Mining and Analysis of Traffic Safety and Roadway Condition Data

By

Dr. Sara J. Graves
Information Technology and Systems Center

Dr. Daniel Rochowiak
Department of Computer Science and Intelligent Systems Lab

and

Dr. Michael Anderson
Department of Civil and Environmental Engineering
The University of Alabama in Huntsville
Huntsville, Alabama 35899

Prepared by

UTCA

University Transportation Center for Alabama
The University of Alabama, The University of Alabama at Birmingham,
and The University of Alabama in Huntsville

UTCA Report 04310
July 2005

Technical Report Documentation Page

1. Report No

2. Government Accession No.

3. Recipient Catalog No.

4. Title and Subtitle

Mining and Analysis of Traffic Safety and Roadway Condition Data

5. Report Date

January 2005

6. Performing Organization Code

7. Authors

Sara J. Graves, Daniel Rochowiak, and Michael D. Anderson

8. Performing Organization Report No.

UTCA Report 04310

9. Performing Organization Name and Address

Information Technology and Systems Center
The University of Alabama in Huntsville
Huntsville, AL 35899

10. Work Unit No.

11. Contract or Grant No.
DRTS98-G-0028

12. Sponsoring Agency Name and Address

University Transportation Center for Alabama
The University of Alabama
Box 870205, 271 H M Comer Mineral Industries Building
Tuscaloosa, AL 35487-0205

13. Type of Report and Period Covered

Final Report: January 1, 2004 – February 1, 2005

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Public transportation systems must constantly balance the drive for improved safety of roadways with constraints of available resources. This project titled "Mining and Analysis of Traffic Safety and Roadway Condition Data" applied advanced data analysis techniques across traffic safety and roadway condition data to explore the feasibility of ultimately providing additional information to transportation policy makers. Successful results of these analyses would be useful in helping to identify potential safety problems based on data mining associations between roadway conditions collected by the Alabama Department of Transportation and traffic safety records that accompany the Critical Analysis Reporting Environment (CARE) system. This project involved researchers from the Data Mining Solutions Center, the Intelligent Systems Lab and the Computer Science and Civil Engineering Departments, at the University of Alabama in Huntsville.

The initial results are very encouraging, and have strongly indicated that data mining is a successful method to perform advanced analysis of traffic safety and condition data. By querying these attributes and their "risk" values among the pavement datasets, roadway locations where crashes are over represented can be identified and targeted for further analysis and evaluation by transportation experts.

17. Key Words

Data Mining, Traffic Safety, Pavement Conditions

18. Distribution Statement

19. Security Class

20. Security Class.

21. No of Pages

22. Price

Contents

List of Figures

Executive Summary

Decision makers working to improve transportation systems must constantly balance the need to improve roadway safety through infrastructure investment with constraints of available resources. This project explored the additional information available to transportation decision makers by integrating two independent datasets: traffic accident and roadway pavement condition data. The roadway condition data are collected by the Alabama Department of Transportation. The traffic accident data were available in the University of Alabama's Critical Analysis Reporting Environment system. The Algorithm Development and Mining [Rushing] system, a data mining toolkit developed by the University of Alabama in Huntsville (UAH) Information Technology and Systems Center, was used to examine the possible relationships between roadway data and traffic safety data. The preliminary study showed that discernable and predictable associations can be identified through analysis of the roadway condition and reported traffic accident events. The analysis results indicate that data mining is a successful method to perform advanced analysis to improve infrastructure investment decisions.

Section 1
Introduction

Accidents on our nation's roadways are a serious threat to the traveling public. Although operator or operational factors are the major cause of most automobile crashes, the road-operating environment can often contribute to an accident. Poor pavement and shoulder conditions can play a role in the occurrence of accidents. Bumps, potholes, pavement roughness and pavement edge drop-off are just a few pavement conditions that could cause difficulty for drivers. Decreasing accidents and improving roadway condition are two potentially connected goals of any transportation administration.

Often, transportation decision makers must balance the need to improve roadway safety conditions with the constraints of available resources. To assist decision makers in making crucial infrastructure investment decisions, large data collection efforts have been undertaken by the Alabama Department of Transportation (ALDOT). Data have been collected on accidents, pavements, bridges, construction, maintenance activities and more. As the size of these databases increases rapidly, both spatially and temporally, it is a challenge to analyze and extract useful information from them without using advanced data analysis tools.

