# Prediction of Maintenance Cost for Road Construction Equipment: A Case Study

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Canadian Journal of Civil Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>cjce-2014-0500.R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>04-Dec-2015</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Bayzid, Sharif; University of Alberta, Department of Civil and Environmental Engineering
Mohamed, Yasser; University of Alberta, Dept. of Civil and Environmental Engineering
Al-Hussein, Maria; University of Alberta, Department of Civil and Environmental Engineering |
| Keyword: | Construction Equipment, Equipment Management, Equipment Maintenance, Data Mining, Case Study |
Prediction of Maintenance Cost for Road Construction Equipment: A Case Study

Sharif Mohammad Bayzid\textsuperscript{1}, Dr. Yasser Mohamed\textsuperscript{2*} and Maria Al-Hussein\textsuperscript{3}

\textsuperscript{1} MSc Candidate, Department of Civil and Environmental Engineering, 1-047 Markin/CNRL Natural Resources Engineering Facility, University of Alberta, Edmonton, AB, T6G 2W2, Email: bayzid@ualberta.ca

\textsuperscript{2} Associate Professor, Department of Civil and Environmental Engineering, 3-011 Markin/CNRL Natural Resources Engineering Facility, University of Alberta, Edmonton, AB, T6G 2W2, Email: yaly@ualberta.ca, Ph: 780-492-1321

*Corresponding author

\textsuperscript{3} System Analyst/Project Manager, Department of Civil and Environmental Engineering, 5-050 Markin/CNRL Natural Resources Engineering Facility, University of Alberta, Edmonton, AB, T6G 2W2, Email: maria.al-hussein@ualberta.ca
Abstract

Equipment maintenance cost is significant in construction operations budgets. This study proposes a systematic approach to predict maintenance cost of road construction equipment. First, maintenance cost data over more than 10 years was collected from a partner company’s equipment management information system. Data was cleaned, and analyzed to obtain a general understanding of maintenance costs trends. Next, traditional cumulative cost models and alternative data mining models were generated to predict maintenance cost based on available equipment and operation attributes. Data mining models were evaluated and validated using portions of the collected data that have not been used in model development. Data collection, analyses, modeling, and validation steps are discussed. The paper also presents the performance of different model types. Based on the case study data, regression model trees performed better than other model types with equipment work hours being the most significant parameter for predicting maintenance cost.

Keywords: Construction Equipment, Equipment Management, Equipment Maintenance, Data Mining, Case Study
Introduction

Equipment management is a crucial topic for equipment-intensive construction domains like earthmoving and road construction. Equipment management focuses on financial, operational and mechanical aspects of equipment. The main responsibility of an equipment manager is to support construction by providing the required equipment on time and at an affordable cost. To make equipment available for different construction projects at a reasonable cost, the financial and mechanical aspects of equipment are very important. The equipment manager has to ensure that all the required mechanical work, such as repair, maintenance, and replacement of equipment is being done at the right time(s) (Vorster 2009). Also, the manager has to confirm that the equipment is owned and operated at the optimum cost. The operational cost includes fuel costs, tire costs, operator’s wages and maintenance costs. Maintenance cost is a big portion of total operating cost that affects budgeting for upcoming years. Maintenance costs differ for different types of equipment, manufacturer, and operating conditions, which makes predicting such costs a challenging task that depends on the company practice and inventory of equipment. This research demonstrates a case study that targets such analysis for a major road construction company.

This study uses data from a major road construction contractor who has been serving Edmonton for more than 40 years. The company does road construction within the city of Edmonton and surrounding areas using different types of road construction equipment. For tracking and maintaining equipment, the company utilizes an equipment management system (M-Track), which is custom developed to company’s needs. The system tracks repair and maintenance work and supports inventory management, labour management and equipment parts purchasing management. However, the system does not provide decision support, such as trend
analysis of equipment maintenance cost, budget of maintenance work, comparison of different Types of equipment and replacement analysis. In the current practice of the company, budgeting for maintenance costs is performed on the basis of the last one or two years’ equipment inspection information and the company’s current need for maintenance work. This budgeting is being conducted without prediction analysis and does not follow any analytical method.

The objective of this study is to find suitable models for predicting maintenance cost of road construction equipment. This paper first presents a review of literature on maintenance cost forecasting, cumulative cost modeling and data mining analysis. Then it describes a summary of the study methodology starting with a background about the information management and data collection process used in the case study. Then, the main phases of the study are further explained under three main sections: Data preparation and pre-processing, Data analysis and modeling, and Model evaluation and validation. Finally a discussion of the results, limitations, future work and main conclusions are presented.

**Literature review**

**Maintenance Cost Forecasting**

Forecasting of construction equipment maintenance cost has been the subject of a number of research studies and can be done in a qualitative or quantitative approach based on past experience or information from a database. The quantitative method is used to predict maintenance and repair costs when past data is available because it is more accurate than basing the decision on the judgment of related experts or people (Kim 1989). Many algorithms have been proposed in previous research for accurate forecasting by quantitative analysis. Quantitative methods that could be used for equipment management include naïve, moving average,
exponential smoothing, time series analysis, and regression. Among these methods, regression is the most common method, which is usually accurate over a medium-range prediction horizon (Makridakis et al. 1989; Manatakis and Drakatos, 1993; Gillespie, 2004). Furthermore, Yip, Fan, and Chiang (2014) also presented a general regression neural network (GRNN) to predict maintenance cost. Although the quantitative forecasting method is more accurate than the qualitative method, as it is based on an actual database, there are some disadvantages. The quantitative method’s main shortcomings are that it depends on historical data for forecasting and that its long-range forecasting is questionable (Mitchell 1998). Despite these drawbacks, the quantitative analysis is mostly used to forecast maintenance costs, and it has also been used in this research.

Most equipment-owning companies use different constant repair cost or modified constant repair cost methods for forecasting equipment maintenance cost. Cox (1971) proposed estimating equipment repair costs as a percentage of purchase prices. Cox included multiplication factors for type of service (easy, medium or severe). The Caterpillar company recommended a method similar to the one proposed by Cox (Caterpillar 1995). For machines whose lifespan would be more than 10,000 hours, Caterpillar added an additional factor, but this factor is applicable for the machine’s entire lifespan. These methods are used to simplify the forecasting analysis, but oversimplification could cause a significant difference between actual and forecasted values.

