

EXPANDING THE USE OF PAVEMENT MANAGEMENT DATA

Vanessa Amado
Department of Civil and Environmental Engineering
University of Missouri – Columbia

ABSTRACT

The process of collecting pavement data has been evolving with advances in technology, thus generating huge amounts of data to be stored in pavement management systems (PMS) databases. The rapid size increases of these databases presents a challenge for state agencies, as they attempt to understand and take advantage of the data to support pavement maintenance and rehabilitation decisions. The knowledge discovery (KDD) process and data mining are being applied to large pavement data sets with the objective of extracting useful information. For example, the data can be used to predict pavement serviceability ratings (PSR), as an indirect way of obtaining the remaining life of pavements. The Missouri Department of Transportation (MoDOT) provided pavement condition data from 1995 to 1999 to be used in the study. Research is being conducted in which results from the whole set of data will be presented and interpreted in order to obtain a better view of the condition of pavements in Missouri and be able to increase the effectiveness of the decision-making process. Parameters needed for future work that would improve the quality of information extracted are also presented.

INTRODUCTION

With advances in technology, more sophisticated equipment is being developed and adopted for the collection of pavement data. Not only has the equipment changed, but decision-making systems are developing to take advantage of the new technology.

Transportation agencies have always managed their pavement. However, it was not until the enactment of the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991 that it was required that all states have a pavement management system (PMS). Although the requirement was amended in 1995 by requiring that only all roadways eligible to receive Federal-Aid Funds would be covered by a PMS, most agencies have and use a PMS (1). The Federal Highway Administration (FHWA), the American Association of State Highway and Transportation Officials (AASHTO), the American Public Works Association (APWA), and the World Bank have provided guidance and encouragement to aid the transportation agencies in the process of developing a more comprehensive and functional PMS structure (1).

An effective PMS must have a comprehensive database. The database must contain reliable, objective and appropriate information in order to assist in any decision-making for planning and budget procedures. In addition, it must contain an inventory of the state's highways. With the increase in automated data collection due to both the developments in technology and growth of the highway system, the size of these databases has increased as well. Therefore, new methods or techniques are needed to assist analysts in the process of discovering useful information and knowledge in these databases (2).

Knowledge discovery in databases (KDD) is a process that provides the techniques and tools needed to understand large data sets. Such techniques provide "intelligent" assistance to humans in order to improve the ability to understand these data sets. The term "data mining" describes the methods used in some of the steps of the KDD process (3). The application of the KDD process and data mining to pavement management data will enable the identification of characteristics common to certain pavements with the goal of predicting future performance of pavements.

The primary objective of this research is to apply a systematic approach to a large pavement management database to extract useful information from pavement management data. This additional information has the potential to further increase the accuracy and efficiency of the decision-making process. In addition, the research will help agencies optimize the use of the data in the PMS databases, and evaluate examples of real data measured and analyzed by state Departments of Transportation (DOT).

The paper provides background on KDD and data mining, PMS, and data integration, followed by a description of the research approach and a case study using data from the Missouri Department of Transportation (MoDOT).

BACKGROUND

The following sections give a summary of the type of problems that data mining is often used to solve, an overview of the concept of PMS, a brief description of the evolution of the data collection process, some examples of the equipment used for the collection of pavement condition data, and a summary of the process of data integration in a PMS.

KNOWLEDGE DISCOVERY AND DATA MINING

In general, the term "data mining" includes all tools engaged to help users analyze and understand their data. However, a more specific definition is given by Moxon, "a set of techniques used in an automated approach to

exhaustively explore and bring to the surface complex relationships in very large databases” (4). The purpose of data mining is to help find useful patterns in large databases by means of algorithms (5). Its applicability to a variety of problems, such as databases containing consumer and transaction information and advanced databases on multimedia, has recently raised the interest of research in this area (6). However, with the rapid growth of databases in the last decade, our abilities to understand and explain the data have decayed, creating opportunities for the use of knowledge discovery in databases (KDD).

KDD is a process that provides the techniques and tools needed to assist in the understanding of large data sets. The techniques are said to provide “intelligent” assistance to humans in order to improve the ability to understand these data sets (3). Fayyad goes on to describe the relationship between KDD and data mining, saying that, “...KDD refers to the overall process of discovering useful knowledge from data while data mining refers to the application of algorithms for extracting patterns from data...” (3).