Data mining, as an emerging data analysis technique, has received a great deal of attention in recent years due to the increasing ability of collecting and storing data. Data mining has been used widely to support decisions in business management, production control, market analysis, engineering design and science exploration. Currently, data mining techniques have been applied to safety [Hardin, 2003; Chong, 2004] or roadway data [Amado, 2000], but it has not been widely applied across the combination of both. Other studies [El-Seoud, 2004] have applied clustering techniques in merged datasets of traffic safety and pavement conditions in analyzing the traffic safety in the state of Florida's transportation system. GIS (Geographic Information Systems) was utilized to identify relevant freeway features at each accident location and to integrate them with the accident database. But they emphasized clustering accident data and considered only six pavement feature attributes: number of lanes, speed limit, local name, median width, median type and shoulder type. Many other key pavement conditions mentioned earlier were not addressed.

Section 2
Background

To fully utilize the database resources, the research team developed an automatic dataset integrating process to merge the pavement condition data and traffic safety data. Advanced multivariate data analysis techniques and data mining algorithms were used to determine whether these techniques could identify inherent roadway safety situations that may have been previously unidentified. This study addressed the relationships using combinations of variables such as roadway conditions, and traffic patterns using classification and association mining techniques. The pavement conditions and traffic safety data were merged spatially based upon the common key attribute of geo-location, known as the "milepost" in both datasets. Considering that pavement conditions were quite different along the different directions of the route, travel direction was also taken into consideration. In the following sections, we provide a brief description of the data mining techniques used in the work, and then explain the data preprocessing necessary for this analytical approach. Finally we describe the results of case studies and discuss the conclusions.

Section 3
Methodology

Data Mining

The increasing amount of data collected and stored in large, and numerous data bases, has far exceeded human ability for comprehension without the use of powerful tools [Han, 2001]. Consequently, important decisions are often based not on the information-rich data stored in databases, but rather on a decision maker's intuitions due to the lack of tools to extract the valuable knowledge embedded in the vast amounts of data. This is why data mining has received great attention in recent years. Data mining refers to extracting or "mining" knowledge from large amounts of data [Han, 2001]. It can be viewed as an essential step in the process of knowledge discovery in databases. This is different from traditional statistical analysis, which typically deals with a relatively small dataset with questions in mind when collecting the data and developing models seeking answers [Hand, 1999]. Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis [Han, 2001]. General data mining principles, including association rules, sequential patterns, classifications, predictions, and clustering, can be applied to many areas.

ADaM (Algorithm Development and Mining) is a data mining toolkit designed and developed in the Information Technology and Systems Center at the University of Alabama in Huntsville (UAH). It provides classification, clustering and association rules mining methods that are common to many data mining systems. The toolkit consists of over 75 interoperable mining and image processing components. Each lightweight and autonomous component is provided with a C++ application programming interface (API), an executable in support of scripting tools (e.g. Perl, Python, Tcl, Shell). ADaM is extensible and scalable, and has been successfully used in several diverse data mining applications [Rushing, 2004]. ADaM is the primary data mining tool used in this pavement condition and traffic safety association study. In particular, the association rules algorithm in ADaM was used to discover the pavement conditions that occur together frequently in the given accident and pavement condition datasets.

Data Preparation

A data mining system normally consists of a sequence of steps, starting with data cleaning, data integration, data selection and transformation to data mining and knowledge presentation. In this section of the report, these steps will be addressed in detail regarding the traffic safety data and pavement condition data. All input data were ultimately transformed into ARFF (Attribute-Relation File Format), for compatibility with ADaM and other data mining tools.

Data Sources and Description

The traffic safety data used in this study were obtained through the CARE system. CARE (Critical Analysis Reporting Environment) is a data analysis software package designed for problem identification, countermeasure development and limited-dimension data mining purposes for traffic safety information by the staff of the CARE Research & Development Laboratory (CRDL) at the University of Alabama. By querying the crash reports on certain counties, detailed accident records on each route can be extracted. These safety data constitute a comprehensive record with up to 231separate attributes. For state routes and Interstate highways, the CARE system has accident location as a distance to the nearest roadway milepost, with a resolution of about 0.1 miles. This is understandable considering the confusing scenario at accident sites. The pavement condition datasets were provided by ALDOT. They included the roadway pavement condition surveys of years 2000 and 2002 with up to 72 attributes. The pavement conditions include such attributes as roughness, elevation, shoulder type, etc. Since pavement data collection is automated using equipment with advance scanning technology, it has a resolution of 0.01 miles or higher.

Data Cleaning and Integration

Noise and inconsistent data are inevitable in large data collections. In accident reports, all information regarding the accident is recorded manually on site. Some important information might be missing, such as milepost information and travel direction of the involved vehicles. As mentioned earlier, milepost and travel direction are two important keys used in merging the traffic safety data and pavement condition data. Hence, the records without geo-location information were useless for this study, and were discarded during the data preprocessing.