According to Nunnally (2000), repair cost is the highest single item of operating cost for most construction equipment. Repair cost is dependent on use of equipment, operating conditions and maintenance standards. Nunnally proposed equation (1) for hourly repair cost:
Hourly repair cost = \( \frac{\text{Year digit}}{\text{Sum of year's digit}} \times \frac{\text{Lifetime repair cost}}{\text{Hours operated}} \)  …………………………… (1)

Where lifetime repair cost = Lifetime repair cost factor X (initial purchase cost – tires cost)

Vorster (2009) proposed a method to prepare database and regression analysis for forecasting maintenance costs. Maintenance costs usually tend to increase with the age of the equipment, so to devising an economical budget for the upcoming year, proper prediction of maintenance cost is crucial. From various literature reviews and also from Vorster’s (2009) point of view, it was found that repair or maintenance costs of equipment can be represented by a second order polynomial as shown in Equation (2). This kind of trend line was observed in some cases; these are described in the general trend analysis section.

\[ MCL_D = A \times W_D + B \times W_D^2 \]  ………………………………………… (2)

Where:

\[ MCL_D = \text{Life-to-date maintenance cost at age } W_D \]

\[ W_D = \text{Life-to-date hours worked by the machine} \]

\[ A = \text{Coefficient that describes the linear increase of cost with age} \]

\[ B = \text{Coefficient that describes the exponential increase of cost with age} \]

**Economic approaches for CCM**

Vorster (1980) proposed a cumulative cost model (CCM), which provides numerical and graphical solutions to many equipment management problems. Graphical solutions help the decision-manager to better understand the problem. As for maintenance cost study, age can take
form of calendar age, age in cumulative hours of use (equipment age), or age in units of production. Practically, fleet managers often use equipment age instead of the other two because it reflects the real usage of a machine. Figure 1-a shows a CCM model where the abscissa of the CCM graph is the equipment age and the ordinate of the CCM graph is the cumulative cost. By drawing tangents, the optimization point can be determined. Optimum economic life, $L^*$, is the tangent to the cumulative cost curve drawn from the origin (Mitchell, Hildreth & Vorster 2011). After that age, the operating cost for equipment usually becomes higher. There are several managerial decisions that can be supported by CCM, including equipment purchase decisions, preventive maintenance strategies, replacement of equipment decisions, and forecasting of repair costs (Mitchell 1998).

**Life-to-Date cost analysis**

The life-to-date (LTD) repair cost is one of the approaches for CCM analysis that is easy to understand, as well as to use, in any managerial decision-making situation. If the data is available, the machines should be analyzed at different ages to spread the data points uniformly throughout the lifetime of the type of the equipment. To avoid the unwanted influence of a machine over other machines, the same number of data points need to be used for each machine (Mitchell 2011). For example, if most of the machines in an equipment class have cumulative cost data up to a 6,000 hour meter reading, and 12 data points are needed, then the data points of each machine should be collected at intervals of every 500 hour meter reading. The graph should be plotted with the cumulative maintenance cost as the ordinate and hour meter reading (age) as the abscissa. A second order nonlinear curve has to be plotted through the origin from the data points (Figure 1-b). Through this graphical analysis, the coefficients A and B for Equation (3) can be determined.
\[ CMC_T = A*H_T + B*H_T^2 \] \hspace{1cm} (3)

\[ CMC_T = \text{Cumulative maintenance cost at any age } H_T \]

\[ H_T = \text{Age of equipment at time } T \]

After plotting the graph of CMC vs. age, the total maintenance cost and average maintenance cost per hour at any time can be extracted. For calculating the average maintenance cost per hour at any age, \( H_T \), a straight line has to be plotted from \( H_0 \) to \( CMC_T \). The slope of the straight line is the average cost of maintenance costs. Though LTD method provides the most accurate picture of maintenance cost over the age of the equipment, this method cannot be used when equipment data is not available from the start of the machine run.

**Period-Cost-Based analysis**

Another approach to CCM is the period-cost-based (PCB) analysis, where it is necessary to have cost data for labour and parts for the same type of machines for any particular period of time. The time period is the number of hours worked by the machine from any starting point, \( H_A \), to the end point, \( H_B \) (Figure 1-c). For a second order polynomial equation, \( H_M \) is the mean of \( H_B \) and \( H_A \). The slope, \( m \), is the marginal maintenance cost at equipment age \( H_m \). The equation of the curve in Figure 1-c is the same as described in Equation 3. The differential equation at \( H_M \) is given in Equation 4:

\[ m = A + 2B*H_M \] \hspace{1cm} (4)

Which is similar to the linear regression equation:

\[ Y = C + D*x \] \hspace{1cm} (5)
From Equations 4 and 5 it can be found that:

\[ C = A \]  \hspace{2cm} (6)  
\[ D = 2B \]  \hspace{2cm} (7)

It is easy to obtain the values of \( A \) and \( B \) from Equations 6 and 7 after obtaining the values of \( C \) and \( D \) from the PCB analysis. Then, the equation of the cumulative maintenance cost, with respect to the equipment age, can be determined from Equation 3.

**Data mining**

In many research studies, data mining analysis was included to explore new knowledge from stored data. Soibelman and Kim (2002) used data mining as a tool in their research to identify valuable and applicable patterns to help construction managers analyze huge amounts of construction management data. Fan (2007) used data mining technology for automated knowledge generation and decision support analysis utilizing large amounts of equipment operational data. He proposed non-parametric outlier mining algorithms for this decision support analysis. Gonzalez-Villalobos (2011) used data mining to analyze similarities between projects during the bidding process. Liao and Perng (2008) found a pattern of occupational injuries in the construction industry from the historical database by using data mining analysis. Their research will help to develop an efficient inspection policy and injury prevention plan. Fan, AbouRizk, Kim and Zaïane (2008) also presented a data mining approach for assessing the residual value of heavy construction equipment, and its potential benefits on the decision making of equipment management. Hammad (2009) proposed that data mining analysis to enhance the efficiency of labour estimating practices. His research proposed a data mining approach which was expected to provide companies with knowledge-based dynamic estimating units that always reveal the
most up-to-date changes. Kumar (2013) predicted scaffold man-hours with respect to different related attributes by creating a linear regression equation using data mining algorithms.