Data mining typically has one of two goals, prediction or description. The term “prediction” refers to the process of using variables from the database to predict unknown values for the variables of interest, and “description” refers to the process of finding “human-interpretable” patterns to describe the data (3). Both are useful in civil engineering applications, for example, in modeling pavement deterioration or in looking for common characteristics of damaged pavements.

Some of the more specific tasks that data mining algorithms can perform on databases include (6):

- **Association** – finding relationships between different attributes, often in large customer databases. For example, an association might be that a customer who buys item A is also likely to buy item B. The idea is to figure out how to determine the presence of some sets of items, given the presence of other items in a transaction;
- **Clustering** – grouping data into ad-hoc categories with considerable similarities between items in the same group. These algorithms are mostly used for descriptive tasks. Such algorithms give the user a good idea of both the identity and nature of the similarity of the different points in each cluster; and
- **Classification** – dividing the data into pre-defined classes that can be either categorical or quantitative. In the quantitative context, classification is viewed as a type of regression. This type of algorithm is generally used in prediction problems since it creates a model based on known data.

Other data mining techniques that are generally accounted for in the classification model include:

- **Decision Tree** – used to examine the data and induce the tree and its rules that will be used to make predictions. It involves "splitting" the information into categories containing examples of a particular class with the purpose of maximize the “distance” between groups at each split (7);
- **Neural Networks** – used as a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions. It involves a network of neurons to process one or more values into a single output, which is called the “class label”; and
- **Bayesian Classifiers** –used mostly for applications in which there is a large number of variables, for example, text classification. It assumes that the outcome of an attribute value of a given class is independent of the values of the other attributes.

Pavement Management Systems (PMS)

Pavement management systems (PMS) are the key instruments that provide decision-makers at all management levels with strategies to maintain pavements in acceptable condition. Various activities are involved in the process, such as planning, design, monitoring, maintenance, reconstruction, budgeting and programming, construction, research, evaluation, and rehabilitation. A comprehensive and efficient PMS will permit agencies to achieve both national and local objectives in delivering acceptable highway services (1).

AASHTO describes a pavement management system as “a set of tools or methods that can assist decision-makers in finding cost-effective strategies for providing, evaluating and maintaining pavements in a serviceable condition” (8). However, selecting the data that will support those decisions must meet certain criteria. That is, even though the tools for collecting huge amounts of data are available, the data to be collected should be chosen cautiously. According to Paterson and Scullion, these criteria deal with relevance, reliability, affordability, and appropriateness (9). The reliability of the data is a key element in a PMS database. The accuracy, spatial coverage, completeness, and currency are the factors that determine such reliability.

Pavement Data Collection

The main goal of highway departments is to provide roadway users with an “acceptable” highway system at the most inexpensive cost. Therefore, a series of investigations and simulations are performed based on the data collected to better understand the behavior of pavements, as well as to improve the conditions of the roadway system. Another use of these data is to establish priorities for the maintenance of each roadway. Therefore, the

collection of pavement condition data facilitates the decision-making process, by providing insights used for highway maintenance, rehabilitation, and reconstruction.

Historically, pavement management decisions have relied on data collected through visual inspection and judgment based on experience (10). Public agencies have found that manual inspection of pavements requires extensive effort and that it is difficult to pass experience along to younger engineers. The collection of pavement condition data is costly and time consuming; as a result, the use of automated methods for data collection has been expanding. Such methods are designed to meet the evolving management requirements of today.

At present, more state of the art equipment is being used, in which data is being collected automatically. However, some manual collection and visual inspection of pavement data is still conducted as integrated parts of the same process.

The type of data collected can be grouped in four major areas (10):

- **Roughness** - irregularities found in the surface of pavements that affect the ride of a vehicle. This type of data is the only measure used by some agencies for network level applications, in order to evaluate the serviceability of pavements.
- **Surface Distress** – the extent of pavement fracture, distortion, and disintegration. The three major components of pavement distress are cracking, rutting, and longitudinal profile. It is used by agencies to evaluate the deterioration, the overall composite index, and the maintenance needs of pavements for both network and project level applications.
- **Structural Evaluation or Deflection** - the ability of a pavement to support traffic. It is used by agencies for the evaluation of material properties and structural capacity of pavements for both network and project level applications.
- **Skid Resistance or Pavement Friction** - “the horizontal force developed when a tire that is prevented from rotating slides along the pavement surface” (11). It is used for the evaluation of safety against skidding in both network and project level applications.

