APPLICATION OF DATA-MINING TO STATE TRANSPORTATION AGENCIES' PROJECTS DATABASES

SUBMITTED: September 2006
REVISED: January 2007
PUBLSIHED: March 2007 at http://itcon.org/2007/8/
EDITOR: C. Anumba

Khaled Nassar, Associate Professor
Department of Architectural Engineering, University of Sharjah
email: knassar@sharjah.ac.ae

SUMMARY*: Data mining is a relatively new data analysis technique that has the ability to discover patterns stored within historical data and is now considered a catalyst for enhancing business processes by avoiding failure patterns and exploiting success patterns. This technique is widely used in business applications including market segmentation, fraud detection, and credit risk analysis as well as many other applications. In the construction domain however, the use of data mining has been extremely limited. Data mining usually requires the availability of a large database of previous cases to be analyzed. Therefore applications in the construction industry must be geared to those situations where such databases are readily available. This paper describes a research effort to explore a potential use of data mining in the construction industry. Real data about asphalt paving projects was collected from various IDOT (Illinois Department of Transportation) sources and analyzed using data mining techniques. The results indicate that data mining can provide information beyond the use of general statistical analysis. Various rules and patterns were derived from the original database, which could be applied to support decision-making. The limitations of data mining are also noted including the need to verify and test the discovered patterns.*

KEYWORDS*: data mining, construction databases, state transportation agencies databases, knowledge discover.*

1. INTRODUCTION

Organizations are increasingly storing large amounts of data generated during their operating activities. Patterns that indicate the effectiveness of the various business processes are usually buried within this historical data. A recently utilized analysis method, data-mining, has the ability to discover patterns stored within historical data and is now considered a catalyst for enhancing business process by avoiding failure patterns and exploiting success patterns. It has been estimated that the quantity of data in the world roughly doubles every year, while the amount of meaningful information decreases rapidly (Adrains 1996). Properly analyzing data and detecting these patterns is therefore of great importance to businesses. The construction industry is no exception. Construction companies collect data on a daily basis for activities and operations. Similarly, State Transportation Agencies (STAs) maintain their own project databases. Public or semi-public access is sometimes provided on the Internet, such as those of Occupational Safety and Health Administration (OSHA), Federal Highway Administration (FHWA) and various STAs. Data mining can provide a great tool for discovering the wealth of information contained in this data (Cabena 1997).

Although several researchers provide definitions for the process of knowledge discovery in databases (KDD) in general and data mining specifically, there is no one definition of these terms. The term "KDD" is generally employed to describe the whole process of extraction of knowledge from data and the term "data mining" is often used exclusively for the discovery stage of the KDD process (Adrians 1996, Hand et al 2001, Jiawei 2001). It should be acknowledged however, that data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods. Consider for example a database of highway construction projects. Traditional statistics such as cost correlations or average production rates during a certain period of time can be easily calculated. However, statistical methods alone can not automatically reveal all the possible hidden relationships within the database. A complex relationship inidicating, for example, that the job overhead increases proportionally with the number of change orders during the summer month for high productivity projects can be hidden within the projects database and can be detected using data mining techniques.

Data mining is widely applied in business applications including market segmentation, customer profiling, fraud detection, evaluation of retail promotions, credit risk analysis insurance policy, and in some military operations (Witten 2000). In the construction domain, the use of data mining techniques has been limited. Nii and Okine presented a data mining approach to pavement rehabilitation and maintenance decision support using rough set theory (Witten 2000). The authors promote the rough set concept as an effective tool for analysis of information systems in a pavement management system (PMS) database gained by both objective and subjective methods. The data used in the study was collected from the Florida Department of Transportation (FDOT) district 6 in 1995 for flexible pavement. It consisted of four measured indices and three derived indices that are related to pavement rehabilitation and maintenance. The conclusion was that the preliminary results indicate that the rough set theory application may well work for a PMS system.