The algorithms built by data mining analysis can be classified by their results, which are clustering, association and classification (Gonzalez-Villalobos 2011). This research work only concentrated on classification. In this case, classification requires a labeled dataset which needs to be split into training and testing datasets. Then a prediction model is built on the basis of the training data set and is evaluated by the testing dataset (Hammad 2009). If the test results are not up to a satisfactory level then the model needs to be changed. The models are usually trained with the help of different algorithms and then evaluated by testing the data set to discover which algorithm’s model could give the best prediction output. The range of algorithms used in data mining is very wide and beyond the scope of this study. However, of particular interest to this study is treed regression family of algorithms (Alexander and Grimshaw 1996). In these algorithms, the data space is divided into smaller partitions by building a decision tree structure. A linear regression model is then created at the leaf ends of that tree. (Chipman et al. 2002).

Different approaches are used to generate the tree structure and induce the linear model at the leafs. Among these approaches are m5 method (Holmes, et. al. 1999, and Quinlan 1992), and REP method (Elomaa and Kaariainen 2001). In these methods, the set of training examples is split progressively. A subset of these examples is either associated with a leaf that contains a model or is split further according to a chosen test. The type of test used, the method used for pruning and smoothing the developed tree characterize the different variations of these algorithms.

In this research, eight data mining algorithms were utilized for numerical prediction: linear regression, second order nonlinear regression, least median square, conjunctive rule,
decision stump, M5Rule, REP tree and multilayer perceptron. After building the models using these algorithms, the models were evaluated to decide which models can be implemented in the field.

**Study methodology**

**Case study background and data collection**

The development of information technology and database systems encourages large contractors to invest and implement these technologies in a data collection and distribution system (Fan 2007). The partner company in this case study took initiative to implement equipment management software called M-Track, which has been in use since 1997. The system is a client-server application with a database on the Microsoft SQL server. The system’s main features are purchasing and inventory management; shop labour time management; and equipment maintenance services such as running repairs, preventive maintenance, and planned maintenance. The company performs three types of maintenance work on its equipment:

- **Running repairs**: repair work that has to be done due to the breakdown of equipment.
- **Planned maintenance**: equipment is inspected annually. On the basis of the severity of the problem and the equipment budget, maintenance work is ordered and performed. If these mechanical deficiencies cannot be repaired, the equipment may break down.
- **Preventative maintenance**: regular maintenance work that is suggested by the manufacturer to keep equipment in the best possible working condition (Nunnally 2000).

Maintenance costs consist of labour and parts costs. The M-Track system supports the time entry of labour work where the labour hours for both running repair and planned maintenance can be found. Parts costs due to running repairs and planned maintenance can be
found in the purchase order and on the inventory tables. It is easy to obtain the labour and parts costs for preventive maintenance, as both are available in a separate table in a good, structured way. The structure to get the total maintenance cost of a piece of equipment from M-Track is shown in Figure 2.

**Main phases of the study**

The methodology followed in this case study can be summarized in three main phases as follows: 1) Data preparation and pre-processing, in which data is extracted from company’s information system, transformed and cleaned to remove any obvious anomalies or incomplete records, and loaded into a format suitable to run the analysis tools. 2) Data analysis and modeling, where basic trend analysis was performed to visualize the actual trend of maintenance cost, cumulative cost modeling was conducted using two economic approaches, and lastly, data mining analysis was done using different algorithms. 3) Model evaluation and validation phase, where the performance of the models generated were compared for each of the equipment classes and then the best algorithm was selected. A summary of the study methodology is shown in Figure 3. The figure is divided into three parts. These parts are discussed in the data preparation and pre-processing, data analysis and modeling, and validations sections in the same order.

**Data preparation and pre-processing**

**Data warehousing**

One of the purposes of data warehousing is the integration of data that is distributed in different systems in the company. For this case study, datasets for maintenance cost of equipment were imported from Microsoft SQL Server into Microsoft Access. From the collected datasets, queries were executed to get different components of the equipment maintenance cost.
The total maintenance cost comes from three different labour costs (running repair (RR), planned maintenance (PL) and preventive maintenance (PM)), and four different parts costs (purchase order (PO), internal transfer (IT), cash purchase (CP) and preventive maintenance (PM)). Purchase order, internal transfer and cash purchases are done for both running repair and planned maintenance. A sample dataset of the different components of maintenance cost is shown in Table 1. The total number of records imported at this stage was over 8000 records. After importing the required data, some of the tables related to maintenance of equipment were exported to Microsoft Excel for further data pre-processing and cleaning as discussed in the following section.

**Data pre-processing**

In the M-Track database, running repair and planned maintenance cost data were stored with respect to timestamps. Just for preventative maintenance, both hour meter reading and timestamp were stored for each of the readings, but for this study, odometer or hour meter reading should always be present with respect to different maintenance cost data. So, by matching up the preventive maintenance’s timestamp to that of the other maintenance, a common dataset for all of the maintenance cost data, with respect to hour meter reading, was prepared.

Inconsistency and missing data is common in most databases due to errors in data entries. This irregularity has to be found and resolved. In this case study, some similar outliers were encountered, such as change of hour meter reading when the hour meter was replaced, mistakes during entry of hour meter readings etc., which is illustrated in Table 2. In the third row of Table 2, problems due to replacement of hour meter are shown. This was resolved by adding the new hour meter readings with the last reading of the previous hour meter. In the sixth row of Table 2, outliers due to misleading hour meter reading are shown. This was solved by taking the average.
of the previous hour meter reading and the next hour meter reading. After solving the problems related to maintenance data, all components of maintenance cost are added to get total maintenance cost. Then cumulative maintenance cost and maintenance cost/hour is calculated as per the following equations:

Cumulative maintenance cost (CMC) at \( H_T = TMC_1 + TMC_2 + TMC_3 + \ldots + TMC_T \ldots \) (8)

Where \( H_T = \) Hour meter reading at any time \( T \)

\( TMC_T = \) Total maintenance cost at any time \( T \)

Maintenance cost/Hour = \( \frac{(CMC_T - CMC_1)}{(H_T - H_1)} \) ................................. (9)

Where \( CMC_1 \) and \( H_1 \) are first available cumulative maintenance cost and hour meter reading for any piece of equipment.