The pavement condition and traffic safety are two completely independent data sets, and are collected and maintained by different transportation administration groups. The same attributes will have different names. The milepost attribute in the safety data is “M__POST,” while in pavement conditions data it is “COND_MILEPOST”. The travel direction attribute in the safety data is “DIR OF TRAVEL_VEH C”, with numbers of 1, 2, 3 and 4 representing four directions (N, S, W, and E), while in the pavement condition data, it is “COND_DIRECTION” with “N, S, W, E” representing four directions respectively. These disparities introduce difficulties in using simple queries to merge these datasets. So, automated procedures were developed to merge the two datasets based on the spatial location information, together with the data cleaning process.

Considering the use of coarse-resolution milepost information for accident location in the accident report, pavement data aggregation was also conducted during the data preprocessing. The maximum, minimum and median values of all the pavement attributes were computed within 0.1 miles of each accident location. The worst case among these values was used in data mining analysis, so the nearby conditions of the accident location could be considered.

Data Categorization

The attribute values of pavement conditions are in both categorical and numerical values. Since the goal was to correlate pavement conditions with accident data, numerical-valued attributes were categorized, such as IRI (International Roughness Index), patching size, elevation, etc. The available pavement datasets were analyzed and the minimum and maximum values were found for all the numerical attributes. Ten categories were used and were equally distributed from the minimum to the maximum.

Section 4
Project Results

In the preliminary study, data were collected for complete roadways spanning Alabama, while at the same time striving to include roadway diversity. To honor privacy concerns, specific roadway identification information was sanitized from all reports, since this effort was conducted to suggest the development of possible future decision making tools rather than focus on specific concerns at this point.

ALDOT provided roadway data concerning distress characteristics of pavement sections for all federal and state routes in Alabama. The data elements contained in the database included location, pavement type, pavement condition, number of lanes, lane width, and shoulder type. Since it would be difficult to study every pavement section in this limited study, two US highways (indexed as route A and C) were selected, one state highway (indexed as route B) and one interstate highway (indexed as route D). Each of these routes spans at least 250 miles across the state. The traffic safety datasets for each of these routes were extracted from the CARE system. The traffic safety and pavement datasets were merged for each route using the procedures explained previously. The final datasets were converted into ARFF formatted text files for compatibility with the ADaM and other data mining tools.

The preliminary analysis compared the occurrence rate (OR for simplicity) of each value of each pavement condition among the accident locations and among the entire pavement condition dataset of each route. The research team concluded that if the ORs were the same for a value of a pavement condition, then that condition played no role regarding the cause of accidents. However, if the OR of an attribute among the accident locations was higher than its corresponding rate in the pavement condition dataset, then this pavement condition may be a good candidate for pavement rehabilitation. The assumption was that the pavement data were sampled evenly across each of the routes.

Association rule mining was applied to the pavement condition datasets of selected routes and ORs were computed for each value (category) of each attribute. The same algorithm was applied to the merged datasets, i.e., the dataset of pavement conditions at the location of accidents. This gave the ORs of each category value of each pavement condition among the accident locations. Further analysis among these “flagged” attributes with higher ORs might help decision makers identify the location(s) with those pavement conditions for a certain route.

Rutting

Rutting, identified in the pavement database as RRUT and LRUT, is an indication of surface depressions. Rutting is a potential safety concern as ruts pool water, leading to vehicle hydroplaning, and tend to pull a vehicle towards the rut path as it is steered across the rut.