Leu et al (Leu et al 2000) investigated the applicability of data mining in the prediction of tunnel support stability using an artificial neural networks (ANN) algorithm. Data from a railway tunnel construction in western Taiwan were used to establish the model. The main objective was to develop a neural network model for the prediction of tunnel support performance. In preparing the database, all information sources were identified and an essential data subset containing 1000 records was selected. After a thorough data cleaning, 470 records remained for the ANN analysis. The number of rock mechanical and construction related attributes used as input variable totaled 14. The data types were both numeric (directly read from rock mechanical logging and daily reports) and logical.

In addition, Soibelman et al. discussed the data preparation process for construction knowledge generation through knowledge discovery in databases (Soibelman 2002), as well as construction knowledge generation and dissemination (Soibelman 2000). Data mining usually requires the availability of a large database of previous cases to be analyzed. Therefore, applications in the construction industry must be geared to those situations where such databases are readily available.

This paper describes a research effort undertaken to explore the applicability of data mining to a potential application in the construction industry. Data mining techniques were applied to a state transportation agency's (STA) database containing information about asphalt paving projects such as cost and schedule data. The goal was to discover any hidden rules of patterns stored within the data. In the next section a brief introduction to data mining techniques is presented. In the following section, data collected from the state of Illinois DOT's projects-database containing typical asphalt paving projects is presented and analyzed. Data mining was used to reveal unknown patterns and trends in the database of paving projects. Examples of the extracted patterns and rules are presented. Finally, the limitations and the conclusions are presented.

2. THE DATA-MINING PROCESS

Several data mining techniques have been developed over the last decade. Generally, the data mining techniques can be categorized in four categories, depending on their functionality: classification, clustering, numeric prediction, and association rules. The main difference between the different techniques is in the way they extract information (algorithms and methods used) and how results (knowledge discovery/rules) are expressed.

Classification problems are essentially predictive models used to analyze an existing database to determine categorical divisions or patterns in the data. Classification problems are focused on identifying the characteristics indicating the group or class to which each record in the database belongs. On the other hand, when there is no pre-identified class or group, the clustering technique is used to group items that seem to fall naturally together. The third data-mining technique is numeric prediction, which is essentially a variant of the classification learning technique where the outcome is a numeric value rather than a category. The outcome to be predicted is not a discrete class but a numeric quantity.

The data mining technique used in this research is association learning. In association learning, the goal is to discover any interesting patterns in the data by discovering association rules. Association rules differ from classification rules in two ways: they can predict any attribute (not just the group or class), and they can predict more than one attribute's value at a time. A typical association rule is represented in the following way:

Cause_1, Cause_2 => Result (or consequence)
That is, if Cause_1 and Cause_2 hold then Result (the association rule) applies, for n% of cases with x% confidence.

Each rule extracted is usually provided with a confidence level and a support. The confidence is the statistical value presenting the probability of a certain rule and the support is the number of cases/projects in which the rule is found. A pattern is defined as several identical or similar rules indicating a trend. Most of the data mining techniques use statistical tests when constructing rules or patterns and also for correcting models that depend too strongly on particular records in producing the rules and patterns (Feldens 2002). Since the goal when analyzing the dataset collected here was to detect any potentially useful patterns, association learning was the data mining technique selected to analyze the dataset collected in this paper. In the next section, an actual application of data mining to an STA paving project databases is presented.

3. APPLICATION OF DATA MINING TO A STA'S PAVING PROJECT DATABASE

In this paper, the application of data mining to a database containing data on construction asphalt paving operations projects was explored. The main purpose was to explore any relationship between relevant variables that might reveal hidden knowledge about the paving projects. Ideally, the most interesting relationships to be identified are those between project cost and other variables, traffic control and traffic control cost and other variables, contractors and any cost variables, as well as any other general relationship between variables undetected. In the following sections the different steps carried out will be discussed.