Data analysis and modeling

General trend analysis

From the prepared database, two types of analyses were conducted: general trend analysis and prediction analysis. Prediction analysis was done by cumulative cost modeling and data mining analysis. The full set of maintenance data available in the company database was utilized in the analysis. Sampling was only performed during the evaluation and validation phase to split data into training and test sets as explained in later sections. After data cleaning and pre-processing, the average number of records used in the analysis for each equipment class was 97 with a minimum of 38 and maximum of 246 records. The overall analysis process is presented in part 2 of Figure 3. Different portions of this part are explained in the following sections.
After data preparation, the first step was to explore the trend of maintenance cost for different equipment classes. The analysis was conducted on the basis of equipment class because it was assumed that within the same class, the behavior of the equipment is almost same. Also, according to Mitchell et al. (2011), variability of repair cost with respect to hour meter reading can be explained more accurately by grouping the equipment as an equipment class. In the statistical analysis, using a larger dataset creates less influence of a particular machine that is performing unexpectedly well or poorly. According to the needs of the company, the research work was done on all available equipment classes between class number 200 and 299 (please refer to Appendix 1 for equipment class descriptions). Fifteen equipment classes were found within this interval with sufficient data for trend analysis.

Cumulative cost was initially used in this study for the trend analysis. Graphs of cumulative maintenance cost vs. hour meter reading showed that the fleet of equipment consists of old and new equipment where maintenance cost data for many of the equipment units was not available from zero hour meter reading. A sample of such trend is shown in Figure 4.

Cost-per-hour trends were also analysed. Figure 5 and Figure 6 illustrates cost per hour trends for two of the available equipment classes. For each of the figures, graph (a) shows the trends of different units within the same equipment class and graph (b) shows the trend of all the units together as an equipment class. The difference between old and new units in the case of cumulative cost is resolved when using cost-per-hour, but the trends of maintenance cost per hour were not always as they are described in the literature. Usually, the trend of maintenance cost increases when the equipment becomes older, but in some equipment classes, the trend is different. Figure 5 presents a normal upward trend for maintenance cost per hour for class 262. However, Figure 6 shows an unusual trend compared to what is usually reported in the literature.
This kind of abnormal trend was found for 4 (equipment class 217, 219, 243 and 253) out of 15 equipment classes. To find out the reasons behind these unfamiliar trends, meetings with an expert in the company were conducted. By analyzing the data and from company point of view, it was found that for most of the cases, high preventive or planned maintenance at different stages of the equipment age could be the main reason behind these downward trends. Following data preparation and trend analysis, alternative modeling approaches were used and evaluated as explained in the following sections.

**Cumulative cost modeling**

Life-to-date (LTD) and part-cost-based (PCB) regression analysis are two approaches of cumulative cost modeling (CCM). For these two approaches, a satisfactory number of equipment units in an equipment class is required, but most of the equipment classes have 5 to 15 equipment units and very few equipment classes contain more than 2 units with enough data instances to carry LTD and PCB analyses. When all the available equipment classes were explored for LTD and PCB analysis, only two equipment classes (class 240 and 262) were found suitable for these analyses. For LTD analysis of equipment class 240 (Graders-150 to 225 hp), the hour meter reading was divided by 1000 to obtain A and B values in Equation 3. Cumulative maintenance cost values were collected for every 1000 hour meter readings. For equipment class 240, maintenance cost data were collected from 1000 hours to 8000 hours (every 1000 hour interval) for all the equipment units so that data could be evenly distributed throughout the range of ages. The plotted dataset for equipment class 240 and the corresponding regression model are shown in Figure 7-a. The corresponding statistical analysis is shown in Table 3. The results of this analysis show coefficient values for A of 1021.7 and B of 4342.1. It should be noted here that an intercept value of 0 was forced in this regression as the maintenance cost at the start of
equipment life should be 0. The statistical analysis of the regression model also shows a good fit of the model with an adjusted $R^2$ value of 0.93. It also shows very low P values for $X$ and $X^2$ which indicates their statistical significance as predictors in this model at a 0.05 alpha level. Also the overall significance of the model (significance F) is well below the 0.05 level. Finally the residual plot (Figure 7-b) does not show a consistent pattern, which indicates the model captures the deterministic behavior in the data. Overall, we can conclude the model is a good fit for this dataset. In PCB regression analysis, an average maintenance cost value has to be calculated. For every piece of equipment of class 240, this average cost was calculated for the period of 2009-2010, along with the midpoint of the hour meter reading for that period, which is shown in Table 4. Then, these values were plotted with the linear regression analysis that is presented in Figure 8. As the dataset for this PCB analysis is relatively small and scattered, the model has low value of $R^2$ (0.536). The values of $C$ and $D$ in Equation 5 were found from Figure 8 to calculate the value of $A$ and $B$ in Equation 3. The values of $C$ and $D$ were found to be 2611 and 2797.2. Then, from Equation 7 and 8, the values of $A$ and $B$ were calculated. These are 2611 and 1398.6. After obtaining the values of $A$ and $B$, PCB and LTD equations are plotted as shown in Figure 9. Although the two methods used different equipment units’ data, the cumulative cost equations for the two methods are quite similar; this is visible in Figure 9. From these two equations, cumulative maintenance costs can be predicted. As an example, to predict cumulative maintenance cost for 16000 hours, LTD and PCB equations provide the values $331028.80$ and $399817.60$, which are not too far from one another. For predicting maintenance cost by regression analysis these two approaches of CCM could be very useful, especially the PCB method, because for much of the equipment, the data from the point of starting the service is not
available. However, this CCM modeling is not practically usable for most of the equipment classes in the database because:

- Within most of the equipment classes very few units have sufficient instances of data to carry out LTD and PCB analyses.
- Only a few equipment classes have more than 2-3 equipment units that have maintenance cost data from zero hour meter reading.
- It was very hard to find any particular one or two years’ maintenance cost data for sufficient equipment units of one equipment class. Only two equipment classes within all of the concerned classes have a sufficient amount of equipment data for a particular one or two years, which is required for PCB analysis.

