University of Alberta

Modeling Maintenance Cost for Road Construction Equipment

By

Sharif Mohammad Bayzid

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Construction Engineering and Management

Department of Civil and Environmental Engineering

Spring 2014
Edmonton, Alberta

To my beloved wife, Nusrat Jahan Dipa, and mother, Sayada Khadiza Begum

(The two women I love most)

Abstract

There are many road construction companies in Canada that possess numerous fleets of heavy equipment. A big portion of the operating cost of these fleets is consumed by maintenance operations. This thesis focused on the cost of maintenance operation for a road construction company. The objective of this research is to propose a systematic approach to predict the maintenance cost of road construction equipment. This research initially intended to collect and pre-process the maintenance work database. Then, trend analysis was conducted in order to obtain a better understanding of maintenance costs. Moreover, this research intended to find the probable correlations between maintenance cost and other attributes of maintenance work. To obtain a better model for each of the available equipment classes, data mining analysis was used to compare different algorithms. These trend analyses and models will help the equipment manager to take decisions related to equipment maintenance cost.

Acknowledgements

During last two years of my graduate study I went through different good and bad situations. I met with many people who inspired me in many ways. I want to give sincere thanks to some of them who were beside me during the hard period of my life and especially to those who helped me a lot in my research work as well as course work to complete my M.Sc. in Construction Engineering and Management.

First, I would like to express my gratitude to Allah: without His blessings it is not possible to achieve anything in life. Then I want to give a special thanks to my supervisor, Dr. Yasser Mohamed. If someone asks me to say just one person who was always beside me, that would be my supervisor. I was not such an efficient and knowledgeable student compared with his other students. But he did not lose his faith in me. If I fell in the middle of some problem in my research work he helped me a lot to overcome whatever problem I was having. He always listens to his students and wants to give constructive suggestions. I have never seen such a cool, calm and intellectual professor in my life. Lastly I want to say "I am thankful to you for supervising me and I feel really honoured to be part of your research team."

I would like to thank Maria Al-Hussein, who helped me a lot to make me familiar with the software M-track. She gave me always constructive suggestions about my research and also she was a very helpful co-author for my first publication at CSCE Conference 2013.

I would like to take this opportunity to thank my parents and sisters who always keep faith in me and feel proud of me. I want to thank my beloved wife, Nusrat Jahan Dipa, who helped me a lot to proofread my thesis.

When I came to the University of Alberta it was really hard for me to adapt to its curriculum. Some of my senior and junior friends helped me at that time. I would like to thank Aniruddha Saha, Ahsanul Karim Sohag, Pejman Alanjari, Chandan Kumar and Mostafa Ali, who helped me a lot in my studies and also helped me to adapt to the new environment.

Lastly I want to mention all my senior and junior friends here in Edmonton who have made me happy on many occasions.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CCM	Cumulative Cost Modeling
CP	Cash Purchase
IT	Internal Transfer
LTD	Life to Date
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
M-Track	Maintenance Track
NSERC	National Sciences and Engineering Research Council of Canada
PCB	Period Cost Based
PL	Planned Maintenance
PM	Preventive Maintenance
PO	Purchase Order
RAE	Relative Absolute Error
RMSE	Root Mean Squared Error
RR	Running Repair
RRSE	Root Relative Squared Error
SDR	Standard Deviation Reduction
TPM	Total Productive Maintenance
WEKA	Waikato Environment for Knowledge Analysis

Chapter 1: Introduction

1.1 Background

In the arena of road construction, equipment management has flourished as a crucial topic. Equipment management mainly consists of financial, operational and mechanical aspects of equipment (Vorster, 2009). The main responsibility of an equipment manager is to support the construction work by providing the required equipment on time and within an affordable budget. To make the equipment available for different construction projects at a reasonable price, 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 and maintenance, and replacement of equipment, is being done according to an established schedule (Vorster, 2009). Also the manager has to confirm that the equipment is owned and operated at the optimum cost. The operational cost consists of fuel costs, tire costs, operator's wages and maintenance costs. As the maintenance cost is a big portion of the total operating cost, budgeting maintenance costs for upcoming years has become very significant. These equipment maintenance costs differ for different types of equipment. Also even if the equipment is the same, the costs can differ depending on the manufacturer. In summary, budgeting or predicting the maintenance cost for upcoming years is complicated, which is the main concern of this research.

The partner in this research, Standard General Inc., is a major road construction contractor that has been serving Edmonton for more than 40 years. The company does road construction for the city of Edmonton and surrounding areas using

different types of road construction equipment. For tracking and maintaining equipment, the company uses an equipment management system (M-Track) developed in collaboration with the National Sciences and Engineering Research Council of Canada (NSERC)/Alberta Construction Research Chair. The system tracks repair and maintenance work as well as supports inventory management, shop labour timesheets and equipment parts purchasing management. The system doesn't provide any decision support such as trend analysis of equipment maintenance costs, budget of maintenance work, comparison of different types of equipment and replacement analysis. For proper equipment maintenance management, it is essential for an equipment manager to budget for maintenance work. In the current practice of the company, the maintenance cost budget is being performed on the basis of the last one or two year's equipment inspection information and the current need of the maintenance work of company. This budgeting is being conducted without any prediction analysis and does not follow any analytical method.

As equipment maintenance is a labour-oriented job and a big portion of the operational budget is consumed by maintenance activity, the company has become concerned with improving its maintenance work system. In the preliminary phase, different strategies for improving the equipment maintenance system for Standard General Inc. were reviewed. Many meetings and discussions were conducted. From these meetings it was determined that equipment maintenance cost could be the area of improvement on which this research should focus.

1.3 Research Objectives

In the maintenance cost database, there are different attributes and these attributes could correlate with the equipment maintenance cost. There are different ways and algorithms to correlate these attributes with equipment maintenance cost. For predicting equipment maintenance cost it is important to find out the systematic approach to correlate the attributes with maintenance cost. The objective of this research is to propose a methodology for predicting maintenance cost of road construction equipment. This objective will be achieved by accomplishing the following sub-objectives:

1. Analyze the trend of maintenance cost for all available equipment classes to assess the behaviour of the cost.
2. Finding general trends of maintenance cost for different types of equipment from the datasets.
3. Apply alternative algorithms or methods for prediction analysis of maintenance cost for each equipment class.
4. Evaluate the performance of alternative algorithms and recommend best performers which are better suited to capture available data.
5. Assess the commonality between different algorithms and whether certain ones can be used for all equipment classes.

1.4 Research Methodology

In order to accomplish the proposed research objective and sub-objectives, the following approach is followed:

- The maintenance operation of road construction equipment is reviewed from general perspective as well as from the perspective of Standard General Inc. The internal procedure of different types of maintenance work is analyzed to understand the pros and cons of the current procedure.
- The database system for the maintenance operation of Standard General Inc. in the MS SQL Server is reviewed. Also, it is necessary to be familiar with M-Track software, which operates the database.
- Literature on different maintenance systems of construction equipment, maintenance cost forecasting methods and concepts of different algorithms that could be used for this research work are studied and reviewed.
- Equipment maintenance data of road construction equipment is imported and then processed according to this research work's requirements.
- A trend analysis of the maintenance cost for each available class of equipment (between equipment class number 200 and 299) is done to visualize the behavior of maintenance cost.
- Two approaches (Life-to-date and Part-cost-based regression analysis) of cumulative cost modeling (CCM) for each available class of equipment (between equipment class 200 and 299) are undertaken.
- A data mining analysis is conducted by running different algorithms through Waikato Environment for Knowledge Analysis (WEKA) software. The main purpose of this analysis is to identify the better algorithms for each available equipment class.

- All the models are evaluated and validated to determine which algorithm is best for each available equipment class.

The total data collection and analysis process which is conducted in this research is summarized in Figure 1. The content of Figure 1 is divided into three parts. Numbers (1) and (2) are elaborated in sections 3.4 and 3.5 of Chapter 3, and number (3) is discussed in section 4.2 of Chapter 4.

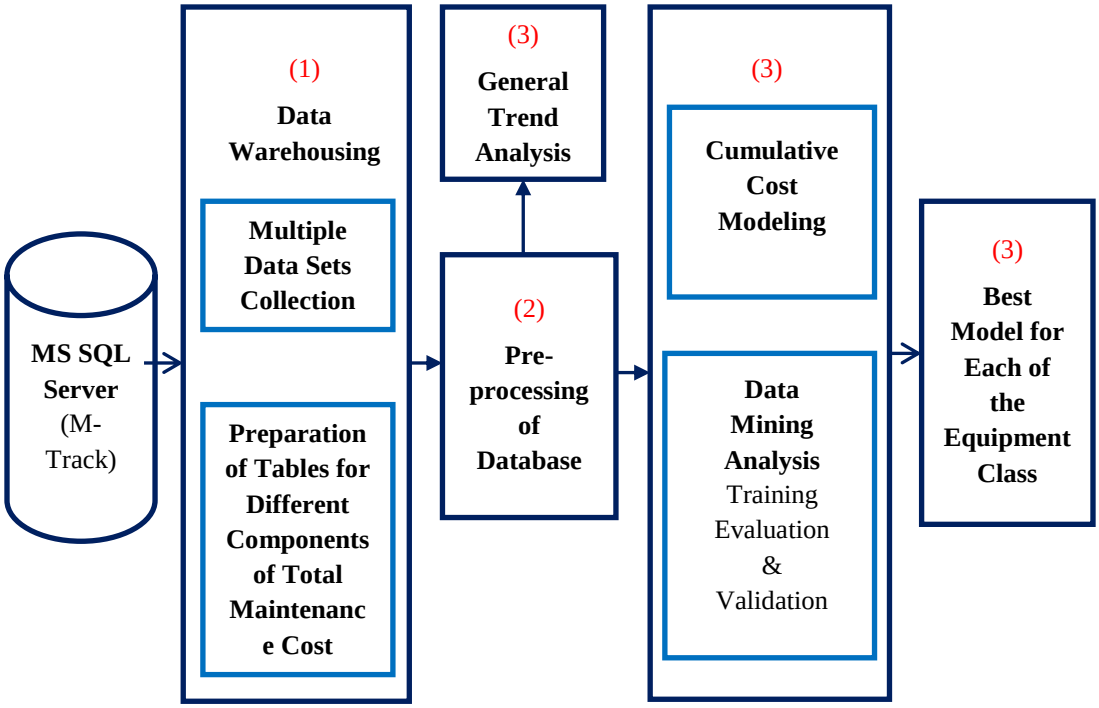


Figure 1: Summary of research work

1.5 Thesis Organization

This thesis contains five chapters. Chapter 1 provides the background, objectives and methodology of this research.