3.1 Data collection

Data available on the paving project in various IDOT sources was explored. In addition, contractors (superintendents, estimators and management personnel) and various IDOT personnel were contacted and interviewed for further knowledge and support. A suitable amount of project instances was needed for the analysis. Therefore, all nine IDOT-districts were contacted and their input requested. The information was mainly collected from two main sources: the IDOT archived bid tabs and contractor's bulletins. Table 1 below lists the type of information collected from each source.

Table 1: Type of information collected

Data source type	Internet -Text files	Internet - Pdf files	IDOT provided
Contractors Bulletins	Project characteristics (partly) Contractors bid	General information Project char.	
Archived Bid Tabs	Bids (bid prices) Contractor information		
Additional information			Cost information Time of day Actual working days

Initially, the contractor's bulletins were downloaded and searched for relevant data. From the bulletins all projects under the paving sections, containing any asphalt concrete paving parts were chosen. These projects were further checked for data, and all projects that were found to have insufficient data, and therefore not fulfilling the categorical criteria (listed below), were left out. In order to ensure consistency, the construction of all the collected projects began and finished between the years 2000 and 2001. These projects had letting days from March until November in the year 2000.

3.2 Data manipulation

Following data collection a dataset was structured in MS Excel. A flat file - dataset consisting of 414 instances (individual projects) and 21 attributes was created. Each instance has a number of attributes. The attributes were divided into four main categories:

- General issues
- Project characteristics
- Traffic control issues
- Contractor's issues

The collected attributes are listed in table 2 along with the type of each variable.

Table 2: Attributes and their state

Attribute	**Type**
General issues	
Contract no.	Numerical
District	Numerical
County	Logical
Location	Logical
Project characteristics	
Type of project	Logical
Distance	Numerical
No. of lanes	Numerical
Planned working days	Numerical
Actual working days	Numerical
DBE	Numerical
Volume of asphalt concrete	Numerical
Surface mix	Logical
Superpave	Numerical/Logical
Time of day	Logical
Traffic control issues	
Traffic control	Numerical/Logical
Total traffic control cost	Numerical
Contractor issues	
Contractors no.	Numerical
Name of contractor	Logical
Contractor's bid	Numerical
Percent change in bid	Numerical
No. of unsuccessful bidders	Numerical

3.2.1 General issues

General issues include attributes such as Contract number, District, County and Location. These have no other meaning other than distinguishing between the instances. There may be certain facts buried in the data that can reveal connections between the general issues attributes and some other attributes.

3.2.2 Project characteristics issues

The “Type of project” attribute describes the type of project being constructed, which can be: Surfacing, Resurfacing, Patching, Widening, or a combination of the two. The Planned Working days, Actual Working days, Length/Distance, Number of lanes, Volume of asphalt concrete, Mixture and Superpave attributes all represents typical aspects of operations. The Volume is represented as QC/QA tons (quality control and quality assurance), which is used as a rough approximation of the total asphalt concrete for each project. The Mixture attribute is used to identify the asphalt concrete mixture used for every project, which in this study includes mixtures, C, D, E and, F (or a combination in case of multiple overlays). The Superpave attribute (Superior Performing Asphalt Pavements) indicates whether the project was a Superpave project or not. This is a factor that can affect the performance and productivity of the contractor. DBE (Disadvantage Business Enterprises) is the participation in % of total estimated contractor cost for minority or women businesses in every project. This attribute can be used to investigate whether DBE influences the contractor cost.

3.2.3 Traffic control issues

“Traffic control” which is a Boolean attribute and associated “Total cost of traffic control” are the only two attributes included in this category. The associated total cost usually consists of several pay-items, which had to

be found and added up to make the total cost for traffic control. These attributes can help in identifying the impact of traffic control costs.

3.2.4 Contractor's issues

This group consists of the Contractor (its name and bidding number), Contractor's bid (bid price), Number of unsuccessful bidders (for each bid) and Percent change from contractor's bid.