**Data mining analysis**

During general trend analysis, it was found that only 15 equipment classes between class number 200 and 299 have adequate realistic datasets for data mining analysis. The following four attributes were considered as candidate variables for predicting maintenance cost per hour:

1) Manufacturer of the equipment (manufacturer)
2) Working year of equipment (year)
3) Hour meter reading of equipment (hour meter)
4) Purchase price of the equipment (purchase price)

It should be noted that all attributes are numerical except the manufacturer attribute, which is nominal. In data mining analysis, nominal attributes are commonly handled using binary comparisons. For example, if equipment manufacturer attribute has a certain value, a binary variable in the model is evaluated to 1, otherwise, the binary variable evaluates to 0. In this data mining analysis, first the most frequently used second order nonlinear regression
analysis was conducted using Microsoft Excel. After that, seven other algorithms were analyzed using Waikato Environment for Knowledge Analysis (WEKA) software (Hall et al. 2009).

The prediction analysis of equipment maintenance cost using a second order nonlinear regression was performed for all equipment classes. Here, as an example of training set, the partial data of equipment class 222 (Wheel Loaders, 4 cy) is presented in Table 5. Also last year’s data, used as the testing set, is shown in Table 5. Training data were used to build the second order polynomial equation (Figure 10-a). To maintain the same basis for comparison with other data mining methods and the preserve as many data points as possible, the data in this phase of the analysis was used in its raw format and was not averaged over work hours intervals as is the case in CCM analysis. The effect of this can be seen in the figure in the form of sudden peaks of maintenance costs usually due to major overhauls or repairs for individual units. This also impacts the smoothness of the trend and accuracy of prediction. From the generated equation, the maintenance cost was predicted for the last year, which is presented in the last column of the testing set (Table 5). There is also actual maintenance cost-per-hour data for each of the instances in Table 5. The graphical comparison of actual and predicted maintenance cost data is presented in the right graph of Figure 10-b. As can be seen in the figure, although the predicted values seem close to the actual values, the \( R^2 \) value of the regression model indicates that hour meter reading does not fully explain the variations in the data.

Alternative data mining algorithms were also used through WEKA to build different models for a given equipment class dataset and make comparisons to determine the best performing model. Initial exploration involved the examination of seven classification algorithms to build the best model for each of the 15 equipment classes. Correlation coefficient and different error measures of each model are generated for each model by comparing test data
with predictions from the built model. The following sections summarize validation method used
to evaluate these models and the outcomes of this evaluation.

Model evaluation and validation

Validation method

One of the steps in building a prediction model using data mining techniques is to decide
on data splitting for training and testing. Two common ways are percentage split and cross
validation. Percentage split involves dividing the dataset into two parts, one part for training and
other part for testing. In cross validation, the data is divided into k-folds where one fold is used
for testing and k-1 folds are utilized for training. The total process is repeated k times and is
usually considered more efficient for validating models (Duan et al. 2003). However, the
objective of this study is to predict an upcoming year’s maintenance cost, which makes cross-
validation unsuitable in this case as it does not maintain the order of data points for forecasting
purposes. On the other hand, with ordered percentage split, we can test the model using last
years’ data and train the model using the previous years’ data. In this analysis, for each class of
equipment, the last one or two years of data (depending on amounts of data available) were used
for testing the model and the rest of the data were used for model training.

Performance of different models

The second order nonlinear regression analysis in Microsoft Excel and seven other
numerical prediction algorithms in WEKA were validated based on five different measures:
mean absolute error, root mean squared error, relative absolute error, root relative squared error
and correlation coefficient (Hall et al. 2009). For any equipment class, the last one or two years’
data were used as a testing set and the rest were used as a training set. The main purpose of the
evaluation was to find out which of the eight algorithms performs the best for an equipment class, and whether there is a particular algorithm that is capable of modeling all or most of the available classes more than others.

Table 6 shows the best case as well as the worst case of the model evaluation and validation step. The best case is shown in Table 6, where for equipment class 222 (Wheel Loader, 4 cy) almost all the algorithm’s error values are very low and the correlation coefficient is almost 1. For equipment class 240 almost all types of error are very high and correlation coefficient is 0 or close to 0. From the comparison of errors in Table 6, it can be seen that M5Rule (Hall et al. 2009) performed better than others in most cases. This algorithm generates a set of rules with a regression model to use under each rule. A sample M5Rule model is shown in Figure 11 for class 222 case. Figure 12 shows the number of equipment classes in which an algorithm performed the best and the number of equipment classes in which the algorithm can still be used even though it did not have best performance. The total count in the figure is the summation of the number of equipment class cases where the algorithm resulted in best or acceptable performance. As shown in Figure 12, the M5Rule performed the best in most equipment classes and the second best algorithm was either the least median square or the multilayer perceptron. As the least median square and M5Rule are both regression-based algorithms, they can be used for most of the equipment classes.

Though M5Rule and Least Median Square can be used for most of the equipment classes, Linear Regression and REP Tree (another regression model tree generator) cannot be ignored, as there are certain equipment classes where root relative squared error and relative absolute error are more than 100% for M5Rule and Least Median Square. For equipment classes 220 and 262, Linear Regression is the best method and for class 253, REP Tree is the best method.
Discussion of case study and analysis results

In this case study, a number of data sets for maintenance costs of different road construction equipment classes were analyzed. The goal of the study is to find the models for predicting maintenance costs for equipment classes for the purpose of annual budgeting and equipment replacement decision making. Traditional PCB and LTD models were applied to some classes. However, the data available were not suitable for carrying the same analysis for all equipment classes. Alternative data mining algorithms were tried and evaluated. The comparison between different algorithms for modeling equipment maintenance cost showed different levels of performance. Some algorithms, namely treed regression, performed better than others for most of equipment classes tested in this study. The advantage of the treed regression models found in this study has also been reported in previous studies for different (mainly auction based residual value) equipment datasets (Fan et. al. 2008). This advantage indicates that these models are viable candidates for modeling equipment maintenance costs. However, datasets for some equipment classes resulted in poor models under all the algorithms tried in this study. While it is not possible to reliably assess the reason of this lower model performance without further research, we can theorize that it could be attributed to higher level of variability in data records in their raw format. This may be due to different equipment use and maintenance patterns that govern this class of equipment a. For example, it was found during data pre-processing that some equipment units are transferred between different branches of the company, which results in a new identifier of the unit and consequently different cost pattern that is difficult to recognize in an automated way. In addition, while some equipment use is limited to construction season, the use of other equipment like graders may expand to winter season affecting their maintenance
cost pattern. Further pre-processing and transformation of data in such cases (e.g. smoothing by averaging expenses over work hour intervals) may be a viable solution, which was not tried in this study.