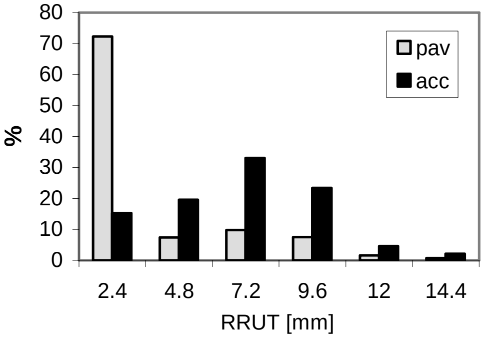

Figure 4-1 Right side wheel path rutting

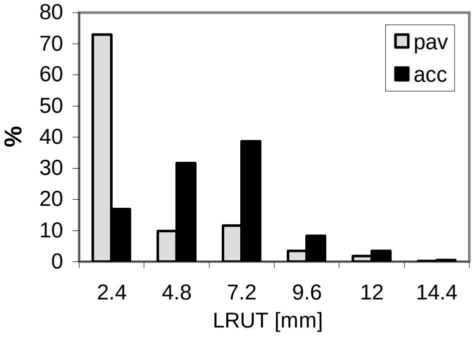

Figure 4-2 Left-side wheel path rutting

Figures 4-1 and 4-2 show the frequencies of RUT values for Route A. Figure 4-1 is rut of right side wheel path and Figure 4-2 is rut of left-side wheel path. The black bars are the frequency of occurrences of each RUT category among the accident locations; the gray bars are the corresponding frequencies of pavement condition along the entire route. As one can see, the higher RUT values occur more frequently among crash sites than the corresponding occurrence along the route. For example, in Figure 4-1 rut values between 4.8 mm and 7.2 mm occurred more than 30 percent among crash sites, while this category only occurred about 10 percent of the time along the route.

Shoulder Type

Shoulder type can have a big effect on automobile crashes [Web]. Having no shoulder or bad shoulder conditions can cause difficulty in operating vehicles in certain circumstances, and could potentially lead to an increase in accident severity as avoidance or escape routes are minimized. Figures 4-3 and 4-4 show the result of difference shoulder types among the accident locations. A shoulder type of zero was considered as not recorded. Among all shoulder types, the grass type had much higher occurrence rate (~ 37 percent) among the accident locations than its corresponding distribution rate (~ 20 percent) along the route. Grass shoulders are often found on low traffic volume, older roads that have steep hills, sharp curves, narrow pavements, etc. They are generally more susceptible to crashes.

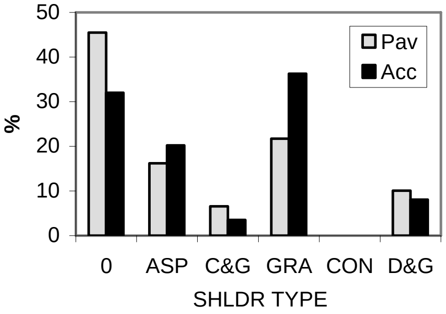

Figure 4-3 Shoulder type of route A

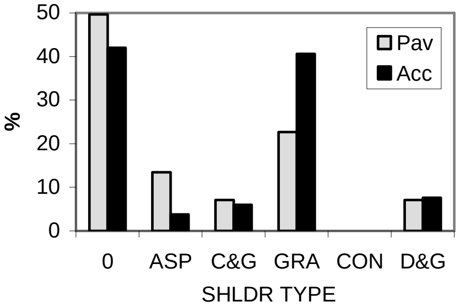

Figure 4-4 Shoulder type of route C

Cracking

Cracking is another index of pavement condition. Figures 4-5 and 4-6 show the result for routes A and B for transverse cracking of severity level 2 on asphalt and concrete pavements. Although more than 70 percent of the route had a count less than or equal to one, accidents were typically more likely to occur in locations with cracked pavement.

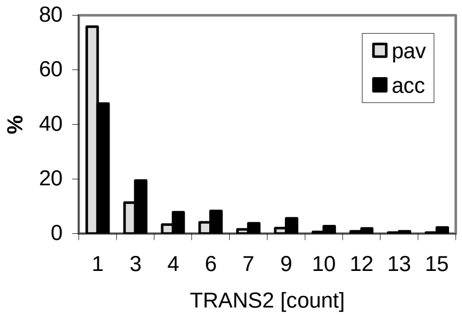

Figure 4-5 Transverse Cracking of A

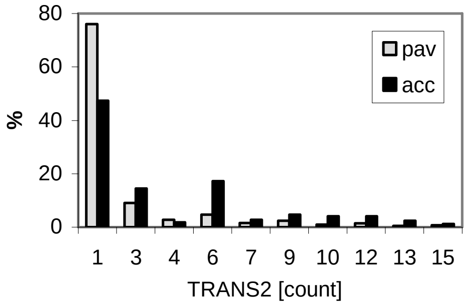

Figure 4-6 Transverse Cracking of B

Roughness

Pavement roughness is an expression of irregularities in the pavement surface that adversely affect a vehicle's ride quality. Every department of transportation in the nation uses the IRI to quantify roughness. IRI values typically range from 0 to 1267 inches/mile. Higher values indicate increased roughness, an unfavorable vehicle operating condition. Figures 4-7 and 4-8 represent the analysis of IRI for route D. Figure 4-7 is for IRI1, the average left IRI and Figure 4-8 is for IRI2, the average right IRI. Route D is an interstate highway and most of the IRI values were in the good value range of less than 100 inches/mile. However, it was apparent that higher roughness values had higher occurrence rates among the accident locations. Higher roughness conditions could be one of the factors contributing to the higher crash rates.