3.3 Statistical analysis

Descriptive statistical analysis was performed on the complete dataset to identify all essential information about the data. Scatter graphs were generated to identify sub-sets that may identify potential correlation between any two variables, check for outliers (see Figs. 1 and 2) and the potential need for normalizing certain attributes. The analysis did not indicate any correlation between the variables nor the need for normalizing any particular attributes. Fig. 1 also shows a frequency plot of the bid price and indicates that almost eighty percent of the bids submitted are for amounts less than 1 million dollars. Fig. 2 shows a coherence graph between tonnage and distance. Although no clear relationship exists between tonnage and distance, other kinds of information that relate the two with other attributes of the database can be detected as will be shown below. Table 3 lists all attributes used for the data-mining process as well as their descriptive statistics.

Table 3: Descriptive statistics of attributes

Attribute	Average	Stdev.	Range	
			Max	Min
Contract no.	*	*	*	*
District [1 - 9]	*	*	*	*
County	*	*	*	*
Location			0	1
Type of project	*	*	*	*
Distance [miles]	3.33	3.40	24.00	0.07
No. of lanes	*	*	*	*
Planned working days	48.0	39.8	310	15
Actual working days	37.8	33.6	191.5	2
DBE	0.06	0.04	0.16	0
Volume of asph. concrete [tons]	9936	17545	152151	0
Surface mix	*	*	*	*
Superpave			0	1
Time of day [D/N]	*	*	*	*
Traffic control			0	1
Total traffic control cost [$]	$ 19,647	$ 38,126	$ 465,250	$ -
Contractors no.	*	*	*	*
Name of contractor	*	*	*	*
Bid price [$]	$ 878,913	$ 1,387,366	$ 15,003,639	$ 33,576
Percent change from bid	4.1	10.9	74.0	-28.2
No. of unsuccessful bidders	1.7	1.4	7	0
Actual wd/Planned wd	0.61	0.42	2.69	0

Nevertheless, the statistical analysis revealed the fact that some variables in this study are not suitable for the data mining. For instance, after the cleaning process, only 2 nighttime projects out of overall 338 were left for data mining, making it impossible to create rules and draw any conclusions about nighttime paving operations. Therefore, the time of day attribute was left out when the dataset was mined.

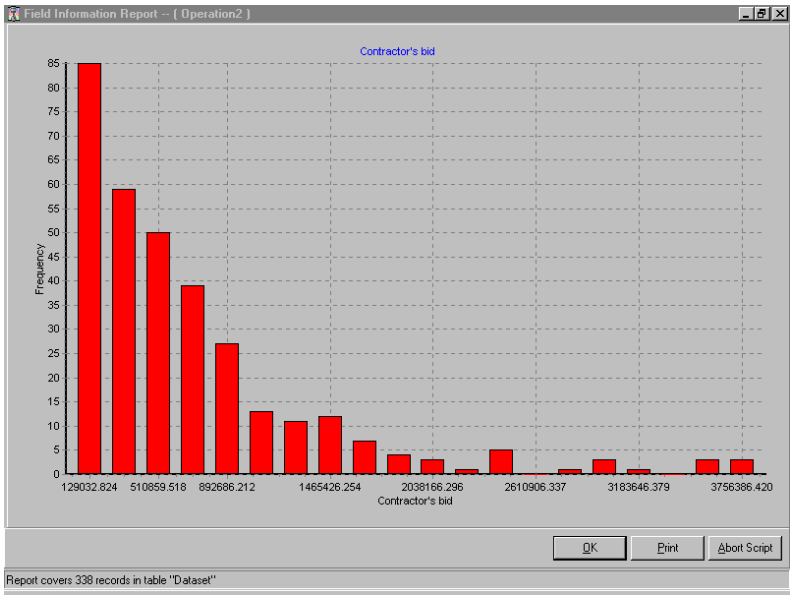

FIG. 1: Histogram of the "Bid price" attributes

3.4 Enrichment

During the analysis, it became obvious that additional attributes were needed. Location as recorded had no meaning to the problem. Some projects took place in the same area (or on the same highway) but were distributed statewide in different locations, which could result in unreliable rules being extracted from the dataset. To include the location of each instance in some form, an attribute was added to the dataset. Each project or instance was coded as either 0 or 1, roadway or highway respectively. Another attribute for the ratio of actual working days to the planned working days was also added to the dataset.