**Study limitations and future research**

This study is based on data from one company. Hence the suitability of the models developed for use by other companies is not tested. Future work may target the generalization of these models for use by any company. The validation of the developed models in this study is based on comparing predictions of last year in the dataset to actual values present in the dataset that have not been used in model training. Predictions beyond the one year window are not tested and if a company needs to use similar models on an annual basis, these models will need to be updated to reflect all current records in the company’s database, which is a process that can be automated easily. The models developed in this study are also equipment class specific. Future work may target the generalization of such models to span multiple classes of equipment. Such generalization will have to take into account the different use patterns and work conditions associated with different equipment classes.

Based on the challenges faced in data extraction and cleaning in this study, it is also suggested that validation of hour meter readings is enforced during data entry in the information management system (M-Track, or similar ones) to avoid discrepancies and inconsistencies in meter readings over time. Finally, use pattern (i.e. what job types and conditions equipment hours were used on) may be a valuable addition to data collection and may explain the differences in maintenance cost trends between different equipment classes and units.
Summary and Conclusions

This study focused on determining a systematic approach to predict equipment maintenance cost for upcoming years for a road construction contractor based on its equipment management data. The data collection and preparation stage consumed a big portion of time in this study. After preparation of dataset trend analysis was conducted to explore the behavior of maintenance cost trend when the equipment becomes older. For prediction analysis CCM and data mining, methods were chosen. Though two approaches of CCM were explored for two of the available equipment classes, it was found that due to lack of sufficient amount of data instances in almost all the equipment classes, CCM is not feasible for this database. In the data mining step, first the common second order nonlinear regression analysis was conducted and compared with other data mining algorithms. Seven algorithms were examined for data mining analysis. After evaluating and validating different algorithms it was found that regression trees based algorithms, particularly model tree builders (e.g. M5Rule and REP Tree) performed better than others for most of the available equipment class datasets. Based on the analysis in this study we conclude that CCM models should be the first choice for their simplicity and ease of development, but in case available data is not suitable for CCM analysis, treed regression models are viable alternatives. This finding aligns with previous studies on different equipment datasets and suggests that the family of treed regression algorithms is a viable candidate for modeling equipment maintenance costs for a wide range of cases.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada through the Collaborative Research and Development Grant CRDPJ 416362 – 11.
References


Table 1: Sample data of different components of equipment maintenance cost

<table>
<thead>
<tr>
<th>Equipment Unit No.</th>
<th>Make</th>
<th>Time Stamps</th>
<th>Hour Meter Reading</th>
<th>Labour Cost ($)</th>
<th>Parts Cost ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Labour Cost ($)</td>
<td>Parts Cost ($)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RR</td>
<td>P</td>
</tr>
<tr>
<td>217-401</td>
<td>Dynapac</td>
<td>9/17/2001</td>
<td>2091</td>
<td>1,076.2</td>
<td>5</td>
</tr>
<tr>
<td>217-401</td>
<td>Dynapac</td>
<td>4/22/2002</td>
<td>2224</td>
<td>4,908.7</td>
<td>5</td>
</tr>
<tr>
<td>217-401</td>
<td>Dynapac</td>
<td>9/16/2002</td>
<td>2406</td>
<td>1,897.5</td>
<td>0</td>
</tr>
<tr>
<td>217-401</td>
<td>Dynapac</td>
<td>5/9/2003</td>
<td>2487</td>
<td>918.75</td>
<td>-</td>
</tr>
<tr>
<td>217-401</td>
<td>Dynapac</td>
<td>10/14/2003</td>
<td>2671</td>
<td>1,653.7</td>
<td>5</td>
</tr>
<tr>
<td>217-401</td>
<td>Dynapac</td>
<td>5/18/2004</td>
<td>2698</td>
<td>2,362.5</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 2: Illustration for hour meter reading corrections

<table>
<thead>
<tr>
<th>Equipment Unit No</th>
<th>Make</th>
<th>Event ID</th>
<th>Time Stamp</th>
<th>Hour Meter Reading (From the Database)</th>
<th>Hour Meter Reading (Corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>205-404</td>
<td>Dynapac</td>
<td>117148</td>
<td>20/03/2009</td>
<td>780</td>
<td>780</td>
</tr>
<tr>
<td>205-404</td>
<td>Dynapac</td>
<td>119452</td>
<td>03/02/2010</td>
<td>820.5</td>
<td>820.5</td>
</tr>
<tr>
<td><strong>205-404</strong></td>
<td>Dynapac</td>
<td><strong>121987</strong></td>
<td>16/02/2011</td>
<td>3</td>
<td><strong>823.5</strong></td>
</tr>
<tr>
<td>205-404</td>
<td>Dynapac</td>
<td>123530</td>
<td>12/01/2012</td>
<td>162</td>
<td>982.5</td>
</tr>
<tr>
<td>230-405</td>
<td>CAT</td>
<td>108686</td>
<td>10/04/2006</td>
<td>10174.5</td>
<td>10174.5</td>
</tr>
<tr>
<td><strong>230-405</strong></td>
<td>CAT</td>
<td><strong>109934</strong></td>
<td><strong>16/10/2006</strong></td>
<td><strong>382.5</strong></td>
<td><strong>10569.75</strong></td>
</tr>
<tr>
<td>230-405</td>
<td>CAT</td>
<td>111093</td>
<td>09/04/2007</td>
<td>10965</td>
<td>10965</td>
</tr>
</tbody>
</table>
Table 3: Statistical analysis for LDT regression model for Class 240

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>59579563406</td>
<td>29789781703</td>
<td>450.6327789</td>
<td>9.79454E-17</td>
</tr>
<tr>
<td>Residual</td>
<td>10</td>
<td>1322131150</td>
<td>66106557.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>60901694555</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0</td>
<td>#N/A</td>
<td>#N/A</td>
<td>#N/A</td>
<td>#N/A</td>
<td>#N/A</td>
</tr>
<tr>
<td>X^2</td>
<td>1021.67825</td>
<td>216.3555707</td>
<td>4.722218367</td>
<td>0.000130521</td>
<td>570.3684377</td>
<td>1472.988062</td>
</tr>
<tr>
<td>X</td>
<td>4342.090653</td>
<td>1375.806481</td>
<td>3.156032999</td>
<td>0.004970133</td>
<td>1472.208625</td>
<td>7211.942682</td>
</tr>
</tbody>
</table>
Table 4: Sample data for PCB analysis (equipment class 240, graders (150 to 225 hp))