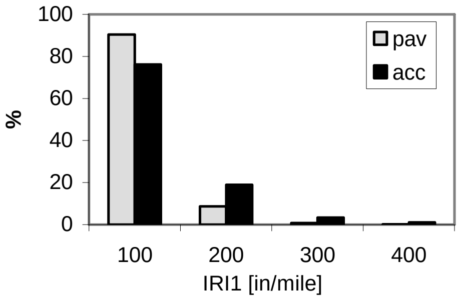

Figure 4-7 Left side of route D

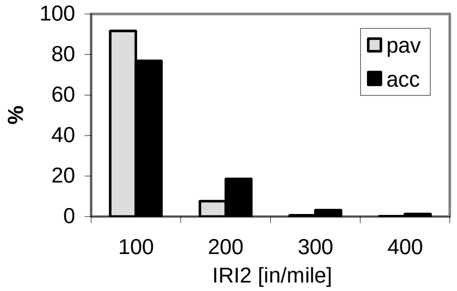

Figure 4-8 Right side of route D

Section 5
Conclusions and Recommendations

This research effort involved merging and associating two different data sources, traffic safety data and pavement condition data, for individual routes based on geo-location information inside these databases. By applying data mining techniques, certain pavement conditions were identified in which there were a disproportionate number of crashes. Preliminary analysis has shown pavement conditions whose values, or value ranges, have much higher occurrence rates among crash locations than the occurrence rates along each of the selected routes. The crash rates on each selected route were not considered due to the data problem mentioned earlier, i.e., a portion of the accident records has to be discarded due to the lack of geo-location information.

Though the data analysis was preliminary at this time, the results were very encouraging, and have strongly indicated that data mining is a successful method to perform advanced analysis on traffic safety and condition data. By querying these attributes and their "risk" values among the pavement datasets, locations were crashes are over-represented can be identified and targeted for further analysis and evaluation by transportation experts. This approach seems capable of providing valuable information for transportation decision makers in roadway administration and maintenance.

In future projects, more attributes from accident data could be considered for additional analyses, such as traffic volume, driver's information, weather conditions, etc. Currently, traffic volume is likely the strongest predictor used for high number of accidents. This may be because high volumes of heavy trucks causes accelerated deterioration of pavement conditions, such as roughness, rutting, cracking, etc. It would also be very interesting to explore whether there are any pavement conditions causing more driving difficulties to one age group than to others.

Suggestions

The quality of the collected data is essential to the data mining and analysis results. As mentioned in the early sections, problems were encountered in the traffic safety data since some records had missing attribute values. This significantly reduced the amount of data available for the analysis. Any record without geo-locations is useless in roadway safety analysis of the type associated with this project. Improved methodologies are needed for collecting accurate geo-location information at accident sites.

Consistency in name and value interpretation of the attributes across different datasets can ease the data preparation process. Improvements towards resolving the inconsistencies existing with the databases utilized in this study would significantly improve the support for future studies.

Section 6
References

Amado, V., “Expanding the Use of Pavement management Data,” *2000 MTC Transportation Scholars Conference,* Ames, Iowa. 2000.

Chong, M., “Traffic Accident Analysis Using Decision Trees and neural Networks,” *IADIS International Conference on Applied Computing*, Portugal, IADIS Press, Pedro, 2004.

El-Seoud, M, Elbadrawi, H. “Data Mining and GIS Technologies to Support Highway Safety Management Systems,” *IAMOT* 2004.

Han, J. and Kamber, M., Data Mining: *Concepts and Techniques,* Academic Press, ISBN 1-55860-489-8. 2001.

Hand, D. “Statistics and Data Mining: Intersecting Disciplines,” ACM SIGKDD, June 1999.

Hardin J. M. and Conerly, M. “Traffic Safety Analyses: A Data Mining Approach,” *UTCA Project # 02115*. 2003.

Hill, S. and Jay K Lindly, J. “Red Light Running Prediction and Analysis,” *UTCA Project # 02112*. 2003.

Randy K. Smith, “Data Mining to Improve Traffic Safety,” *UTCA Project # 04107*. 2004.

Rushing, J., et al, “ADaM: A Data Mining Toolkit for Scientists and Engineers,” *Computers and Geosciences,* accepted in Nov. 2004.

Web, New York Court Finds That Defective Shoulder Caused Death in Automobile Accidents, URL: *http://www.usroads.com/journals/rilj/9702/ri970202.htm*