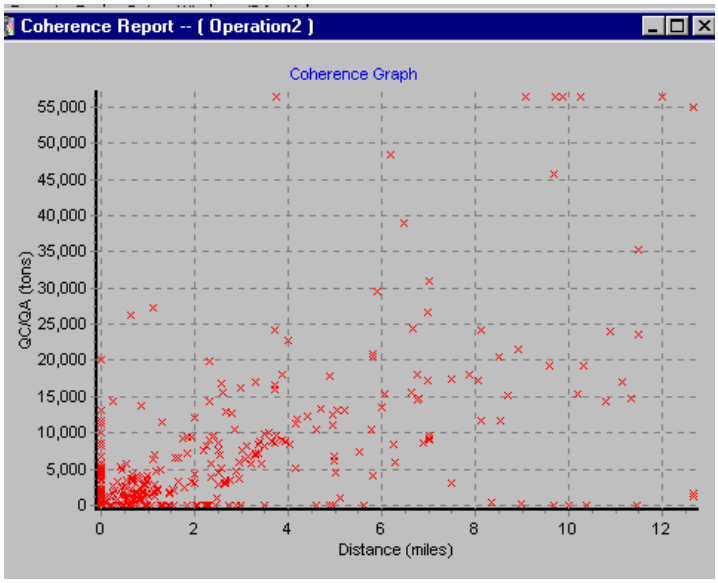

FIG. 2: A coherence graph between tonnage and Distance

3.5 Data mining

Before actual data mining was carried out, a careful consideration of the four different data-mining techniques (discussed in section two) was done to decide which would be the most suitable. Given the fact the attributes in the dataset collected here consist of both logical and numerical values, the classification technique was chosen. An appropriate software tool was selected. AIRA, an add-in to Excel, AiraExcel.xla from Hycones Information

Technology (TCC 1999) was used. It uses a combination of classification - neural networks and statistical analysis to generate rules.

The advantages of using this software include the ability to handle unlimited amounts of instances, its simplicity in manipulating and generating rules, its suitability for relatively small datasets like the one created for the study, and its ability to handle missing data. For example, data collected from one of the state districts was missing in three fields (time of day, actual working days and % change from contractor bid). However, the software's ability to disregard redundancies or noisy data made data mining possible. Therefore data mining process was then carried out using the software and the results were displayed in text file similar to the one shown in Fig. 3.