Chapter 2 contains the literature review on different types of maintenance work for construction equipment, forecasting maintenance costs, cumulative cost modeling, data mining analysis and the algorithms that could be used in this

research work. Also different methods of model evaluation and validation are reviewed in this chapter.

Chapter 3 consists of data collection and the pre-processing part. This chapter first describes the data management system in the M-Track software and the equipment maintenance operation of Standard General Inc. Then the data warehousing system for this research is illustrated. Lastly the pre-processing of dataset is described with illustrations.

Chapter 4 presents the analysis part of this research. Here, the first summary of the analysis part is presented and then it is elaborated in different sections. The trend analysis part is presented with graphs of maintenance cost vs. hour meter reading for different equipment classes. Then the cumulative cost modeling part is presented with the description of pros and cons of this process from this particular dataset's point of view. After that, the data mining part is illustrated. Lastly, all the models are evaluated and validated to determine which model or algorithm/algorithms is/are best for this maintenance dataset.

Chapter 5 presents the summary, contributions and limitations of this research. It also provides recommendations for future research.

Chapter 2: Literature Review

2.1 Introduction

Each and every company has its own way of doing maintenance work for equipment. As maintenance cost depends on the total process of maintenance activities, each company wants to figure out its own approach about predicting or forecasting maintenance cost for budgeting purposes. Different scholars have offered different theories and concepts not only about general forecasting but also about maintenance cost forecasting. This chapter includes descriptions of different researchers' points of view about maintenance activity and costs, forecasting, different algorithms used for maintenance cost forecasting, and model evaluation and validation.

2.2 Maintenance Activity and Maintenance Costs for Heavy Equipment

In the early period many companies used to do run-to-failure repairs. This is known as reactive maintenance. This approach restores equipment to working order in the least possible time. These companies that rely on reactive maintenance usually keep reserve machines, large spare parts inventories, and use worker reassignments to deal with breakdowns (Sheu, 1994). But breakdown, in many cases, causes delays in completing a project, which causes reductions in hourly production. Many factors influence the hourly production rate of a machine, e.g., weather condition, operator efficiency, and equipment availability. Within these factors the most controllable factor is equipment availability (Rapp

& George, 1998). Therefore, good companies conduct periodic inspections and service to identify and eliminate potentially time-consuming breakdowns. These periodic inspections and servicing, which are scheduled by an inspector, are called Planned Maintenance (PL) or Predictive Maintenance. Also, in the 1950s, to reduce unscheduled breakdowns, progressive companies have introduced preventive maintenance (PM) (Peng, 2012). PM is scheduled maintenance work suggested by manufacturers to keep equipment in the best possible operating condition (Nunnally, 2000). According to Panagiotidou & Tagaras (2007) preventive maintenance policies helps to improve reliability and to reduce maintenance related cost. PM was a major advance which gave managers some control over equipment breakdowns. Previously equipment failures were taken as acts of nature, and there was nothing to do before breakdown (Peng, 2012).

There is another type of maintenance work which is called total productive maintenance (TPM). TPM was first practiced by Japanese firms in the 1970s, but American companies became familiar with TPM in late 1980s, when two of Seiichi Nakajima's books, *Introduction to TPM* (Nakajima, 1988) and *TPM Development Program* (Nakajima, 1989) became available in English. Nakajima (1989) said that the two objectives of TPM are zero breakdowns and zero defects. To achieve these objectives, operators are involved in maintenance work which makes them care as much about the equipment as they do about their jobs (Peng, 2012). Machine operators are trained and become responsible for some basic maintenance work, but still there is a maintenance department in the company to

handle major repairs and PM, sets the standard for maintenance work and trains operators (Peng, 2012).

Though TPM is very useful in many cases, most of the equipment-owning road construction companies do mainly three types of maintenance work: PM, predictive maintenance and reactive maintenance. These maintenance works become the most important part of the operating cost calculation. Depending on the methodology used in each company, the amount of maintenance cost varies. Maintenance cost is considered to be the highest percentage of costs for operating a piece of equipment (Peurifoy & Schexnayder, 2002). Also maintenance cost is hard to predict, and decisions regarding maintenance costs affect the hourly rate and economic life of a machine (Vorster, 2009). If economic life is increased, then the ownership cost will decrease, but to follow this approach, maintenance costs have to be increased. An equipment manager cannot control many influencing factors of maintenance costs such as weather conditions and unexpected breakdowns. On the other hand, maintenance cost estimates are greatly affected by working conditions, type of work, the operator's skill and the policy regarding operators (Vorster, 2009). Therefore proper maintenance cost estimates or accurate forecasting of maintenance budgets for coming years has become a big challenge for most equipment managers. The research for this thesis is being conducted to help equipment managers to forecast future equipment maintenance costs.

2.3 Forecasting Analysis

A forecasting analysis can be qualitative or quantitative. This analysis is used to determine what will happen in the future and is based on past experience or information from a database. A forecast is basically a prediction of what will take place in the future and according to Makridakis (1989), planning is an important aspect of forecasting. A plan is a decision-making tool that can be used to shape the future of an event. Forecasting, on the other hand, is basically used to understand whether or not that plan will work properly.

2.3.1 Qualitative Forecasting

Qualitative forecasting is made on the basis of opinion and judgment of related people or experts. Qualitative forecasting is most important when a past database is not available or when it has to be forecasted far into the future (Kim, 1989). The Jury of Executive Opinion is a formalized qualitative forecasting method in the equipment management arena (Wilson et. al., 1994). Most of the time the jury can give a better forecast than any one jury member can. There are also other qualitative methods and most do not require extensive understanding of mathematical methods. There are some basic disadvantages of qualitative forecasting. According to Kim (1989), in most cases qualitative analysis is biased and not always accurate over time. Kim also says that providing judgment for any good forecasting requires years of experience. Also, Makridakis (1989) said that people who give judgment on forecasting generally become overly optimistic. However, qualitative judgment is required in most of the decision-making

problems, such as selecting the best method of prediction analysis for a particular company (Mitchell, 1998).

2.3.2 Quantitative Forecasting

The quantitative method is used to predict maintenance and repair costs when past data is available because it is more appropriate than basing the decision on the judgment of related experts or people (Kim, 1989). There are lots of algorithms that have been proposed in last couple of years to get 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 (Makridakis et al., 1989). Among these methods, regression is the most common method, which is usually accurate over a medium-range prediction horizon (Makridakis et al., 1989). Regression analysis is a statistical analysis which shows the relationship of a dependent variable to one or more independent variables. A relationship with only one dependent variable is called a simple regression analysis. A relationship with more than one independent variable is called a multiple regression analysis. There are also linear and nonlinear regression analyses. The selection of the trend line depends mostly on the data points. Although the quantitative forecasting method is more accurate than the qualitative method, as it is based on an actual database, there are some disadvantages, too (Makridakis et al., 1989). The qualitative method's main shortcomings are that it depends on the previous database 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.

2.4 Maintenance Cost Forecasting

Most of the equipment owning company uses 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 close 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 huge difference between the actual and forecasted value.

Blaxton, Fay, Hansen & Zuchristian (2003) proposed a formula for calculating unscheduled maintenance cost as following

Unscheduled Maintenance Cost =

$$\frac{\textbf{\textit{Scheduled operating hour X Hourly Maintenance labour cost X Mean Time to Repair}}}{\textbf{\textit{Mean Time to Failure}}} \ldots\ldots (1)$$

Halpin & Senior (2011) proposed a guideline to forecast repair costs using the following formula:

Repair factor X (delivered price-tire) / 1000= estimated hourly repair reserve... (2)

Repair factors are also mentioned in a table in section 13.10 in his book.

Herbert Nichols (1976) said that an hourly repair cost can be calculated by multiplying factors for type of equipment, work conditions, total hours of use,

years of useful life, temperature, operator style, maintenance quality, type of use, luck, equipment quality and pace of work. These factors are multiplied with each other and then multiplied by 1/10,000th of the purchase price of the machine to obtain an hourly cost.

According to Nunnally (2000), the repair cost is the highest single item of operating costs for most construction equipment, which is dependent on the use of the equipment, operating conditions and the maintenance standard. He proposed an equation for hourly repair cost which is

$$\text{Hourly repair cost} = \frac{Year\ digit}{Sum\ of\ years\ digit} X \frac{Lifetime\ repair\ cost}{Hours\ operated} \quad \text{..... (3)}$$

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

Lifetime repair cost factors are given in a table of his book.

Hours operated= Expected equipment life in hours

Year Digit= Year of operating.

Sum of years digit= Sum of the years digit of the expected life. Such as, if one equipment has 5-year life, then sum of years digit=1+2+3+4+5= 15.

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 for 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 with the following form:

$MCL_D = A \times W_D + B \times W_D^2$ ..... (4)

Where

MCL_D = life-to-date maintenance cost at age W_D

W_D = life-to-date hours worked by the machine

A = coefficient that describes the linear increase of cost with age

B = coefficient that describes the exponential increase of cost with age

An example of the corresponding trend line of the equation is presented in Figure 2. In this research, similar trend lines have been built based on available data, which is described in Chapter 4.

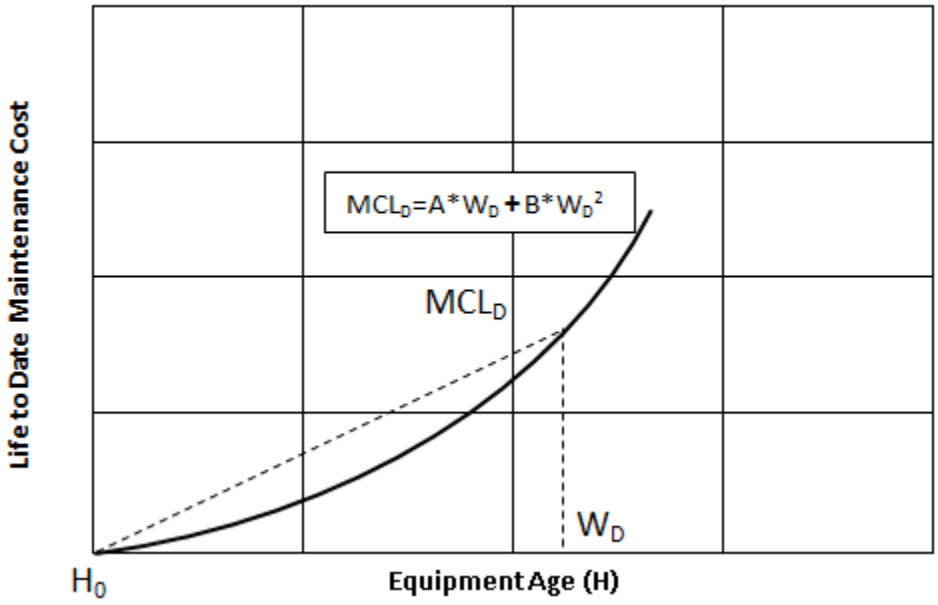


Figure 2: Trend of maintenance cost (Vorster, 2009)

2.5 Equipment Cumulative Cost Modeling (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.

Figure 3 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 and higher.