Other types of data, such as ride quality, appearance, traffic, costs (e.g. construction, maintenance, user), location reference, geometric and structure data, and environment (e.g. climate, pavement temperature, drainage, water below surface, freeze/thaw), are also monitored and evaluated by agencies (12).

Automated Data Collection

Agencies have been changing their methods of collecting pavement data from manual to automated, as technology has evolved. Another significant factor in the use of more automated methods for the collection of data has been the growth of the highway system. The staff needed to inspect and collect data manually from today's system would be tremendous, but trying to accomplish it within a reasonable time frame would be even more challenging. Other benefits of automated data collection are the accuracy of the data and efficiency of the method. In addition, it allows easy and flexible output of multiple parameters (13). Therefore, the implementation of an automated data collection method ensures a more comprehensive database, which is the key element of a pavement management system.

Boettcher explains that an automated method of data collection is needed to “meet evolving management requirements” (14). That is, with the development of new technologies, more advanced techniques for PMS will be employed, and thus better methods of data collection will be developed.

Equipment

There are several types of automated equipment for collecting pavement surface distress and/or pavement structural strength information. Some of the equipment that are currently being used by state DOTs include:

- **Profilometer / Rutbar Van** – collects ride quality data using lasers and transverse profile data using ultra-sonic sensors,
- **Skid Systems** – measures surface friction characteristics,
- **Falling Weight Deflectometer (FWD)** – collects pavement surface deflection data,
- **Ground Penetrating Radar Van (GPR)** – collects data, in a non-destructive manner, to determine the upper surface layer thickness, and presence of certain types of pavement anomalies such as stripping or presence of moisture (15),
- **Multi-Function Vehicle (MFV)** – collects video images of the pavement surface and the right-of-way, and other information, such as ride quality, rut depth, and global positioning system (GPS) data, and
- **Automatic Road Analyzer Vehicle (ARAN[®])** – collects longitudinal and transverse profiles, grades, cross-slope, pavement texture, pavement distress, GPS coordinates, and video images of the right-of-way and pavement in a single pass (16).

The practice is to combine some of these equipment to gather the pavement surface distress and pavement structural strength information in order to obtain a complete set of data.

Data Integration

Data integration refers to the issues related to obtaining more information from the data collected (17). The general issues include data collection technologies and practices. Technologies are combined to improve the quantity and/or quality of the data collected. For example, GPS coordinates are being used to link the data collected from the ARAN[®] vehicles to state highway locations to reproduce the location of “link points” between the two data sets (18). Likewise, different agencies use the data collected from pavements in a variety of combinations and circumstances.

The use of new technology as a consequence of the need for more accurate and reliable pavement data has raised questions about how much data should be collected; that is, since computing power and memory are no longer significant constraints, other criteria must be used to determine the type and frequency of data collection. Likewise, concerns related to the historical data in the PMS databases have also raised some questions, such as what type of data are being used for which decision-making processes, what patterns can be extracted from the data, and from which data can these patterns be extracted.

RESEARCH APPROACH

As part of the knowledge discovery process, data mining is applied to extract patterns of information in PMS databases, which can be used to predict the present serviceability ratings (PSR), as an indirect way of obtaining the remaining life of pavements. The process consists of building the data mining database by exploring the data and preparing it for the given applications, evaluating the different types of data mining applications, building the model, and evaluating the model. Figure 1 illustrates the knowledge discovery process as described in the paper.

Building Data Mining Database

Building the data mining database, together with exploring and preparing the data to be mined, takes anywhere from 50 percent to 90 percent of the time and effort in the entire knowledge discovery process. The process of building the database is usually divided in eight tasks (7):