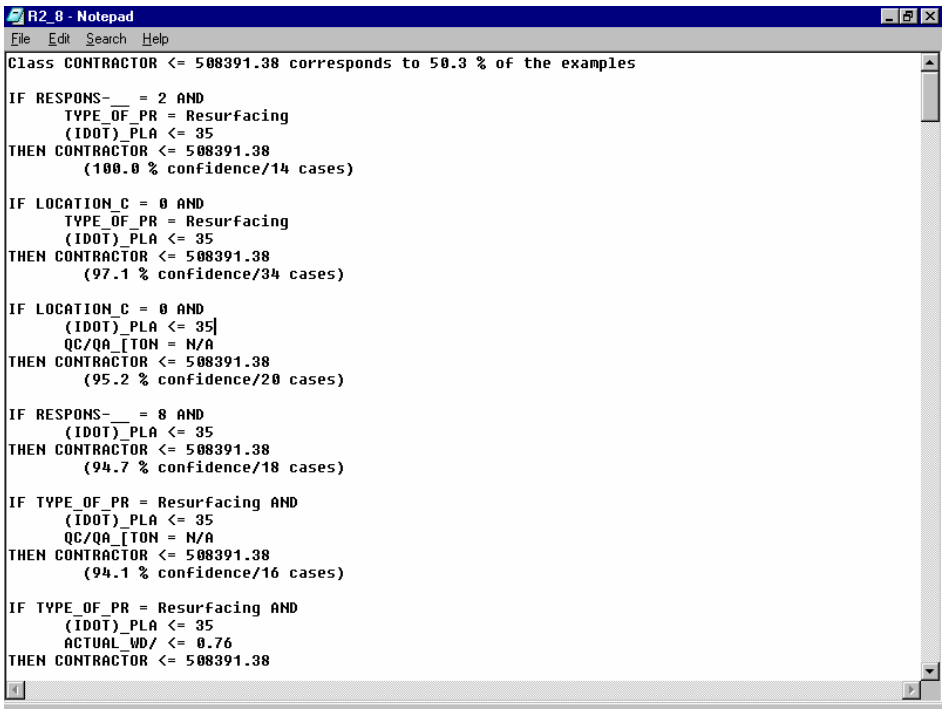

FIG. 3: Sample Rules Generated

4. RESULTS

The software output of generated rules is presented in the following format:

IF *CAUSE_1* AND
CAUSE_2 (and/or)
CAUSE_3
THEN *CONSEQUENCES* and associated *CONDITION*
(with X % confidence and support in n cases)

The first step in extracting these rules is usually to define the targets and the causes (the causes are the attributes that should be considered by AIRA, and are candidate to appear in the IF-side of the rules). It is important to have all relevant attributes available, and not select any irrelevant ones. The target attribute (spreadsheet column) has the information that will appear in the THEN-side of the rules. It is generally the information one wants to be able to predict based on data mining-discovered rules (TCC 1999). The CONSEQUENCES part of the generated rules is presented with the corresponding X % confidence interval and the number of cases in which this rules applied. One of the generated rules for example was:

IF *Location = 0* AND
Type of project = Surfacing (and)
Bid <= 508391
THEN *Total Traffic Cost* <= 9125
(92.9% confidence/13 cases)

That is for Roadways where new surface is being laid and contractor's bid price was less than or equal to $508,391 the total traffic cost was less than or equal to $9125. This can be stated with approximately 93% confidence and is supported with 13 cases. The most significant rules that were found are listed in table 4, (with a minimum confidence value of 75% identified as useable). The goal is not just to find rules similar to the one above, but instead to extract a number of similar rules that would resemble a trend, i.e. a pattern.

Table 4: Major rule findings

Target: ACTUAL/PLANNED WORKING DAYS *Class: ACT./PLAN. WORKING DAYS <= 0.75 corresponds to 46.2 % of the examples*
Rule no. 1 IF DISTRICT 7 AND WORKING DAYS <= 35 VOLUME <= 5134 TONS THEN ACT./PLAN. WORKING DAYS <= 0.75 (84.6 % confidence/11 cases)
Rule no. 2 IF HIGHWAY AND TYPE OF PROJECT = Surfacing VOLUME > 5134 TONS THEN ACT./PLAN. WORKING DAYS <= 0.75 (80.0 % confidence/16 cases)
Rule no. 3 IF ROADWAY AND VOLUME <= 5134 TONS TOTAL TRAFFIC COST > $15878 THEN ACT./PLAN. WORKING DAYS <= 0.75 (76.9 % confidence/20 cases)
Target: ACTUAL/PLANNED WORKING DAYS *Class:0.75 < ACT./PLAN. WORKING DAYS <= 1* *corresponds to 45.6 % of the examples*
Rule no. 4 IF DISTRICT 3 AND HIGHWAY TYPE OF PROJECT = Resurfacing (overlay) THEN 0.75< ACT./PLAN. WORK. DAYS <= 1 (84.6 % confidence/22 cases)
Rule no. 5 IF DISTRICT 3 AND TYPE OF PR = Resurfacing 0.75< THEN ACT./PLAN. WORK. DAYS <= 1 (82.8 % confidence/24 cases)
Rule no. 6 IF TYPE OF PROJECT = Patching/Resurfacing PLANNED WORK. DAYS <= 35 $5250 < TOTAL TRAF. COST <= $15878 THEN 0.75< ACT./PLAN. WORK. DAYS <= 1 (91.7 % confidence/11 cases)
Rule no. 7 IF TYPE OF PROJ. = Patching/Resurfacing PLANNED WORK. DAYS <= 35 $282969 < BID PRICE <= $761317 THEN 0.75 < ACTUAL_WD/ <= 1 (81.3 % confidence/13 cases)
Target: BID Price *Class: $282969 < BID price <= $761317 corresponds to 33.4 % of the examples*