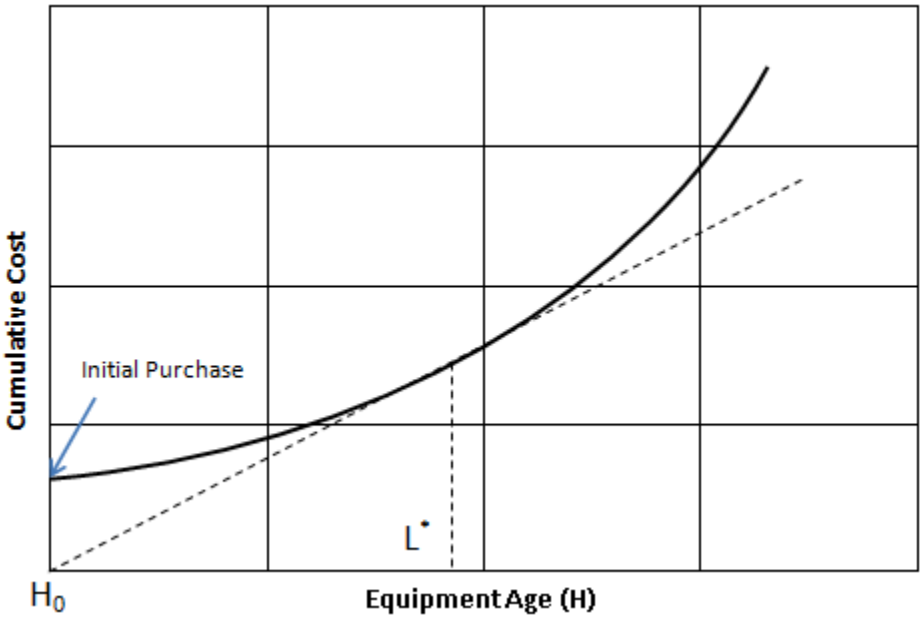


Figure 3: Cumulative cost model (Vorster, 1980)

There are several managerial decisions that can be supported by CCM such as equipment purchase decisions, preventive maintenance strategies, decisions about replacing equipment, and forecasting repair costs (Mitchell, 1998). Two approaches of CCM are used in this research; the literature review of those two approaches is described in the following two sections.

2.5.1 Life-to-Date Cost Analysis

The life-to-date (LTD) repair cost is one of the approaches for CCM analysis which is easy to understand as well as to use in any managerial decision-making situation. The first condition for this analysis is that the equipment for which the data will be analyzed by the LTD method should be of a similar type and size working under the same operating conditions (Mitchell et al., 2011). If the operating condition changes, the repair cost could also be changed on the basis of the equipment status (Nunnally, 2000). When the equipment is grouped in this way, the maintenance cost varies as a function of the equipment's age.

For each data point in the LTD method, the cumulative maintenance cost of a machine should be paired with hour meter reading or equipment age (H). 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 et al., 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 4). Through this graphical analysis the coefficients A and B for Equation 5 can be determined.

$CMC_T = A*H_T + B*H_T^2$ ..... (5)

CMC_T= cumulative maintenance cost at any age H_T

H_T= age of equipment at any time T

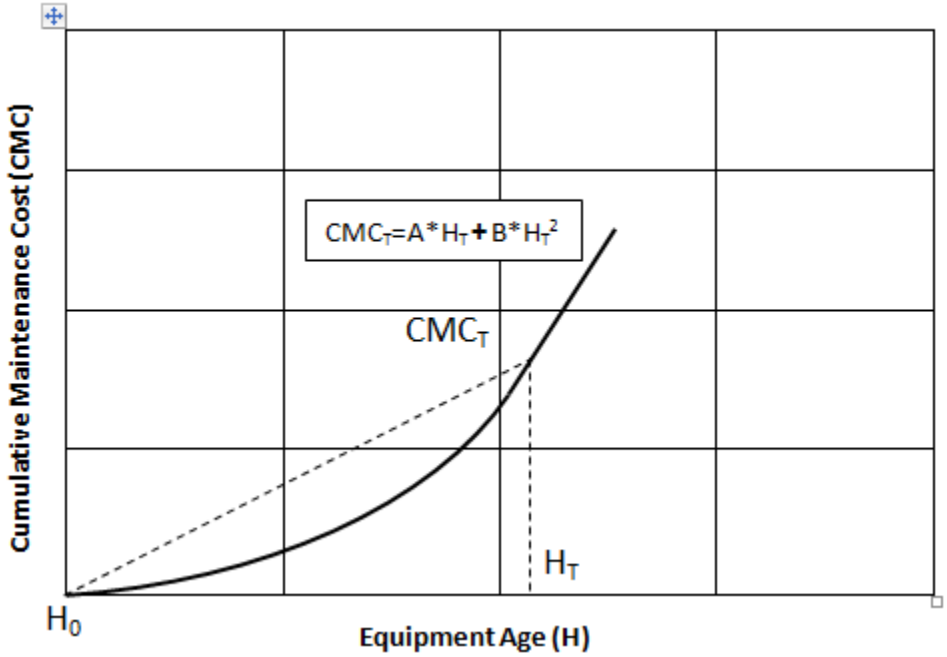


Figure 4: Life-to-date approach for cumulative cost model

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 of maintenance costs.

There are some drawbacks to the LTD approach. The first is that equipment data should be available from zero hour meter reading. That means that if the data is not available from the starting of the machine's run, then that machine's data cannot be used. In many cases a company buys used equipment from another company. It is common that many companies do not store data about their

equipment, but when they start storing data it is determined that none of the existing equipment has databases from the zero hour meter reading. The LTD approach is not workable in these cases, though it provides the most accurate picture of maintenance cost over the age of the equipment.

2.5.2 Period-Cost-Based Analysis

Another approach of the CCM is the period-cost-based (PCB) analysis, where it is necessary to have data about maintenance costs 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 5).

For a second order polynomial equation, H_M is actually 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 5 is the same as described in Equation 5. The differential equation at H_M would be

$$m= A+2*B*H_M \quad \text{..... (6)}$$

This is similar to the linear regression equation

$$Y=C + D*x \quad \text{..... (7)}$$

From Equation 6 and 7 it can be found that

$$C=A \quad \text{..... (8)}$$

$$D=2*B \quad \text{..... (9)}$$

It is easy to obtain the value of A and B from Equations 8 and 9 after obtaining the value of C and D from PCB analysis. Then, the equation of the cumulative maintenance cost with respect to the equipment age can be determined from Equation 5.

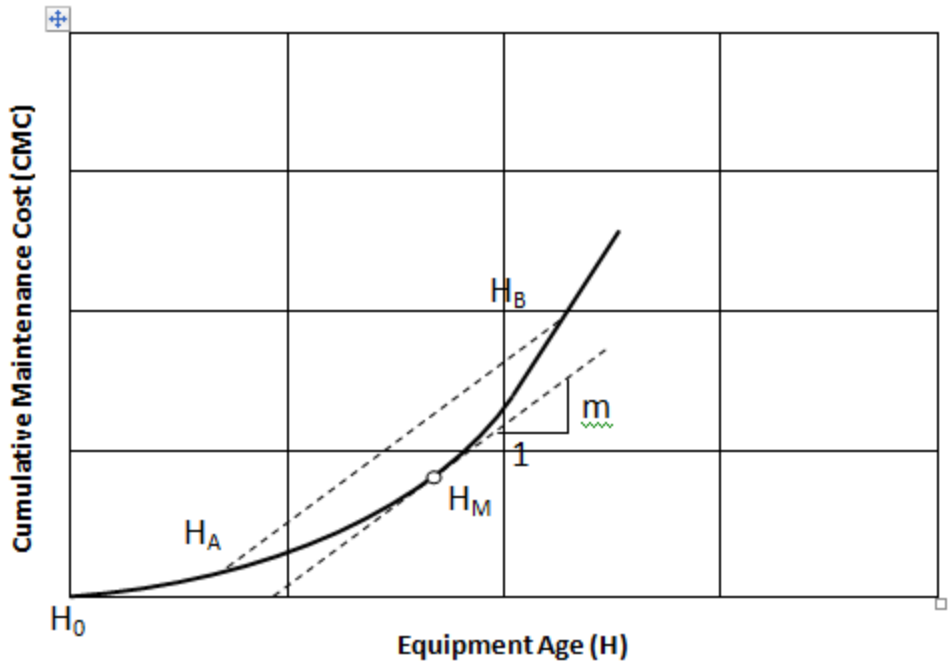


Figure 5: Period-cost-based approach for cumulative cost model

2.6 Data Mining Analysis

The world is being overwhelmed by data. It has been found in one study that the amount of data stored in all the databases in the world is being doubled every 20 years (Witten, Frank & Hall, 2011). But having a huge database is not very helpful if no knowledge can be extracted from it. Discovering new knowledge about an attribute from the pattern analysis of a database and /or prediction analysis is called data mining (Tan, Carrillo, Anumba, Bouchlaghem, Kamara & Udeaja, 2007). Fayyad, Piatetsky-Shapiro & Smyth (1996) state that discovered knowledge which is drawn from data mining analysis has to be formerly unknown, non-trivial and valuable to the customer.

In many research studies, data mining analysis was included to explore new knowledge from stored data. To evaluate the applicant's credit score Huang, Chen

& Wang (2007) built a credit scoring model using data mining analysis. For developing prediction models for breast cancer survivability Delen, Walker & Kadam (2004) used two popular data mining algorithms (artificial neural networks and decision trees) along with a more commonly used statistical method (logistic regression). Saraee, Waheed, Javed & Nigam (2005) used data mining techniques including relevance analysis, association rules mining and clustering to identify the general trends and probable solutions related to heart disease or heart attacks. In other words, data mining analysis has been utilized in many diversified fields.