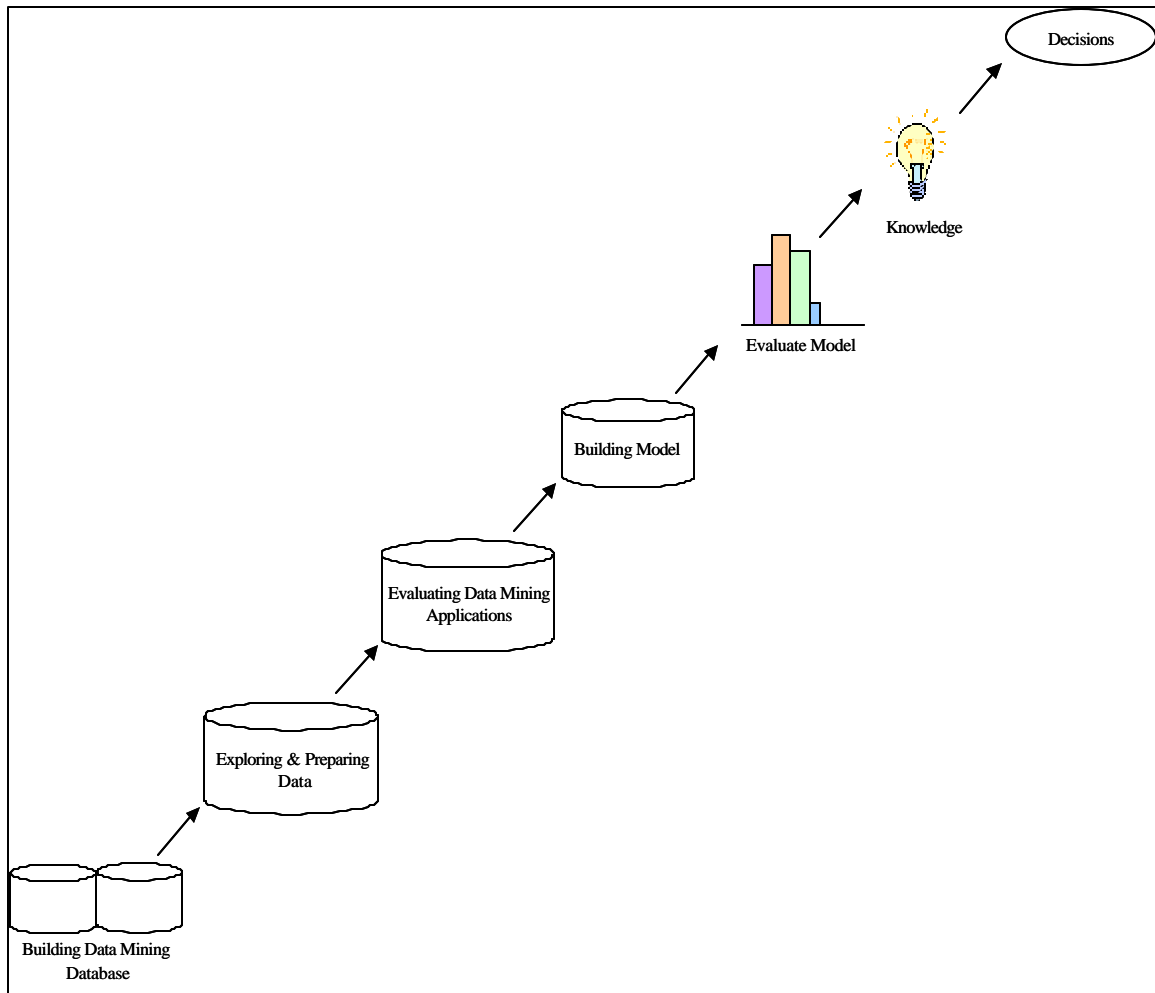


Figure 1. Knowledge Discovery Process

1. *Data collection* – identifies the sources of the data to be mined. The data needed may not be available, in which case the process for collecting it must be arranged. Data collection usually includes items such as the source of data, owner, cost (if purchased), size of tables (i.e. records/attributes), size of file(s), physical storage, security requirements, restrictions on use, and privacy requirements.
2. *Data description* – describes the content of each file. In the description of data, a list of items, for example, number of records/attributes, number or percentage of records with missing values, and field names, is usually presented. For each field name, a list of descriptions is usually provided as well. Some of the items in the list may be data type, definition, description, source of field, unit of measure, range of values, list of values, and time frame.

