99 results." *FHWA-OP-00-016*, US Department of Transportation and FHWA ITS Joint Program Office, Washington, D.C.
- Greenshields, B. D., and Weida, F.M. (1978). "Statistics with applications to highway traffic analyses." *Eno Foundation for Transportation, Inc.*, Westport and Connecticut.

- Guía Urbana del Área Metropolitana. (2000). "2000 san juan metropolitan area metro data." CD-ROM.
- HCM. (2000). *Highway capacity manual*, Transportation Research Board, Washington, D.C.
- IBM Corporation (1996). *Using the intelligent miner for data*, version 6 release 1, SH12-6394-00
- ITS Data Archiving Five-Year Program Description. (2000). *ADUS Program*.
- Khisty, C. J. (1990). *Transportation engineering: an introduction*, Prentice Hall, New Jersey, NY.
- Kotsiantis, S. and Kanellopoulos, D. (2006). "Association rules mining: a recent overview." *GESTS International Transactions on Computer Science and Engineering*, 32(1), 71-82.
- Lee, D.H., Jeng, S.T., and Chandrasekar, P. (2004). "Applying data mining techniques for traffic incident analysis." *Journal of the Institution of Engineers, Singapore*, 44 (2).
- Manila, H. (2000). "Theoretical framework for data mining." *SIGKDD Explorations*, 1(2), *ACM SIGKDD*, pp. 30-32.
- Mannering, F. L., and Kilareski, W.P. (1998). *Principles of highway engineering and traffic analysis*, John Wiley & Sons, Inc., New York.
- Margiotta, R. (1998). "ITS as a data resource: preliminary requirements for a user service." *Report FHWA-PL-98-031*, FHWA, Washington, D.C.
- McCown, S. (2005). National Climatic Data Center (NCDC). February 2005, electronic communication.
- Merriam-Webster's Collegiate® Dictionary. (2000). <<http://www.m-w.com>> (June 7, 2000).
- Microsoft (2001). "Microsoft Encarta Encyclopedia Standard 2001." Microsoft Works Suite 2001.
- Morales, J. M. (2003). J.M. Morales & Associates Transportation Engineering Consultants, electronic communication, March 2003.
- Moss, L.T. (2007). "Defining data mining." *BusinessIntelligence.com*, <<http://businessintelligence.com>> (September 15, 2007).

- Musick, R., Catlett, J., and Russel, S. (1993). "Decision theoretic subsampling for induction on large databases." *Proceedings of the tenth international conference on machine learning*, Morgan Kaufmann, San Mateo, Ca. 212-219.
- Nassar, K. (2007). "Application of data-mining to state transportation agencies' projects databases." *ITcon*, 12, 139-149.
- NCDC (2005). "Local climatological data, san juan luis munoz marin international airport (SJU) station." *National Climatic Data Center*. <<http://www.ncdc.noaa.gov/oa/ncdc.htm>> (February 1, 2005).
- Oates, T. and Jensen, D. (1998). "Large data sets lead to overly complex models: an explanation and solution." *Proceedings of the fourth international conference on knowledge discovery and data mining*, AAAI Press, Menlo Park, Ca. 294-298.
- Office of Highway Policy Information (OHPI). (2000). "Transportation Acronyms." US Department of Transportation Federal Highway Administration. <<http://www.fhwa.dot.gov/ohim/acronym.htm>> (January 25, 2000).
- Pearce, V. (1999). "Transportation management centers-bringing it all together through staff coordination." *ITE Journal*, (40-42), 243-244.
- Pérez, I. A. (2004). Puerto Rico Highway and Transportation Authority (PRHTA), personal communication, February 2004.
- Pesquera, C. I., and Gonzalez, S.L. (1996). "Tren urbano on track to relieve congestion." *Railway Gazette International*, 152(1).
- Pilot Software. (2000). "*Data mining white paper, glossary of data mining terms.*" <<http://stuart.iit.edu/course/ecom540/summer00/week3/Glossary%20of%20Data%20Mining%20Terms.htm>> (June 5, 2000).
- Poch, M., and Mannering, F. (1996). "Negative binomial analysis of intersection-accident frequencies." *Journal of Transportation Engineering*, 122(2), 105-113.
- Pyle, D. (1999). *Data preparation for data mining*, Morgan Kaufmann Publishers, San Francisco, CA.
- Quiñones, L. (2003). Department of Transportation and Public Works of Puerto Rico (DTOPPR), personal communication. January 8, 2003.
- Smith, L. (2003). "Archived Data." *ITS Decision Report*. <www.calccit.org/itsdecision> (September 12, 2007)