Data mining analysis is becoming a popular decision-making support tool in a lot of research related to construction management. Soibelman & Kim (2002) used data mining as a tool in their research to identify valuable, applicable and unidentified 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 for Standard General Inc. He proposed non-parametric outlier mining algorithms for this decision support analysis. Gonzalez-Villalobos (2011) helped the construction managers at an industrial construction enterprise in the bidding process by doing data mining to extract embedded trends and arrangements of the bidding process. Liao & 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. Hammad (2009)

proposed that data mining analysis can be used to enhance the efficiency of labour estimating practices. His PhD thesis 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), in his M.Sc. thesis, predicted scaffold man-hours with respect to different related attributes by creating a linear regression equation in data mining software. Much research has already been conducted to improve the efficiency of data mining in industry, and studies are still being carried out. The algorithms built by data mining analysis can be classified by their results (Gonzalez-Villalobos, 2011), which are:

2.6.1 Clustering

Clustering is an unsupervised learning method that makes it possible to group records based on similarity (Foss & Zaiane, 2002). It depends mainly on the perception of minimizing the distances between data points in the same group and maximizing the distances between data points in different groups. Ankerst, Breunig, Kriegel & Sander (1999) mentioned that clustering is appropriate to initially organize a set of data into different groups to apply further analysis by other data mining algorithms.

2.6.2 Association

The second kind of unsupervised technique is association analysis. It is used most frequently to discover the shopping pattern in different stores and credit card transaction databases (Hammad, 2009). For Association Rules, two factors must be measured (Witten & Frank, 2005): the support and the confidence of the rule. Support calculates the number of occurrences for which the association rule can

forecast accurately. On the other hand, confidence measures the strength of the rule by a percentage which actually shows the precision of the rule (Gonzalez-Villalobos, 2011).

2.6.3 Classification

Classification is a supervised data mining technique where a class or attribute needs to be predicted on the basis of a model which has to be trained by the previously known dataset (Oracle, 2008). 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 data set to know which algorithm's model could give the best prediction output. The best model is used to predict output for unlabeled new data. In this research work, eight algorithms have been applied to build models and then the models have been compared to each other. A brief literature review of these algorithms is described below.

Linear Regression

Linear regression is a method that creates a relationship between a dependent and one or more independent variables by developing a linear equation that fits best to observe data (Crossman, n.d). If the relationship is with only one independent variable then it is called "simple linear regression". If the relationship is with multiple variables then it is called "multiple linear regression". In the software WEKA (Waikato Environment for Knowledge Analysis) linear regression is

basically multiple linear regression where an attribute is predicted on the basis of best suitable other attributes. The common equation for multiple linear regression is

$Y= a_0+a_1x_1+ a_2 x_2+\ldots+a_nx_{n+} \varepsilon$.....(10)

Where x_1, $x_2 \ldots x_n$ are the attributes , a_0, a_1, $\ldots a_n$ are unknown parameters and ε is a random variable (Sahoo & Jha, 2013).

Second Order Nonlinear Regression

Nonlinear regression is a nonlinear model where a dependent variable depends on one or multiple independent variables nonlinearly (IBM, 2011). Nonlinear regression is usually used where linear regression does not fit properly with the observational data. If the observational data seems to have a curvature relationship then it is better to use nonlinear regression rather than linear regression. Within nonlinear regression, second order nonlinear regression is the easiest and most popular method to use in any trend analysis problem.

The usual equation for second order nonlinear regression analysis is

$y= ax^2+bx+c$..... (11)

Where dependent variable y has a curvature relationship with independent variable x, and a, b and c are constants.

Least Median Square

Least Median Square is a regression analysis. One of the ways to make a regression more robust is to minimize the median of the squares of the difference between the data points and the regression line, which is called the Least Median Square method (Witten et al., 2011). Let's consider a set data point (x_i, y_i) Where,

$y_i = x_i a_1^* + a_2^* + e_i$ ..... (12)

$i = 1,\ldots\ldots,n$

Here $a_1^*$and a_2^* are unknown parameter vectors that have to be estimated. Let's assume an arbitrary parameter vector (a_1, a_2) and ith residual $r_i = y_i - (x_i a_1 + a_2)$. The minimization of the median of the squared residuals is entitled by the Least Median Square method (Mount, Netanyahu, Romanik, Silverman & Wue, 2007).

Conjunctive Rule

The Conjunctive Rule is a decision making rule which implements a single rule learner that predicts either a numeric or a nominal class value (Witten et al., 2011). In the conjunctive rule method, least values for many attributes have to be assigned and reject any result which does not meet the minimum value on all of the attributes (Devasena, Sumathi, Gomathi & Hemalatha, 2011). The conjunctive rule utilize the AND logical to correlate the attributes. Here the resultant is the distribution of the available classes in the dataset or mean for a numeric value of the classes. The test instances, which are usually not used to build the model, are utilized to check the accuracy of the model by the default class distribution/value (Witten et al., 2011).

Decision Stump

Decision stump is a machine learning model which is actually a one-level decision tree. It has only one interior node and uses only one attribute to predict (Sammut, 2011). This method is a weak learner or base learner which is in most of the cases used in a combined classifier (Witten et al., 2011). But decision stumps individually performs unexpectedly well on some commonly used

datasets. Its main properties are high biasness and low variance which can make the method perform much better as they are less intending to over-fitting (Sammut, 2011). In this method the main thing is selecting the best attribute which has best score defined as

$$\text{Score (A)} = \frac{\max(A=C, A\neq C)}{n} \quad \text{(Ai \& Langla, 1992)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (13)$$

For this equation on a training set of size n, the number of times the concept (C) and the attributes (A) have the same value (A=C) and different values (A≠C) have to be calculated.

M5Rule

The M5Rule is basically a regression rule obtained from model trees. In this method, a model tree is applied to the full training dataset, and the best leaf is selected and made into a rule. Then that tree is redundant. All data instances utilized by the rule are also discarded and removed from the dataset. This procedure is applied to the remaining instances and finished when all instances are enclosed by one or more rules (Holmes, Hall & Frank, 1999).

There are mainly two stages in M5Rule analysis- building the initial tree using a splitting criterion, and pruning the tree. A decision tree algorithm is used to build an initial tree. Then a splitting criterion is used. The splitting criterion is based on calculating the reduction of standard deviation at a node for each possible test (Wang & Witten, 1996). This splitting criterion can be called the standard deviation reduction (SDR)

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (14)$$

Here, T is the set of instances that reach the node and T_i donates the ith subset out of one potential test. After investigating all possible tests, the method chooses the test which maximizes the SDR value (Quinlan, 1992). In this way, sometimes an expanded tree which needs to be pruned backward can be built up.

In the pruning process, first the average of the absolute difference between the actual target value and the predicted value from the model has to be calculated for each of the training examples that reach the node. This absolute difference has to be multiplied by a factor (n+v)/(n-v) to reduce the effect of unseen cases (Wang & Witten, 1996). Here, n is the number of the training set and v is the number of parameters in the model at the same node. This is called error estimation in the pruning process.

In this method a linear regression model is computed for each interior node of the unpruned tree using the parameters that are tested in the sub tree below that node. This linear model is simplified by dropping parameters to reduce the estimated error. The parameters have to be eliminated one by one until the error estimate decreases. On the basis of the lowest error estimate, the M5Rule chooses either the sub-tree or the simplified linear model. When the estimated error is lower for the linear model, then the sub-tree at this node is pruned to a leaf (Wang & Witten, 1996).

REP Tree

The REP Tree is a decision tree which uses information gain or variance reduction and prunes in the prediction process (Ali, Tickle & Pang, 2008). As the decision tree in most of the cases faces an over-fitting problem caused by the

noised training data, a pruning process is required to remove the sub-trees resulting from the noise (Park, Hsiao-Rong Tyan, & Kuo, 2006). The REP tree is a fast-pruning algorithm which can deal with the noisy training data in a very effective way.

Multilayer Perceptron

Multilayer Perceptron (MLP) is a popular neural network. It is an artificial neural network that is used widely for pattern recognition and interpolation (Noriega, 2005). MLP is a supervised learning algorithm, so it requires a sufficient amount of input and output data. MLP is broadly used for pattern classification because it gains knowledge about how to convert input data into a desired output (Panchal, Ganatra, Kosta & Panchal, 2011).

2.7 Model Training and Testing Options

There are several options or methods for using datasets for training and testing models, such as using the same dataset for testing and training, percentage split, cross-validation. Among these methods, cross-validation and percentage split are used in most cases.

Cross-validation is a popular statistical method of evaluating algorithms or models. There are several forms of cross-validation but the basic one is a k-fold cross-validation. In this method, the data is randomly split into k-folds of approximately equal size. Then, k-1 folds are used for the training model and the remaining fold is used for testing. This procedure is repeated k times and each fold is utilized for testing once (Duan, Keerthi & Poo, 2003). For estimating and predicting scaffold man-hours, Chandan (2013) used cross-validation as a testing

and training option. To estimate the performance of classifiers, Delen et al. (2004) used a stratified 10-fold cross-validation approach. Duan et al. (2003) also used a 10-fold cross-validation approach in data mining to predict breast cancer survivability.

In the percentage split method, the total dataset is divided into two parts. One part is used as testing and the other part is utilized as training data. This split of data can be random or ordered. The main advantage of percentage split is that the training model cannot see the testing data, so proper testing or evaluation of a model is possible in this mode. For the application of data mining techniques in the medical field, specifically in the areas of heart disease or heart attacks, Saraee et al. (2005) split the medical data into two parts for training and testing. For medical image classification, Antonie, Zaiane & Coman (2001) used 90 percent of the data for training models and the remaining 10 percent for testing in data mining analysis.

2.8 Model Evaluation and Validation

Models can be built according to the developer's needs. But to implement this in the real world, the model developer must first conduct a model evaluation and validation. A model evaluation compares model results with the data from field experiment results or the real data from observations (Drinking Water Source Protection, 2013). The main goal of evaluation is to provide useful feedback about the model results to different users, including sponsors, clients, supervisors, team leader and other relevant parties (Trochim, 2006). The evaluation is useful if it helps the users to make constructive decisions.

The model validation is the process of calculating the accuracy of a simulation or prediction model and its associated prediction, compared to what the output would be in the real world (Gore, 2010). A model is basically developed for a specific purpose or purposes. A model can be said to be satisfactorily validated if it can satisfy all the purposes within a certain range (Sargent, 2013). It might happen that a model can satisfy some purposes or conditions but not all of them. In these cases it has to be specified first which conditions have to be satisfied and the range of satisfaction for each condition. If any of the required conditions cannot be satisfied by the model, then the model has to be deemed invalid (Sargent, 2013).

There are many different ways of evaluating a model depending on the model and the function of the evaluation. The most important basic types of evaluations are formative and summative evaluation (Trochim, 2006). A formative evaluation usually improves or strengthens the model, but a summative evaluation examines the outcome of the model. Some examples of formative evaluations are need assessment and implementation evaluation, and some examples of summative evaluations are outcome evaluation, cost-benefit analysis, and impact evaluation (Trochim, 2006).