Rule no. 8 IF PLANNED WORK. DAYS <= 35 VOLUME > 5134 TONS TOTAL TRAFFIC COST <= $5250 THEN $282969 < BID price <= $761317 (75.0 % confidence/15 cases)
Rule no. 9 IF PLANNED WORK. DAYS <= 35 VOLUME <= 5134 TONS TOTAL TRAFFIC COST <= $5250 THEN BID price <= $282969 (84.9 % confidence/62 cases)

Certain patterns were identified from the hundreds of rules extracted. For example,

- There is a general trend for new surface type of projects for highways to be more often complete within scheduled time. Any combination of rules in this regard (and there are several) has usually a high confidence level and is largely supported.
- In district 3, projects tend to be within scheduled time when planned working days are less than 35. This is in particular the trend when the project is of the resurfacing type (overlay). Moreover, the bid price tends to be less than $283,000 if working days are less than 35 working days.
- If projects in district 2 are scheduled longer than 35 days, it can be implied with more than 50% confidence that projects overrun the time-schedule. Also if the volume of asphalt concrete is more than 5130 (QC/QA) tons the bid tends to be more than $760,000.
- For resurfacing types of projects in general, when projects are scheduled for less than 35 working days, tend to be either low in volume (tons of asphalt concrete) or the bid price is lower than $ 283,000.
- For district 7, if volume of asphalt concrete is less than 5130 tons and either traffic control cost lower than $5250 then bids tend to be lower than $283,000.
- Moreover, the bid price tends to be less than $283,000 if working days are less than 35 working days. Also if the volume of asphalt concrete is more than 5130 (QC/QA) tons the bid tends to be more than $760,000.

After these patterns were generated, IDOT was contacted to verify some of these patterns. IDOT personnel could confirm some of the patterns and did provide explanation for them. For example, one of the reasons suggested for the extracted rules about exceeding schedules in some rural IDOT districts relates to generous mobilization times needed in those districts, which has to be a minimum of 15 days.

5. DISCUSSION

It is important to note a couple of key issues relating to the implementation of data mining in the construction industry. Firstly, there is an obvious lack of standardization in the construction industry as it relates to collecting and storing project and company specific data. This industry fragmentation significantly hinders the uptake of data mining techniques in practice. Furthermore, the way in which the information is stored in the construction industry is generally not very organized and patchy. In the research presented here, for example, various pieces of information had to be collected from bulletins, reports, as well as electronic databases and then re-structured into one database in order to facilitate data mining.

In order to overcome this problem, the construction industry can learn from the state-of-the-art in data mining in other industry sectors. There is a need for a unified data model for construction data. This unified data model would be similar to the current building product model utilized in the Industry Foundation Classes (IFC) but would focus primarily on construction specific data. We envision this data model to be organized in three layers; one layer would capture project-specific data such as cost, estimate, schedule and productivity. The second would capture company-specific data, such as profitability, bids and bonding capacity. These two layers would be obviously interrelated so that information can be indexed from one layer to another. The third layer would capture industry-specific data such as employment rates, industry wide productivity rates and financial ratios.