3. *Selection of data* – selects subsets of data to be mined. It is the “gross elimination” of irrelevant or unnecessary data, which is a very important task when limiting factors like space and time are a concern.
4. *Data quality assessment and data cleansing* – identifies characteristics of the data that will affect the quality of the model. Missing values or incorrect values account for many of the problems encountered while performing a data mining application, and these need to be fixed or accounted for in order to provide reliable results.
5. *Consolidation and integration* – combines data from different sources into a single mining database and requires the differences in data values to be reconciled. For example, U.S. dollars cannot be combined with Canadian dollars without a conversion, because there would be inconsistencies and the knowledge discovery process would be of a poor quality.
6. *Metadata construction* – provides information that will be used in the creation of the physical database as well as information that will be used by analysts in understanding the data and building the model. Kimball refers to metadata as having a “back-room that guides the extraction, cleaning, and loading processes, and a front-room that makes query tools and report writers function smoothly” (19).
7. *Load the data mining database* – stores the data in databases. Depending on the quantity and complexity of the data, it may require storage in a database management system (DBMS) as opposed to a flat file.
8. *Maintain the data mining database* – preserves/cares for the data, for example, making backups of parts of the process already completed and reorganizing the data to reclaim disk storage or to improve performance.

Exploration of the Data

Exploring the data to be mined is an extremely important part of the process of building the data mining database. It consists of identifying the most important fields/attributes in predicting an outcome and determining which derived values may be useful. Since the data sets may contain hundreds of columns (fields), exploring the data can be as time consuming and labor intensive as it is informative. It is extremely important to have a good interface and fast computer response because the nature of the exploration will depend on the time it takes for results to be obtained. Depending on the system, this task can take anywhere from seconds to minutes or even hours.

Data exploration is basically used to discover and evaluate appropriate problems in the data, define solutions and implement strategies, and produce measurable results. It involves a series of stages, including exploring the problem space, exploring the solution space, specifying the implementation method, and mining the

data. Each of the stages demands a certain amount of time, and usually the importance of these stages to success is inversely proportional to the time consumption during the exploration (7). That is, by just exploring the data set one can determine if the data are “noisy”; this process might take only minutes, but the advantages of knowing the data quality will be enormous.

Preparation of the Data

The process of preparing the data is frequently confused with the process of exploring the data. Even though one process is needed to obtain the other (that is, it is hard to define how to extract value from the data mining activities that follow without identifying the problem to solve), these are two very different processes (20). The preparation of data involves the selection of variables and rows, the construction of new variables, and the transformation of variables with the purpose of manipulating and transforming raw data into data that are more easily accessible. Variables and rows are selected carefully to minimize the time it takes to build a model and to optimize the output obtained from the model. The construction of variables is useful since some variables have little effect when used alone and might need to be combined with other variables. One way of constructing variables is by using algebraic operations, like addition, subtraction, and ratios. Likewise, the transformation of variables is useful to represent the data according to the tool one chooses; for example, variables may be scaled to fall within limited ranges, such as 0 to 10. Another example of variable transformation would be the variable classification one gives to certain fields; for example, the records in a field of *last year worked* might need to be classified as categorical data in order for the application to understand the value of the record as a year.

Building the Model

Model building is an iterative process, in which alternative models are explored to find the most useful model for solving the problem in question. Models replicate some useful features of the original object in some “more-convenient-to-manipulate” way. Furthermore, models can be computationally manipulated to answer questions posed about the model’s behavior, and thus, about the behavior of the real world (20). By manipulating the model, engineers, scientists, and economists can determine how well the proposed explanation “works”.

The first part of building the model is deciding what type of data mining task is needed, that is, descriptive or predictive. The next step is to choose a model type for making the corresponding task. There are several model types, such as decision trees, neural networks, and rule induction, and the model type chosen will influence the kind of data preparation that must be conducted.

The two requirements for building a predictive model are training and testing the data mining model. Training and testing the data is usually referred to as supervised learning. It involves training or estimating the model on a portion of data, and then testing and validating it on the remainder of the data. The resulting accuracy rate of the test database is used to estimate how the model will perform on future databases that are similar to the training and test databases. It does not guarantee that the model is correct, but rather that if the same techniques were used on a series of databases with similar training and test data, the average accuracy would be close to the one obtained. Training stops when the accuracy rates on the test database no longer improve with additional iterations.