There are qualitative and quantitative approaches for model validation. For the quantitative approach of validating a model, the occurrence distribution for all conditions of the real world system have to be compared with the occurrence distribution for the same conditions of the simulated model (Viet, Fourichon, Jacob, Guihenneuc-Jouyaux & Seegers, 2006). The qualitative approach is one of

the most commonly used to validate a model which is most likely biased in many occasions and for which no solid conclusions can be found from relatively complex models (Campbell & Bolton, 2005). Though qualitative methods such as graphical comparison of model output and experimental real data are commonly used in engineering, the quantitative method provides a systematic way to calculate errors and uncertainty of a prediction model with the occurrence in the real field (Ling & Mahadevan, 2013). For quantitative validation, usually the total database has to be divided into two parts: one part for training and the other part for testing. Hammad (2009) used 85 percent of the database randomly for building a model by data mining. He used the rest for testing.

The choice of proper model evaluation and validation approach depends on the model, purpose of the model and the database. Hammad (2009) calculated estimating error of his data mining model for validating it. Mitchell (1998) used a coefficient of determination (R^2) for evaluation and cross-validation of his model. As the database is not large enough, he preferred cross-validation instead of splitting the database. Poveda (2008) compared the crisp output obtained for each data with the actual output by calculating the percentage error. There are many other ways of evaluating or validating a model. In this research work, five model evaluation or validation methods have been used: mean absolute error, root mean absolute error, relative absolute error, root relative squared error and correlation coefficient. All are briefly discussed below.

2.7.1 Mean Absolute Error and Root Mean Squared Error

The mean absolute error (MAE) is a parameter which is used to measure how close predictions are to the corresponding observation. It is the average difference between the values which are obtained from the prediction or forecasting model and the corresponding observed real value (Willmott & Matsuura, 2005).

$$MAE=\frac{(P_1-A_1)+\cdots+(P_n-A_n)}{n} \quad \text{..... (15)}$$

Mean squared error is the average of the squared difference between the model prediction value and the corresponding observed real value. The root mean squared error (RMSE) is just the square root of the mean square error (Willmott & Matsuura, 2005). So the equation is

$$RMSE=\sqrt{\frac{(P_1-A_1)^2+\cdots+(P_n-A_n)^2}{n}} \quad \text{..... (16)}$$

The RMSE provides relatively high value to large errors because the errors are squared before they are averaged. This means that the RMSE is most valuable when large errors are unwanted.

The RMSE is always larger than or equal to the MAE. If the RMSE is equal to the MAE, then all the errors or individual differences are of the same magnitude. According to Willmott & Matsuura (2005), the RMSE is badly chosen for any validation because it depends on three characteristics of a set of errors (error magnitudes, square root of the number of errors ($n^{1/2}$) and the average of the error magnitude), whereas the MAE depends on only one (the average of the error magnitude).

2.7.2 Root Relative Squared Error and Relative Absolute Error

The Root Relative Squared Error (RRSE) is calculated by dividing the Root Mean Squared Error (RMSE) of the prediction model by the RMSE obtained from predicting the mean of the actual value, and then multiplying by 100.

$$RRSE=\sqrt{\frac{(P_1-A_1)^2+\cdots\ldots+(P_n-A_n)^2}{(\bar{A}-A_1)^2+\cdots\ldots+(\bar{A}-A_n)^2}}\,\%$$ (Witten et al., 2011)..... (17)

Where, P= Prediction from the model

A= Actual value

$\bar{A}$= Mean of Actual Value

RRSE means how much better is the prediction of the developed model with respect to the prediction of the mean of the actual value. Smaller values for the RRSE are always better and values greater than 100 percent indicate that the model is doing worse than predicting the mean of actual value.

Relative absolute error (RAE) can be calculated in the same way. RAE acquires the total absolute error and divides it by the total absolute error of the mean of the actual value (GeneXpro Tools, n.d.).

$$RAE=\frac{(P_1-A_1)+\cdots\ldots+(P_n-A_n)}{(\bar{A}-A_1)+\cdots\ldots+(\bar{A}-A_n)}\%$$ (Witten et al., 2011) …….. (18)

Where, P= Prediction from the model

A= Actual value

$\bar{A}$= Mean of Actual Value

2.7.3 Correlation Coefficient

Correlation coefficient is the comparison between the variance of the prediction value and the variance of the actual value. It is a single number between +1 to -1

that gives a good idea about how closely one variable is related to another variable and it is denoted by "r" (Higgins, 2006). It is also called Pearson's Correlation Coefficient because the calculation method was developed by Karl Pearson. The correlation coefficient will be +1 or –1, if the two variables are in an ideal linear relationship, and will be 0 if there is no linear relationship between the variables (Witten et al., 2011). The equation for the correlation coefficient is given below.

$$\text{Correlation Coefficient} = \frac{\sum_i (P_i - \overline{P})(A_i - \overline{A})}{\sqrt{\sum_i (P_i - \overline{P})^2 \sum_i (A_i - \overline{A})^2}} \quad \text{..... (19)}$$

Where,

$\mathbf{P_i}$ = Prediction value for the ith test instance

$\overline{\mathbf{P}}$ = Average of the prediction value.

$\mathbf{A_i}$ = Actual value for the ith test instance

$\overline{\mathbf{A}}$ = Average of the actual value

From equation 19 it can be understood that the Correlation Coefficient evaluates predicted values (P_i) by comparing them with actual values (A_i).

2.9 Conclusion

This chapter has reviewed many different scholars' methods and algorithms. The chapter also included discussions about the way in which many researchers have forecasted maintenance costs. It also reviewed the methods and algorithms that were used for maintenance cost forecasting, model evaluation and validation. On the basis of this knowledge, the main analyses were conducted. These analyses will be discussed in Chapter 4.

Chapter 3: Data Collection and Pre-processing

3.1 Introduction

Alberta is the fourth largest province in Canada. It has 31,000 km of highways which form an extensive network throughout the province. This province has 226,000 kilometres of public roads – approximately 22% of the total national network (Government of Alberta, 2013). To build and maintain these road networks, many road construction contractors have been working in Alberta for a long time. One of the contractors is Standard General Inc., which has many years of experience in road construction. The company usually owns and sometimes rents equipment for this road construction work. Standard General Inc. has to spend huge amounts of money every year to maintain its fleet of different types of equipment. The company stores its data in an MS SQL server and needs to utilize the database to forecast maintenance costs for budgeting and replacement purposes. A primary step of this knowledge-discovering research work was to extract the data from the database and maintain a data warehousing system. Various obstacles came up during the data extraction and pre-processing stage of the research work. These obstacles consumed large portion of the research time. In this chapter, the background and the whole process of the data extraction and pre-processing are discussed. The complications faced during the first 10 to 12 months of this research for data preparation are described in the following sections.

Section 3.2 discusses how Standard General Inc. manages the equipment management database with the software M-Track. This section explains the total structure of M-Track.

Section 3.3 explains how Standard General Inc. usually maintains its heavy equipment. The section also describes the current practice that M-Track uses to obtain the total cost of the different maintenance work.

Section 3.4 introduces this research work's data warehousing system. It explains how different types of maintenance cost data were collected to obtain the total maintenance cost of the equipment.

Section 3.5 describes the total work of pre-processing the database. To modify the database according to the research interest, several steps of modification were taken. These steps are explained with examples.

Section 3.6 presents the conclusion of this chapter.

3.2 Construction Equipment Data Management by M-Track

Data collection is a pre-requisite of every job for any big construction company. A construction contractor collects data on a daily basis for management of operation, repair, maintenance and purchase. The development of computer technology and digital control encourages large contractors to invest and implement these technologies in a data collection and distribution system (Fan, 2007). Standard General Inc. had taken an initiative to implement an equipment management software called M-track. The system has been in place since 1997 and since then it has been redeveloped several times. The system is a client-server application with the server part as a database on the MS SQL server. The system's

main features are purchasing and inventory management; shop labour time management; equipment maintenance services such as running repairs, preventive maintenance, and planned maintenance. The M-track data management system is shown in Figure 6.

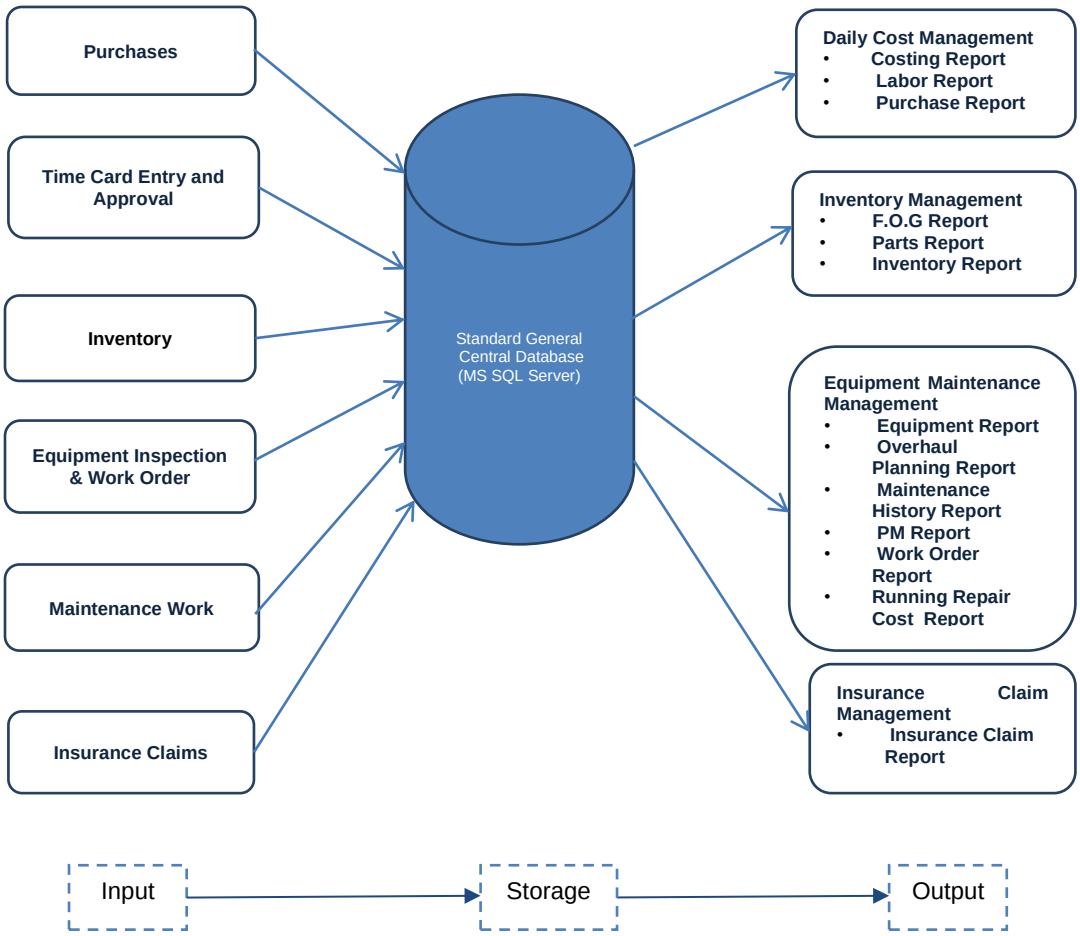


Figure 6: Data management system of M-track

3.3 Maintenance of Heavy Equipment in Standard General

Standard General Inc. has a large fleet of construction equipment. They have trucks, loaders, dozers, scrapers, graders and many other types of equipment which are mostly required for road construction and maintenance. Whether the company purchases or rents equipment for the fleet, it places importance on the

maintenance of all of the equipment. The company mainly performs three types of maintenance work on 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 and periodic maintenance work that is suggested by the manufacturer to keep the equipment in the best possible working condition (Nunnally, 2000).