<table>
<thead>
<tr>
<th>Equipment Unit No</th>
<th>2009-2010 Maintenance Cost ($)</th>
<th>Start 2009 Meter Reading</th>
<th>End 2010 Meter Reading</th>
<th>Step</th>
<th>Midpoint (1000 hr)</th>
<th>2009-2010 Maintenance Cost ($/1000 hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>240-404</td>
<td>44,961.00</td>
<td>5986</td>
<td>7845</td>
<td>1859</td>
<td>6.9155</td>
<td>24,185.58</td>
</tr>
<tr>
<td>240-405</td>
<td>89,481.00</td>
<td>10936</td>
<td>13102</td>
<td>2166</td>
<td>12.019</td>
<td>41,311.63</td>
</tr>
<tr>
<td>240-406</td>
<td>50,323.50</td>
<td>5150</td>
<td>7481</td>
<td>2331</td>
<td>6.3155</td>
<td>21,588.80</td>
</tr>
<tr>
<td>240-407</td>
<td>41,670.00</td>
<td>3134</td>
<td>5236</td>
<td>2102</td>
<td>4.185</td>
<td>19,823.97</td>
</tr>
<tr>
<td>240-481</td>
<td>54,099.00</td>
<td>11880</td>
<td>12625</td>
<td>745</td>
<td>12.2525</td>
<td>72,616.10</td>
</tr>
<tr>
<td>240-709</td>
<td>61,864.50</td>
<td>3924</td>
<td>5336</td>
<td>1412</td>
<td>4.63</td>
<td>43,813.38</td>
</tr>
</tbody>
</table>
Table 5: Sample of training and testing data sets for equipment class 222 (wheel loaders, 4cy)

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Year</th>
<th>Hour Meter Reading</th>
<th>Purchase Price ($</th>
<th>Maintenance Cost ($/hr)</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komatsu</td>
<td>2001</td>
<td>11093</td>
<td>127,500.00</td>
<td>9.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komatsu</td>
<td>2002</td>
<td>12081</td>
<td>127,500.00</td>
<td>12.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komatsu</td>
<td>2003</td>
<td>12694</td>
<td>127,500.00</td>
<td>47.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>2010</td>
<td>1507</td>
<td>203,087.43</td>
<td>8.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>2010</td>
<td>1713</td>
<td>203,087.43</td>
<td>13.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>2010</td>
<td>2168</td>
<td>203,087.43</td>
<td>12.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komatsu</td>
<td>2012</td>
<td>22374</td>
<td>127,500.00</td>
<td>37.82</td>
<td>42.39</td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>2012</td>
<td>2744</td>
<td>304,631.14</td>
<td>9.55</td>
<td>13.09</td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>2012</td>
<td>3644</td>
<td>304,631.14</td>
<td>17.55</td>
<td>15.45</td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>2012</td>
<td>3897</td>
<td>304,631.14</td>
<td>16.71</td>
<td>16.09</td>
<td></td>
</tr>
</tbody>
</table>
Table 6: Comparison of algorithms performance for equipment class 222 and 240

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Algorithm</th>
<th>Mean Absolute Error</th>
<th>Root Mean Squared Error</th>
<th>Relative Absolute Error</th>
<th>Root Relative Squared Error</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset for class 222 (wheel loader, 4cy)</td>
<td>Nonlinear Regression</td>
<td>1.81</td>
<td>2.1</td>
<td>19.96</td>
<td>21.58</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Least Median square</td>
<td>2.42</td>
<td>2.67</td>
<td>26.8</td>
<td>28.1</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Linear Regression</td>
<td>2.94</td>
<td>3.3</td>
<td>32.6</td>
<td>34.6</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Conjunctive Rule</td>
<td>2.55</td>
<td>2.56</td>
<td>28.3</td>
<td>26.9</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Decision Stump</td>
<td>2.5</td>
<td>2.56</td>
<td>27.6</td>
<td>26.9</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>M5Rule</td>
<td>1.5</td>
<td>1.71</td>
<td>16.6</td>
<td>17.9</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>REP Tree</td>
<td>2.22</td>
<td>2.41</td>
<td>24.6</td>
<td>25.3</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Multilayer Perceptron</td>
<td>2.88</td>
<td>3.37</td>
<td>31.89</td>
<td>35.32</td>
<td>0.97</td>
</tr>
<tr>
<td>Dataset for class 240 (grader, 150 to 225 hp)</td>
<td>Nonlinear Regression</td>
<td>3.4</td>
<td>4.4</td>
<td>148.4</td>
<td>150.45</td>
<td>-0.45</td>
</tr>
<tr>
<td></td>
<td>Least Median square</td>
<td>3.31</td>
<td>4.52</td>
<td>143.58</td>
<td>154.6</td>
<td>-0.53</td>
</tr>
<tr>
<td></td>
<td>Linear Regression</td>
<td>3.85</td>
<td>4.58</td>
<td>166.96</td>
<td>156.5</td>
<td>-0.39</td>
</tr>
<tr>
<td></td>
<td>Conjunctive Rule</td>
<td>3.01</td>
<td>3.54</td>
<td>130.4</td>
<td>121.3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Decision Stump</td>
<td>2.84</td>
<td>3.38</td>
<td>123.4</td>
<td>115.6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>M5Rule</td>
<td>2.17</td>
<td>3.06</td>
<td>94.3</td>
<td>104.5</td>
<td>0.0324</td>
</tr>
<tr>
<td></td>
<td>REP Tree</td>
<td>2.85</td>
<td>3.38</td>
<td>123.38</td>
<td>115.6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Multilayer Perceptron</td>
<td>7.01</td>
<td>9.08</td>
<td>212.36</td>
<td>215.47</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Figure Captions

Figure 1: Cumulative cost model

Figure 2: Structure of total maintenance cost of equipment in M-Track

Figure 3: Summary of study methodology

Figure 4: Cumulative cost trends

Figure 5: Cost per hour trend for equipment class 262

Figure 6: Cost per hour trend for equipment class 217

Figure 7: LTD analysis for equipment class 240, graders (150 to 225 hp)

Figure 8: PCB analysis for equipment class 240 graders (150 to 225 hp)

Figure 9: Comparison of LTD and PCB analysis for equipment class 240 graders (150 to 225 hp)