Evaluation of the Model

Once the model is built and validated, it can be used by analysts to recommend actions based on its results or to apply the model to different data sets. Remembering that the accuracy rate applies only to the data on which the model was built, it is important to learn more about the type of errors found and the costs associated with them. By evaluating and interpreting the model one can learn about the types of error as well as the accuracy rate of the model.

CASE STUDY

The methods described are being applied to a large pavement condition data provided by the Missouri Department of Transportation (MoDOT). The objective is to predict the PSR value as an indirect approach to obtaining the remaining life of a pavement. For PSR values, a pavement with a threshold of 24 is considered for replacement (21). Pavement condition data from 1995 to 1999 from MoDOT was provided for the study. MoDOT gathered a set of data from two different databases, data collected with the ARAN® van and structural data from a separate database, to present a better set of data given the objectives of the study. The data were provided in a database format consisting of 28,231 records and 49 fields.

Modeling

The first step in building the model was to convert the database file to an Excel® file in order to facilitate the preparation and exploration process of the data. Table 1 shows the field names, definitions, and types of data present in each field. The data types are related to the appropriate scales of measurement, for example, *gender* would be the measurement and *categorical* would be the scale type for that measurement. During the exploration of the data set, several problems were identified. It had three extremely noisy columns, NHS, LSHLDTYPE, and RSHLDTYPE, and missing data for every other year from the RIDE, CRACKING, RUTTING, PSR, and

CONDITION fields. The three noisy columns were not considered relevant for the objective of the study, and therefore they were not selected for the model. A first attempt at restoring the missing data was made by using the values from previous years in the place of missing values. A second attempt will be made by using the average of previous and subsequent years of the same fields. In addition, 54 percent of the whole data was found to be a better representation for the model. That is, from the five-year condition data reported, only the 1999 PSR data was fully complete and the second most complete set of data was the 1998 PSR data. Therefore, the selection of records was based on the amount of complete data for 1998. The advantage of making this selection was to obtain a higher percentage of good data to build the model.

Table 1. Pavement Condition Data from 1995 to 1999

Field	Meaning	Type
DIST	District	Numeric
CNTY_NO	County number	Numeric
RTE_DESIG	Route designation	Categorical
ROUTE	Route number	Categorical
DIR	Direction	Categorical
CONT_BLOG	Mileage from the beginning of county	Numeric
CONT_ELOG	Mileage from the end of county	Numeric
AREA	Area (i.e., urban/rural)	Categorical
ST_SYS	Street system	Categorical
FC	Functional classification	Categorical
LANES	Number of lanes	Categorical
TRAV_WAY	Width of travel way	Numeric
ORIG_SURF	Original surface	Categorical
OVERLAYS	Quantity of overlays	Numeric
SURF_TYPE	Surface type	Categorical
YR_LSTWK	Year last worked	Categorical
AADT	Average annual daily traffic	Numeric
RIDE (95 to 99)	Ride quality	Numeric
CRACKING (95 to 99)	Cracking	Numeric
PSR (95 to 99)	Present serviceability rating	Numeric
RUT (95 to 99)	Rutting	Numeric
CONDITION (95 to 99)	Condition	Numeric
DESGN_CLS	Design classification	Categorical
DIVCLS	Division classification (i.e., divided/undivided)	Categorical
LSHLDWIDTH	Left shoulder width	Numeric
RSHLDWIDTH	Right shoulder width	Numeric
NHS	National highway system	No Data
LSHLDTYPE	Left shoulder type	Categorical
RSHLDTYPE	Right shoulder type	Categorical

As part of the data preparation process, the software tools to be used were selected. The IBM Intelligent Miner for Data (Intelligent Miner) was selected according to its connectivity features and data mining characteristics. The features of the IBM Intelligent Miner for Data provide the use of large sets of relational data to

be mined. The data sources that the software accepts are ASCII text, DBase, Oracle, and Sybase. The software is able to perform several data mining applications, such as prediction, data preprocessing, regression, classification, clustering, and associations. In addition, it uses decision trees and neural networks for its discovery methodology (22).

Based on the features provided by Intelligent Miner, the data for the model were converted to a text file. The text file was then corrected so that it would not include any tabs; this was an extensive part of the process given the number of records and the irregularities of each field in the data. Once the tabs were eliminated, the data set was ready for the applications of the Intelligent Miner.