Maintenance costs consist of costs for labour and parts. The system supports the time entry of labour work where the labour hour for both running repair and planned maintenance can be found. Parts cost due to running repair and planned maintenance can be found in the purchase order and on the inventory tables. It is easy to obtain the labour and parts cost 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 7.

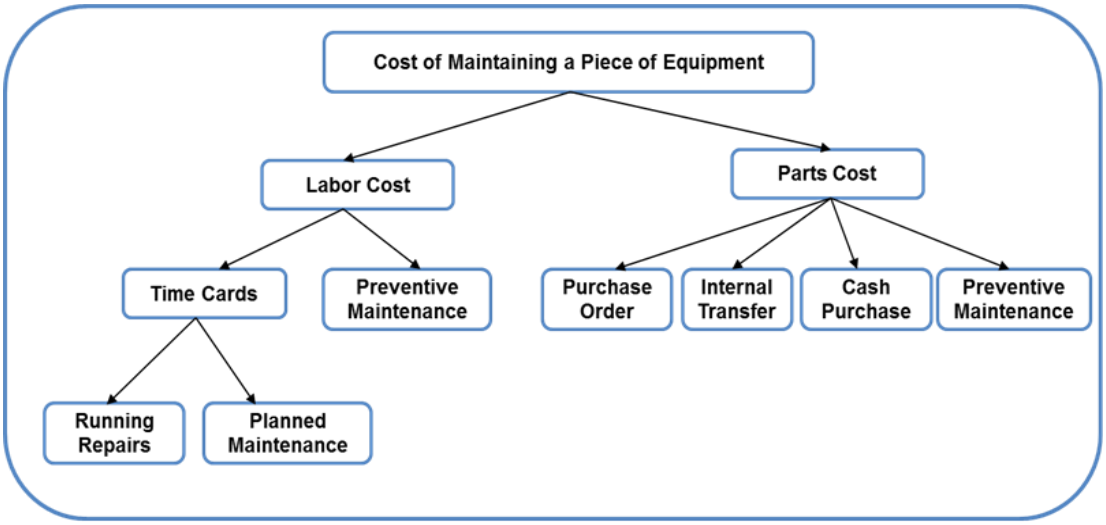


Figure 7: Structure of total maintenance cost of equipment (M-Track)

3.4 Data Warehousing

One of the advantages of the data warehousing system is the integration of data that is distributed in different systems of the company. For this research work only one database was used. From this database, required datasets for obtaining the total maintenance cost of equipment were imported into Microsoft Access. From the collected datasets, a different query was done to obtain 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)) as shown in Table 01. The purchase order, internal transfer and cash purchases are done for both RR and PL. A sample dataset of different components of maintenance cost is shown in Table 1. The process of preparing tables for different components of total maintenance cost from the SQL server is shown in Figure 8.

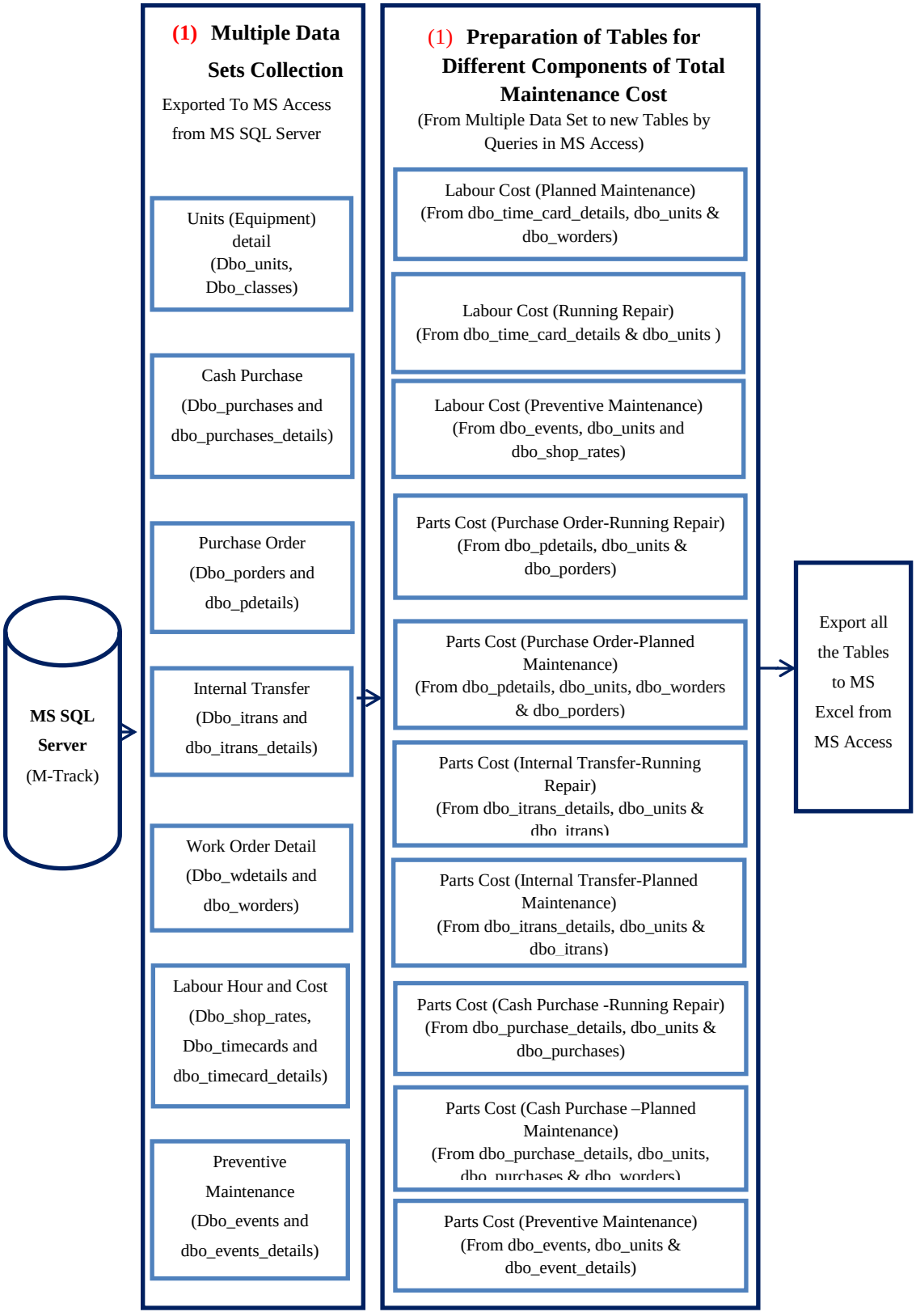


Figure 8: Summary of data warehousing system

Table 1: Sample data of different components of equipment maintenance cost (Bayzid, Al-Hussein & Mohamed, 2013).

			Labour Cost ($)			Parts Cost($)						
Equipment Unit No	Time Stamps	Hour Meter Reading	RR	PL	PM	PO_ RR	PO_ PL	IT_ RR	IT_ PL	CP_ RR	CP_ PL	PM
217-401	9/17/2001	2091	1,076.25	-	105.00	549.55	-	35.85	-	-	-	55.23
217-401	4/22/2002	2224	4,908.75	-	420.00	3,189.30	-	453.60	-	-	-	514.66
217-401	9/16/2002	2406	1,897.50	-	105.00	45.96	-	-	-	-	-	38.52
217-401	5/9/2003	2487	918.75	-	315.00	-	-	63.45	-	-	-	312.15
217-401	10/14/2003	2671	1,653.75	-	105.00	6,987.00	-	27.60	-	-	-	44.67
217-401	5/18/2004	2698	2,362.50	-	315.00	754.69	-	807.60	-	-	-	531.43

3.5 Data Pre-processing

In the database of M-Track, running repair and planned maintenance cost data were stored with respect to timestamps. Just for preventative maintenance, both the hour meter reading and timestamp were stored for each of the readings, but for this study the 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 database for all of the maintenance cost data with respect to the hour meter reading was prepared. This process is discussed in Figure 9 and a sample of the dataset is presented in Table 1.

Data entry is error-prone. So, data inconsistency and missing data are common in most of the database. These kinds of irregularities have to be figured out and then

proper ways need to be found to resolve them. In this application some similar outliers were encountered, such as a change of the hour meter reading when the hour meter was replaced, mistakes during entry of hour meter readings, etc., which are illustrated in Table 2. In the third row of Table 2, problems due to the replacement of the hour meter are shown. This was resolved by adding the hour meter reading with the last reading of the previous hour meter. In the sixth row of Table 2, an outlier due to misleading hour meter reading is shown. This was solved by taking the average of just the previous hour meter reading and the next hour meter reading. Although these approaches do not give the accurate hour meter reading, it is close to the actual number.