Figure 10: Second order regression for equipment class 222 (wheel loaders, 4cy)

Figure 11: Sample M5Rule model for equipment class 222

Figure 12: Performance of different data mining algorithms for 15 classes of equipment
Cost of Maintaining a Piece of Equipment

Labor Cost

- Time Cards
- Preventive Maintenance
- Running Repairs
- Planned Maintenance

Parts Cost

- Preventive Maintenance
- Purchase Order
- Internal Transfer
- Cash Purchase
- Preventive Maintenance
(1) Data Preparation and Pre-processing

- Multiple Data Sets Collection MS SQL Server (M-Track)
- Preparation of Tables for Different Components of Total Maintenance Cost
- Pre-processing of Database

(2) Data Analysis and Modeling

- General Trend Analysis
  - Cumulative Cost
  - Cost Per Hour
- Cumulative Cost Modeling (CCM)
  - Life-To-Date (LTD) Cost Analysis
  - Period Cost Based (PCB) Analysis
- Data Mining Analysis with Different Classifiers
  - M5Rule
  - REP Tree
  - Least Mean Square
  - Linear Regression
  - Conjunctive Rule
  - Decision Stamp
  - Multilayer Perceptron
  - Second Order Regression

(3) Model evaluation and validation

- Training and Testing Scheme
  - Ordered percentage split
  - Testing: by last one or two year data
  - Training: by rest of the data
- Performance Measures
  - Mean absolute error
  - Root mean squared error
  - Relative absolute error
  - Root relative squared error
  - Correlation coefficient
- Compare Performance and Identify Best Models for Each Equipment Class
240, Graders (150 to 225 hp)

262, Cement spreader and concrete paver
262, Cement spreader and concrete paver

(a)

262-Cement spreader and concrete paver

(b)
217- Vibratory rollers (doubles) drum of 80+

(a)

217-Vibratory roller(doubles) drum of 80+

(b)
240, Graders (150 to 225 hp), LTD Analysis

Cumulative Maintenance Cost in $

\begin{align*}
y &= 1021.7x^2 + 4342.1x
\end{align*}

Hour Meter Reading (1000 hrs)

Residual Plot

(a)

(b)
240, Grader (150 to 225hp), PCB Analysis

\[ y = 2797.2x + 2611 \]

\[ R^2 = 0.5361 \]
Draft

\[ y = -4E-08x^2 + 0.002x + 3.5438 \]

\[ R^2 = 0.4162 \]

(a) Class 222-Wheel Loaders, 4 cy

(b) Class 222-Wheel Loaders, 4 cy
Rule: 1 IF 
   hour_Corrected <= 12663.5
THEN
   Maintenance Cost/Hour =
       -5.5036 * Make=Komatsu + 0.0009 * hour_Corrected + 6.6982

Rule: 2 IF 
   hour_Corrected > 12779
   Year > 2006
THEN
   Maintenance Cost/Hour =
       0.5434 * Year - 0.0003 * hour_Corrected - 1061.3203

Rule: 3 IF 
   hour_Corrected <= 12779
   Year > 2002
THEN
   Maintenance Cost/Hour =
       0.9067 * Year - 0.0027 * hour_Corrected - 1762.1323

Rule: 4 Maintenance Cost/Hour =
       5.3496 * Year - 0.0032 * hour_Corrected - 10653.8151
The bar chart illustrates the number of data sets for various machine learning models. The models include M5Rule, REP Tree, Linear Regression, Least Median Square, Multilayer Perceptron, Decision Stump, Conjunctive Rule, and Second Order Nonlinear Regression.

- **M5Rule** shows the highest number of data sets, with a significant portion in the best performance category.
- **REP Tree** has a moderate number of data sets, with a notable fraction in acceptable performance.
- **Linear Regression** displays fewer data sets, primarily in the best performance category.
- **Least Median Square** and **Multilayer Perceptron** show similar patterns, with a balance between best and acceptable performance.
- **Decision Stump** has a moderate number of data sets, distributed across performance categories.
- **Conjunctive Rule** and **Second Order Nonlinear Regression** have comparable data sets, with a focus on acceptable performance.

The chart indicates that M5Rule and REP Tree have the highest variability in performance, while other models show more consistent performance across categories.
Appendix 1

Appendix Table 1: List of Utilized Equipment Classes in Analysis

<table>
<thead>
<tr>
<th>No</th>
<th>Equipment Class Number</th>
<th>Name of Equipment Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Class 202</td>
<td>Tire Compactor, 100+ hp</td>
</tr>
<tr>
<td>2</td>
<td>Class 205</td>
<td>Tandem Roller, less than 5 ton</td>
</tr>
<tr>
<td>3</td>
<td>Class 213</td>
<td>Vibratory Compactor, 50+ hp</td>
</tr>
<tr>
<td>4</td>
<td>Class 216</td>
<td>Vibratory Rollers (doubles), 7000+ kg</td>
</tr>
<tr>
<td>5</td>
<td>Class 217</td>
<td>Vibratory Rollers (doubles) drum of 80+</td>
</tr>
<tr>
<td>6</td>
<td>Class 219</td>
<td>Wheel Loaders, 0 to 1 cy</td>
</tr>
<tr>
<td>7</td>
<td>Class 220</td>
<td>Wheel Loaders, 1 to 2 cy</td>
</tr>
<tr>
<td>8</td>
<td>Class 222</td>
<td>Wheel Loaders, 4 cy</td>
</tr>
<tr>
<td>9</td>
<td>Class 223</td>
<td>Wheel Loaders, 4 1/2 to 5 cy</td>
</tr>
<tr>
<td>10</td>
<td>Class 240</td>
<td>Graders (150 to 225 hp)</td>
</tr>
<tr>
<td>11</td>
<td>Class 243</td>
<td>Elevating Scrapers over 9 cy</td>
</tr>
<tr>
<td>12</td>
<td>Class 253</td>
<td>Wheel tractors (backhoe)</td>
</tr>
<tr>
<td>13</td>
<td>Class 256</td>
<td>Asphalt Pavers, track 70 to 175 hp</td>
</tr>
<tr>
<td>14</td>
<td>Class 262</td>
<td>Cement spreader and concrete paver</td>
</tr>
<tr>
<td>15</td>
<td>Class 265</td>
<td>Soil reclaimer/stabilizer</td>
</tr>
</tbody>
</table>