- Soibelman, L. and Kim, H. (2000). Generating construction knowledge with knowledge discovery in databases. Proceedings of the 8th International Conference on: Computing in Civil and Building Engineering, Stanford, California.
- Texas Transportation Institute (TTI) and The Texas A&M University System. (1999). "ITS data archiving: Case study analyses of san antonio TransGuide[®] data." *US Department of Transportation and Federal Highway Administration. Report No. FHWA-PL-99-024.*
- Texas Transportation Institute (TTI). (2001). "Guidelines for developing ITS data archiving systems." *Report 2127-3 Texas Department of Transportation (TxDOT), US DOT, and FHWA.*
- Toivonen, H. (1996). "Sampling large databases for association rules." *Proceedings of the 22th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., San Francisco, Ca. 134-145
- TransGuide[®]. (2002). <<http://www.transguide.dot.state.us/docs/section2.html>> (June 25, 2002).
- Turner, S. (2001). "Archived its data serve multiple purposes." *Texas Transportation Researcher*, 37(2).
- Turner, S., Albert, L., Gajewski, B., and Eisele, W. (2000). "Archived ITS data quality: Preliminary analysis of san antonio TransGuide[®] data." *Texas Transportation Institute (TTI) The Texas A&M University System*. Jan. 2000.
- Turner, S. (2002). "Using historical traffic data to meet future urban transportation needs." *Texas Transportation Researcher*. 38(2).
- Two Crows. (1999). "*Introduction to data mining and knowledge discovery*." Two Crows Corporation. 3rd ed.
- US Census (2008). <<http://www.census.gov>> (April 6, 2008).
- US Department of Transportation (USDOT) and Federal Highway Administration (FHWA). (1999). "Tracking the deployment of the integrated metropolitan intelligent transportation systems infrastructure in the USA: FY 1997 results." *Joint Program Office for Intelligent Transportation Systems, USDOT and FHWA*. Washington, DC.
- US Department of Transportation (USDOT). (2000). "The changing face of transportation." *Bureau of Transportation Statistics, US DOT*. Washington, DC.

US Department of Transportation (USDOT). (2002). "ITS deployment tracking, 2000 survey results." <<http://www.its.dot.gov/>> (November 7, 2002).

Villalba, J. (2004). Edwards and Kelcey, personal communication, January 2004.

Winick, R. M. (2002). Motion Maps, telephone communication, July 2002.

Zaki, M.J., Parthasarathy, S., Li, W., and Ogihara, M. (1996). "Evaluation of sampling for data mining of association rules." *UR CSD; TR 617 University of Rochester, Computer Science Department*, Rochester, Ny.

VITA

Vanessa Amado graduated with a Bachelor's Degree in Civil Engineering from the Polytechnic University of Puerto Rico in 1998. She began working for Guillermet, Ortiz & Associates as a civil engineer in training before graduation. In August of 1999 she was admitted with a full fellowship into the Graduate School of the University of Missouri (MU) in Columbia, MO. In May 18, 2001 she obtained the title of Master's Degree in Civil Engineering. In the spring of the same year, she was admitted with another full fellowship into the doctoral program of the Department of Civil & Environmental Engineering at MU. She was selected as the 2001 Outstanding Student of the Year in Region 7 of the US Department of Transportation and was one of the 2002 Eno Fellows. Mrs. Amado obtained her engineer in training (EIT) license in 2003. She is a member of the College of Engineers and Surveyors of the Commonwealth of Puerto Rico (CIAPR for its Spanish acronym). Currently, Mrs. Amado is employed by CMA Architects & Engineers, LLP as a civil engineer for the firm's transportation department. She is married to Mr. Gustavo A. García and lives in San Juan, Puerto Rico.