Table 2: Illustration for hour meter reading correction (Bayzid et al., 2013)

Equipment Unit No	**Event Id**	**Time Stamp**	**Hour meter Reading (From the database)**	**Hour Meter Reading (Corrected)**
205-404	117148	20/03/2009	520	520
205-404	119452	03/02/2010	547	547
205-404	**121987**	**16/02/2011**	**2**	**549**
205-404	123530	12/01/2012	108	655
230-405	108686	10/04/2006	6783	6783
230-405	**109934**	**16/10/2006**	**255**	**7046.5**
230-405	111093	09/04/2007	7310	7310

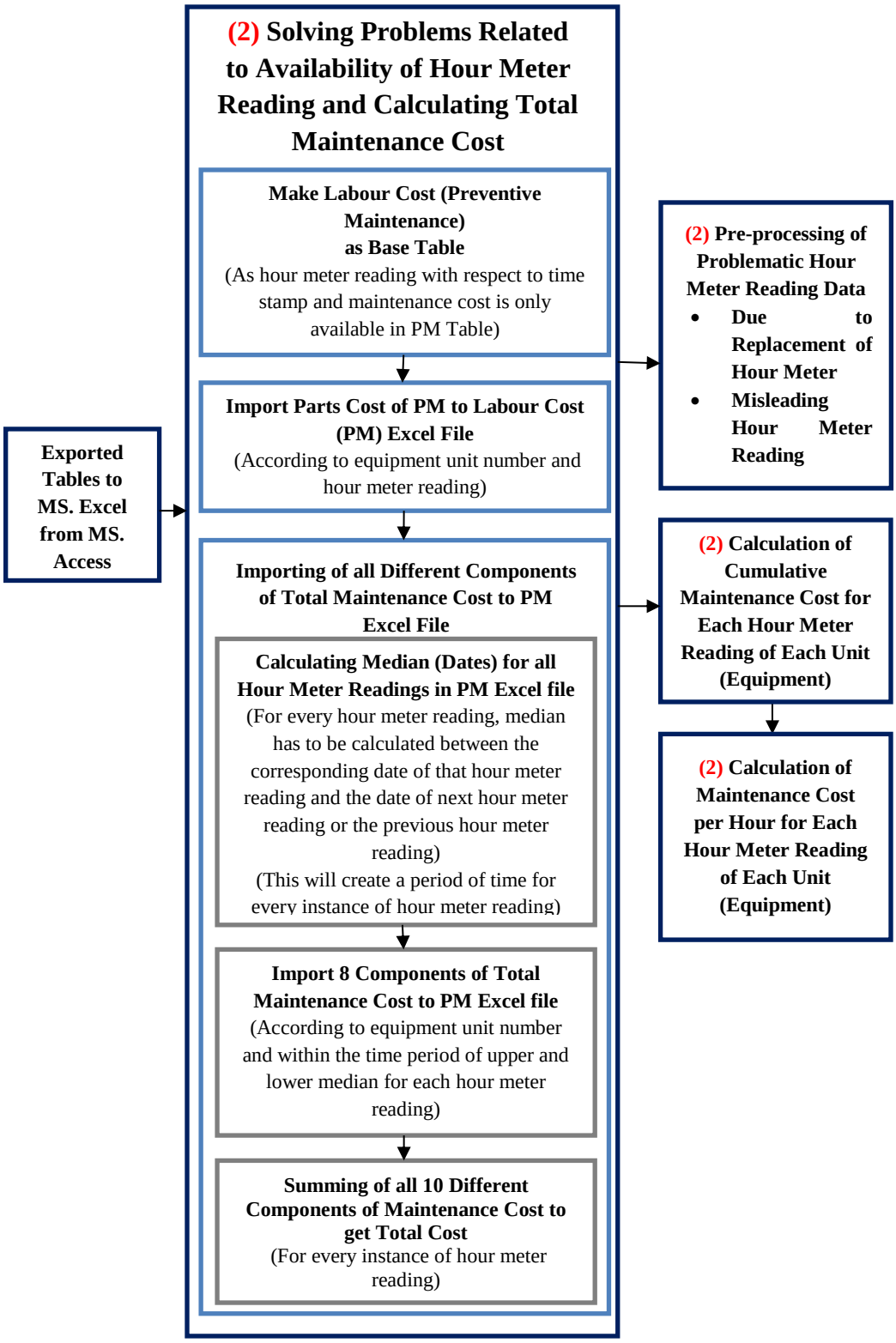


Figure 9: Summary of data preparation

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 $H_T = TMC_1 + TMC_2 + TMC_{3....} + TMC_T$ ..... (20)

Where, H_T = Hour meter reading at any time T.

TMC_T= Total Maintenance Cost at any time T

Maintenance Cost/ Hour= $(CMC_T - CMC_1)/ (H_T - H_1)$..... (21)

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

An example of the calculations is given in Table 3 which is extension of Table 1. Total maintenance cost of Table 3 is obtained from the summation of all the components of maintenance cost from Table1.

Table 3: Sample of calculation of maintenance cost in $/ hour

Equipment Unit Number	Hour Meter Reading	Total Maintenance Cost in $	Cumulative Maintenance Cost in $	Maintenance Cost in $/Hour
217-401	2091	1,214.63	1,284.63	
217-401	2224	6,324.25	7,608.88	47.55
217-401	2406	1,391.33	9,000.21	24.49
217-401	2487	1,072.98	10,073.19	22.19
217-401	2671	5,878.80	15,951.99	25.29
217-401	2698	3,180.87	19,132.86	29.40

3.6 Conclusion

In most cases, the hardest part of quantitative research work is collecting data and pre-processing it according to the research interest. This chapter described the scenario of the database, how the database is being managed by M-Track, and M-Track itself. Then examples were given for describing the complications regarding the database and how it was solved. In the beginning of the research work it was thought that with only one database, data warehousing and pre-processing of database would not be complicated. But afterwards it was found that though the research is based on only one database, it was a bit hard to figure out the way to rearrange or pre-process it for this research work. After several discussions with experienced personnel, the problems were solved one by one. Though it took a long time, developing the pre-processed database provided a strong foundation to go forward with the research work.

Chapter 4: Prediction Models for Equipment Maintenance Cost

4.1 Introduction

Prediction models of equipment maintenance cost were proposed by many researchers in different ways. Some of those models were briefly discussed in Chapter 2. In this chapter, first a general trend analysis is discussed and then the prediction analysis is explained. The general trend analysis is done to organize the procedure of prediction analysis and to obtain a basic concept about which equipment classes have sufficient and reasonable datasets to use for prediction purposes. After trend analysis, prediction analysis was done first by cumulative cost modeling (CCM) and then data mining analysis was conducted. The last step was to compare, evaluate, and validate these methods and algorithms. The total analyses of this research and the outcomes of all analyses are discussed in the following sections:

Section 4.2 presents the steps of the analysis work for this research. Here the whole analysis work is summarized in one figure.

Section 4.3 describes the general trend of different equipment classes. The usual and unusual trends are discussed here.

Section 4.4 displays how CCM can be used for this database.

Section 4.5 presents the data mining part. In this section, prediction analysis by most popular algorithm, second-order nonlinear regression, is discussed first. Then, data mining analysis using the Waikato Environment for Knowledge Analysis (WEKA) for seven other algorithms is elaborated.

Section 4.6 explains how these algorithms or models were evaluated and/or validated. This part is fully based on the illustration of the results from the analyses of the previous sections.

Section 4.7 presents the conclusion about the outcomes of these analyses.

4.2 Summary of Analysis Work

From the prepared database that is described in Chapter 3, two types of analyses were conducted: general trend analysis and prediction analysis. Prediction analysis was done by CCM and data mining analysis. The total process of these analyses is presented in one figure (Figure 10). This figure is elaboration of part (3) of Figure 1. Also, different portions of this figure will be elaborated in the following sections of this chapter.

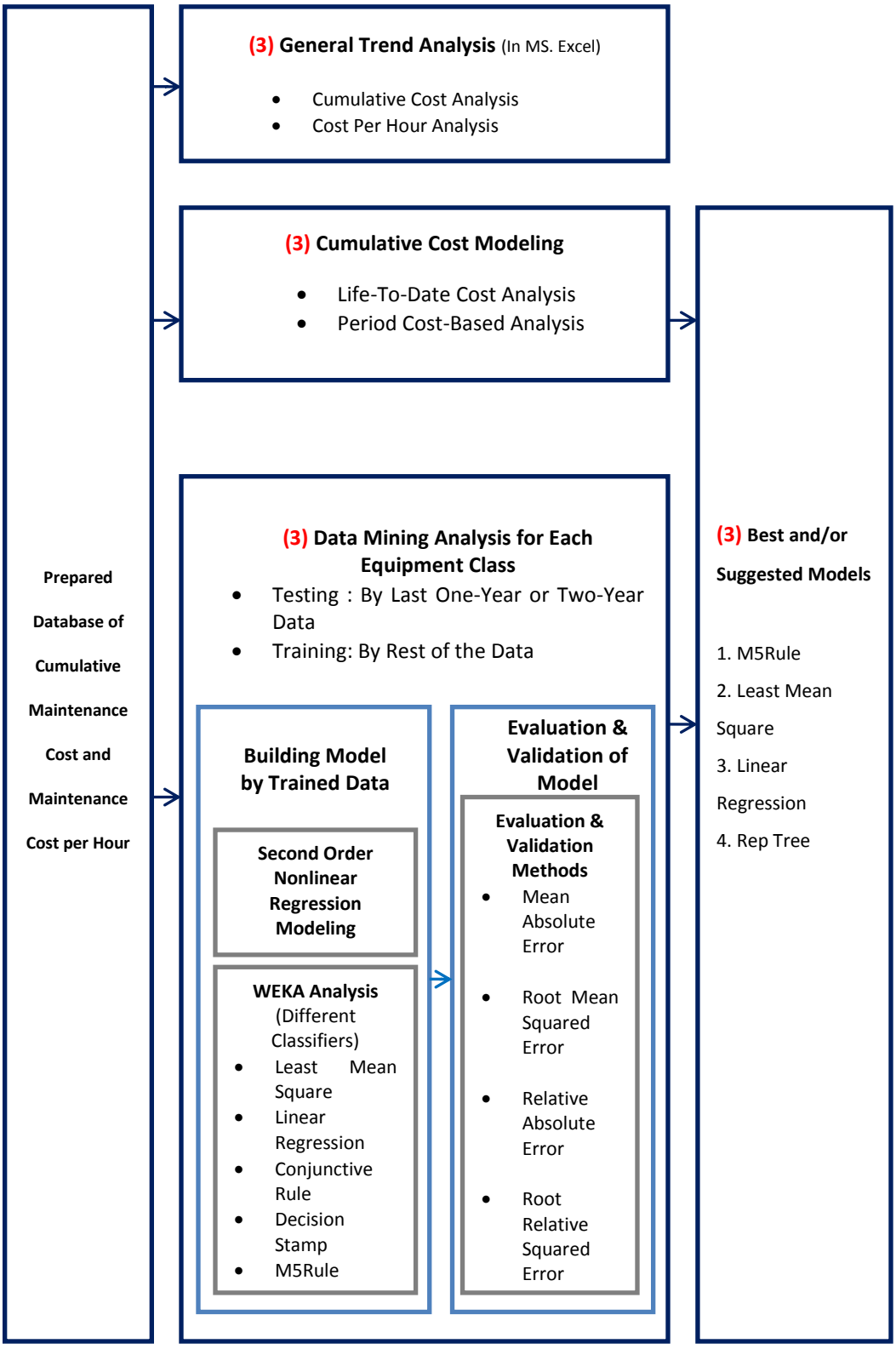


Figure 10: Summary of analysis work study

4.3 General Trend Analysis

When the pre-processed database was ready, the first step was to explore the trend of maintenance cost for different equipment classes. Similar equipment was grouped into equipment classes. In this research work, the analysis was conducted on the basis of equipment class, because it was assumed that within the same class the behaviour of the equipment was almost the same. According to the needs of the company, the research work was driven on all available equipment classes between class numbers 200 and 299. Fifteen equipment classes were found, which provided sufficient data for trend analysis within this interval.