The three data mining applications considered for the research are association, neural clustering, and tree classification. The basic understanding of the data mining applications used in the study is that they should be able to produce a description summarizing the characteristics of pavements, that have a PSR greater than 24, for example; and that they should be able to compare two groups of pavements, such as those which have PSR greater than 24 and those which have PSR less than 24 (23).

Given that associations are used to show attribute-value conditions that occur frequently together in a given set of data, this method will be used to find the PSR value of a pavement given certain characteristics of a pavement, such as YR_LSTWK, CONDITION, RUTTING, RIDE quality, number of OVERLAYS, AADT, SURF_TYPE, etc. Neural clustering applications are used to find centers for each cluster once the records are grouped together due to similar characteristics. This type of data mining application will be used to find the PSR of a pavement given the particular characteristics of such pavement and the similarities it has with a given cluster, i.e. a new pavement can be distributed to the cluster whose center is the most similar.

Since tree classification applications create a model based on known data, this method will be used to classify pavements as belonging to a “vulnerable” group (i.e. $PSR > 24$) or not belonging to a “vulnerable” group (i.e. $PSR < 24$). The process for the tree classification is divided into three parts, training mode, testing mode, and application mode. In training mode, a mining run learns the fields of each of the defined pavement “vulnerable” classes. Testing mode will then be used to test the accuracy of the model created in the training mode by applying this model to test data with known “vulnerable” pavement classes, and application mode will then be used to predict which pavement will obtain a high PSR value in the future.

Preliminary Results

Due to the extensive process of data preparation, the results obtained in this paper are from 32 percent of the portion of data selected for the completion of the research from the condition data collected in 1999. Even though the results obtained here are from a small segment of the entire data set provided by MoDOT, it is still a good representation of the condition of pavements for the State of Missouri since it consists of 4816 records.

The results indicate that 89 percent of the pavements were in “vulnerable” conditions in 1999; that is, having PSR values greater than or equal to 24. The tree classification method was used to study the data. PSR was used as the class label, and RUT, RIDE, COND, CRCK, YR_LSTWK, AADT, OVERLAYS, ORIG_SURF, and SURF_TYPE were used as active fields. Table 2 shows the preliminary results obtain from the condition data collected by MoDOT in 1999.

Table 2. Preliminary Results from the Condition Data Collected in 1999

		PSR > 24	PSR < 24
		Range of Values	
Rut	Min.	0.255 – 0.295	0.195 – 0.285
	Max.	0.295 – 0.345	0.325 – 0.335
Ride	Min.	4.435*	2.295*
	Max.	6.15 – 7.49	4.435*
Condition	Min.	18.45*	18.45*
	Max.	**	**
Cracking	Min.	3.445 – 3.805	3.285*
	Max.	3.805 – 4.050	3.75 – 4.050

* Constant values

** No data was found

The rest of the parameters, YR_LSTWK, AADT, OVERLAYS, ORIG_SURF, and SURF_TYPE, need further study to obtain better information.

The results shown here are preliminary results that are intended to demonstrate the type of information expected, further research is being conducted in which results from the whole set of data selected for building the model will be presented and interpreted in order to obtain a better view of the condition of pavements in Missouri.

The methods for restoring the missing values in the data are expected to cause some problems. It is likely that the replacement of missing values with values from previous years will not show the corresponding behavior of the deterioration of pavements, but it is a good way of avoiding inconsistencies in the results obtained from performing the data mining applications. Likewise, the replacement of missing values with the average of values obtain from previous and subsequent years will probably not show the corresponding behavior either, but again it is a way of avoiding inconsistencies in the results obtain. The results obtained from these two approaches will be compared to the results obtained from leaving the data just as it is in order to learn from the percentage of errors found as well as to point out the need for the collection of pavement data.

CONCLUSIONS

The collection of pavement data is an extensive process that is done both manually and automatically. The data collected are currently being used by state agencies to assess the decision-making process for the maintenance and rehabilitation (M&R) of our nations highways. However, the amount of data stored in the PMS databases today is immense, which generates the need for knowledge in order to use more of the data for any M&R decisions.