This unified data model would greatly facilitate the implementation of data mining techniques in the construction industry.

Another issue to note is the importance of having procedures in place to decide on which aspects of the construction data are suitable for mining, and to develop rational basis for data mining. This procedure must include a system for evaluating the results by potential end-users such as project managers and upper level executives. In the research presented here, informal discussions were carried out with IDOT personnel to decide on key aspects of the available data that can be mined. The results were then reviewed and some of the new rules discovered confirmed existing perceptions, such as the trend for new highway surfacing projects to be completed within scheduled time. Other pieces of information were suspected but the data mining procedure placed accurate probability values to them. Several new rules were completely new and provided new insight into the data, such as the relation between bid price and working days.

6. CONCLUSIONS AND RECOMMENDATIONS

The study presented here describes an application of data-mining analysis to a typical construction database containing information about asphalt projects in Illinois. A case study was presented to test the applicability of data mining as an analysis method. A database was constructed with collected data from IDOT sources. Data-mining technique was utilized to analyze the created dataset and rules generated. Based on the generated results and interpretation, certain previously unknown patterns were discovered. The study shows that data mining can provide information on a dataset/database beyond statistical methods only and provide a source of valuable information (that could not have been detected otherwise) to support decision-making. If the time-consuming data collection process can be reduced, the method can extract information faster than other analysis methods.

Suggestions for future research include increasing the size of the dataset used, as well as trying other software and techniques to verify the extracted rules and trends. One of the main characteristics of data mining is the large amount of data needed to generate rules. The major rules generated in this study usually have high confidence but only limited amount of cases supporting the rule. Therefore, increasing the data set will significantly enhance the quality and reliability of the generated rules and trends.

Another important extension to this research is exploring the validity of the previously unknown patterns that were discovered. This could entail mining the data using other software as well as conducting a long-term study to check to verify those rules. Furthermore, other applications of data mining to the construction industry could be developed. One of the applications suggested by IDOT personnel is collusion detection among contractors, similar to the fraud detection application of data mining currently used by many insurance companies.

This study has shown that data mining can provide information on a dataset or databases beyond the use of general statistical analysis only and potentially provide a source of valuable information that could not have been detected otherwise. Moreover, the results imply that the rules derived from the original dataset could be applied to support decision-making.

7. REFERENCES

Adrians P., Zantinge D. (1996). Data mining. Addison-Wesley Longman, England.

Cabena P. (1997). Discovering data mining: From concept to implementation. Prentice Hall, NJ.

Han J. (2001). Data mining: Concepts and techniques. Morgan Kaufmann Publishers, San Francisco.

Hand, D.J, Mannila, H., Smyth, P. (2001). Principles of data mining. MIT press, Massachusetts

Miguel F. (2002). URL: http://www.softlookup.com/

Nii O. Attoh-Okine, (1997). Rough set application to data-mining principles in pavement management database, J. Comput. Civ. Eng., Am. Soc. Civ. Eng. 11 (4) 231-237.

Soibelman L., Hyunjoo K. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. Journal of Computing in Civil Engineering, ASCE, 16 (1), 39-47.

Soibelman L. (2000), Construction knowledge generation and dissemination. Berkeley-Stanford CE&M workshop: Defining a research agenda for AEC process/product development in 2000 and beyond.

Leu S., Chee N., Shiu-Lin C. (2000). Data mining for tunnel support: neural network approach. Journal of automation in construction, Volume 10, Number 4, pp. 429-441(13).

TCC, Two Crows Corporation (1999). Introduction to data mining and knowledge discovery. Third edition. Two Crows Corporation.

Witten, I. H., Frank, E. (2001). Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufman, California.