The KDD process is being applied to a set of data provided by MoDOT as an initial step towards the goal of expanding the use of pavement management data. The process here described is part of an ongoing study. Among the parameters used for the assessment of the pavements by means of the KDD process and data mining applications, the most relevant were the present serviceability rating, rutting, ride quality, condition, cracking, year the pavement was last worked, AADT, overlays, original surface, and surface type. However, for future work, parameters like the percentage of trucks and passenger vehicles, design life of each pavement, type of pavement (i.e. concrete/asphalt, and if concrete, reinforced/unreinforced or continuously reinforced), and environment, should also be considered in order to extract better information, thus make better decisions.

ACKNOWLEDGMENTS

The author would like to thank the Missouri Department of Transportation, which provided Missouri pavement condition data for the research, and the IBM Corporation, for providing the Intelligent Miner for Data software to be used in the research. In addition, the author thanks Dr. K. Sanford Bernhardt for her suggestions.

REFERENCES

1. National Cooperative Highway Research Program. *NCHRP Synthesis 203: Current Practice in Determining Pavement Condition: A Synthesis of Highway Practice*. Transportation Research Board, National Research Council, National Academy Press, Washington, D.C., 1994.
2. Stumme, G., R. Willie, and U. Willie. Conceptual Knowledge Discovery in Databases Using Formal Concept Analysis Methods. *Lecture Notes in Artificial Intelligence 1510*. 1998. pp. 450-458.
3. Fayyad, U., P. Shapiro, and P. Smyth. *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, Massachusetts, 1996. pp. 1-34.
4. Moxon, B. Defining Data Mining. *DBMS Online*. 1996. <http://www.dbmsmag.com/9608d53.html>. Accessed February 25, 2000.
5. Méndez, J., M. Hernández, and J. Lorenzo. A Procedure to Compute Prototypes for Data Mining in Non-structured Domains. *Proceedings Second European Symposium on Principles of Data Mining and Knowledge Discovery*, Sept. 1998.
6. Aggarwal, C.C. and P.S. Yu. Data Mining Techniques for Associations, Clustering and Classification. *Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining*, April 1999.
7. Two Crows. *Introduction to Data Mining and Knowledge Discovery*. Two Crows Corporation, Pontomac, MD, 1999.
8. American Association of State Highway and Transportation Officials. *Guidelines for Pavement Management Systems*. Washington, D.C., July, 1990.
9. Paterson, W.D.O. and T. Scullion. *Information Systems for Road Management: Draft Guidelines on System Design and Data Issues*. The World Bank, Washington, D.C., Sept. 1990.
10. National Cooperative Highway Research Program. *Synthesis of Highway Practice 76: Collection and Use of Pavement Condition Data*. Transportation Research Board, National Research Council, Washington, D.C., 1981.
11. National Cooperative Highway Research Program. *Synthesis 14: Skid Resistance*. Transportation Research Board, National Research Council, Washington, D.C., 1972.
12. Hudson, W. R., R. Haas, and W. Uddin. *Infrastructure Management*. McGraw Hill, New York, NY, 1997.
13. Fukuhara, T., K. Terada, M. Nagao, A. Kasahara, and S. Ichihashi. Automatic Pavement-Distress-Survey System. *Journal of Transportation Engineering*, ASCE, Vol. 116, No. 3, May/June 1990, pp. 280-286.
14. Boettcher, G. Automated Road Data Collection. 2000, <http://www.esri.com/library/userconf/proc95/to400/p355.html> , Accessed January 25, 2000.

15. Murphy, M. Texas Department of Transportation, telephone communication, Jan. 2000.
16. Roadware, Inc. Products. <http://www.roadware.com/products.htm>. Accessed January 25, 2000.
17. Sanford, K.L. *Improving Condition Assessment: Data Requirements For Bridge Management*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA., 1997.
18. Block, E. D. Connecticut Department of Transportation, telephone communication. Dec. 1999.
19. Kimball, R. Meta Meta Data Data. DBMS Online March 1998. <http://www.dbmsmag.com/9803d05.html>. Accessed February 25, 2000.
20. Pyle, D. *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, San Francisco, California, 1999.
21. Franks, G. Missouri Department of Transportation, personal communication. April, 2000.
22. Goebel, M. and L. Gruenwald. A Survey of Data Mining and Knowledge Discovery Software Tools. *ACM SIGKDD Explorations*, Vol. 1, No. 1, June 1999, pp. 20-33.
23. Han, J. and M. Kamber. *Data Mining: Concept and Techniques*. Morgan Kaufmann Publishers, San Francisco, California, 2001.