Cumulative cost analysis was taken as a primary initiative of this research work for the basic trend analysis. After generating some graphs of cumulative maintenance cost vs. hour meter reading, it was realized that for this trend analysis the approach was not a good option. As the fleet of equipment consists of old and brand new equipment, maintenance cost data for many equipment units were not available from the zero hour meter reading. For this reason the trends in the graph show up individually by piece of equipment, not by the class of equipment, as seen in Figure 11 for equipment class 240-Graders (150 to 225 hp).

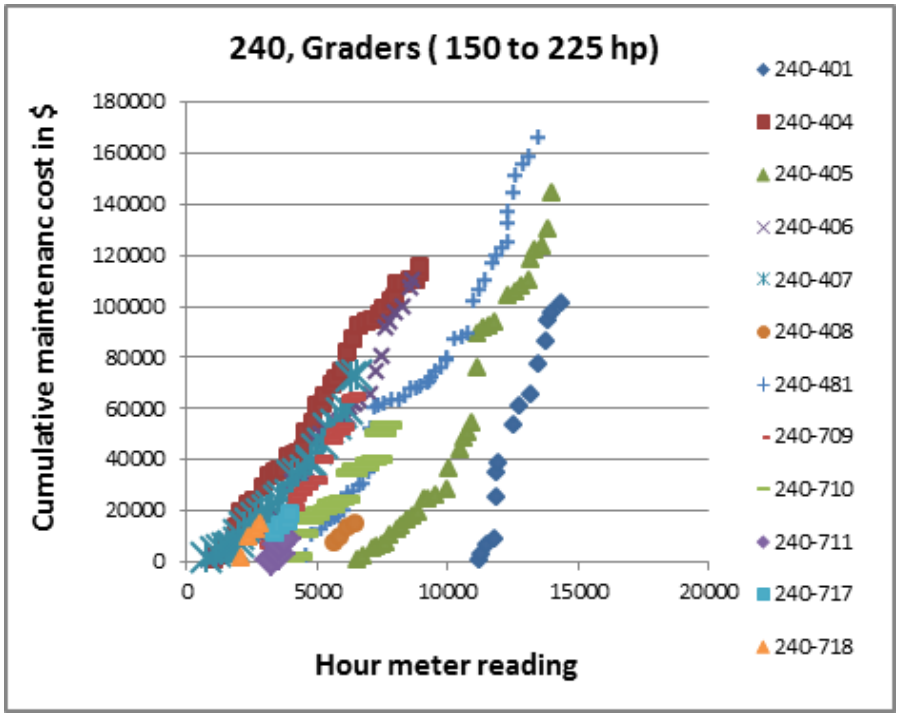
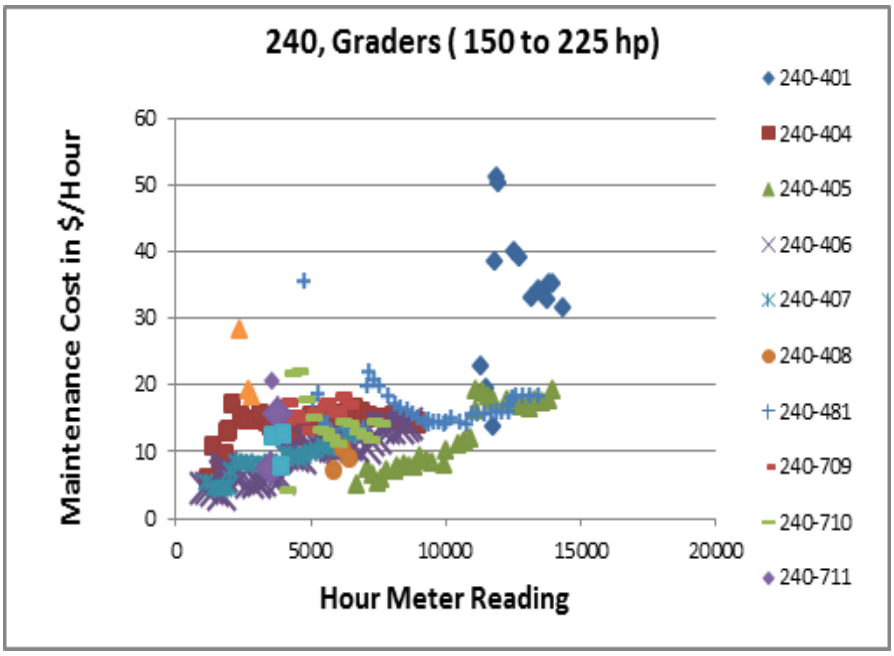


Figure 11: Cumulative cost analysis

After cumulative cost analysis, a cost-per-hour analysis was attempted. In Figure 12 to Figure 14 the trends of maintenance cost per hour are illustrated for three of the available equipment classes. For each of the figures, the top graph shows the trends of different equipment units within the same equipment class and the bottom graph shows the trend of all the equipment units together as an equipment class.

According to Mitchell et al. (2011) variability of repair cost with respect to hour meter reading can be explained more rightly by grouping the equipment as an equipment class. In the statistical analysis, using larger dataset creates less influence of a particular machine which is performing unexpectedly well or poor. Also according to Vorster (2009) when the group size is large, more confidence can be provided on the target output. For this reason, more importance was given on the equipment class for trend analysis, CCM and data mining analysis.

The illustrated problem mentioned previously in this section was solved in a cost-per-hour analysis, but the trends of maintenance cost per hour were not always as they are described in the literature review. Usually the trend of maintenance cost increases when the equipment becomes older, but in some equipment classes of equipment this did not happen. Figure 12 presents an upward trend for the maintenance cost per hour of graders (150 to 225 hp). Figure 13 presents the same upward trend for a cement spreader and concrete paver. However, in Figure 14 the trend is unusual compared with the one in the literature review, because it shows a trend of decreasing maintenance cost per hour with respect to the hour meter reading. This kind of unfamiliar trend was found for 4 (equipment class 217, 219, 243 and 253) out of 15 equipment classes. These trend analyses for all of the available equipment classes are presented in Appendix 1. To find out the reasons behind this unfamiliar trend, a couple of meetings with an expert in this particular field were conducted. By analyzing the data and from their own points of view it was found that, for most of the cases the initial high preventive or planned maintenance could be the main reason behind these downward trends. This maintenance works is usually done when the company buys used equipment.



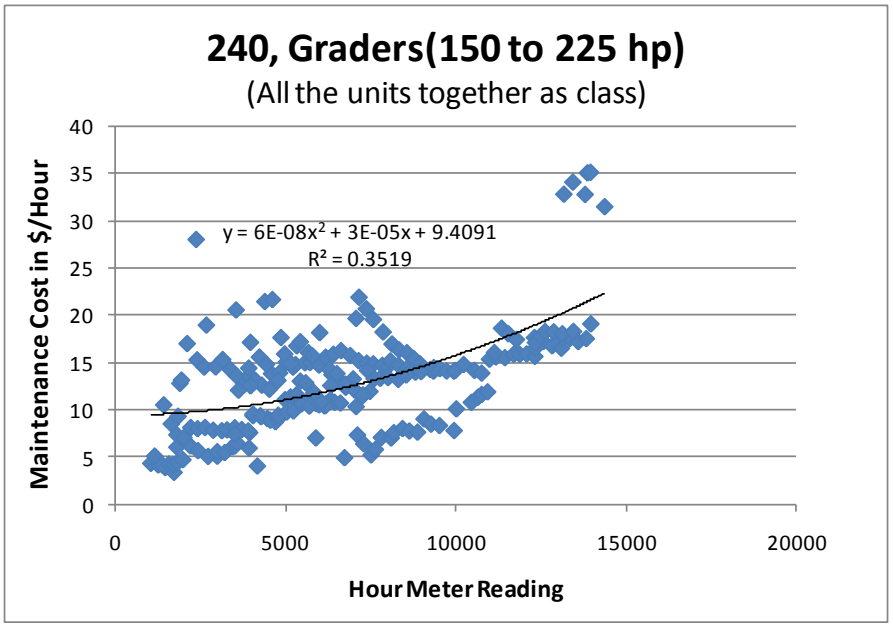
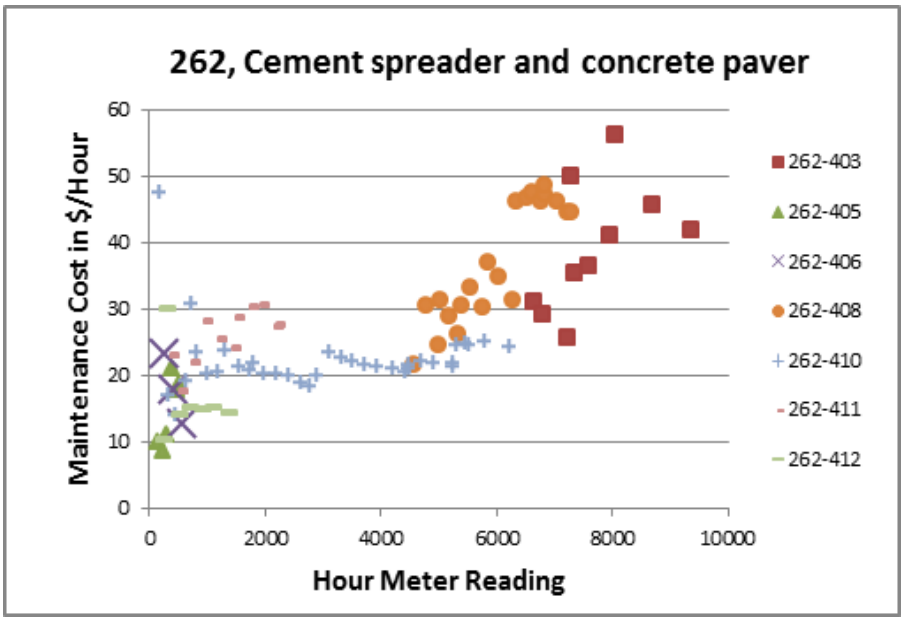


Figure 12: Cost-per-hour trend analysis for equipment class 240, graders (150 to 225 hp)



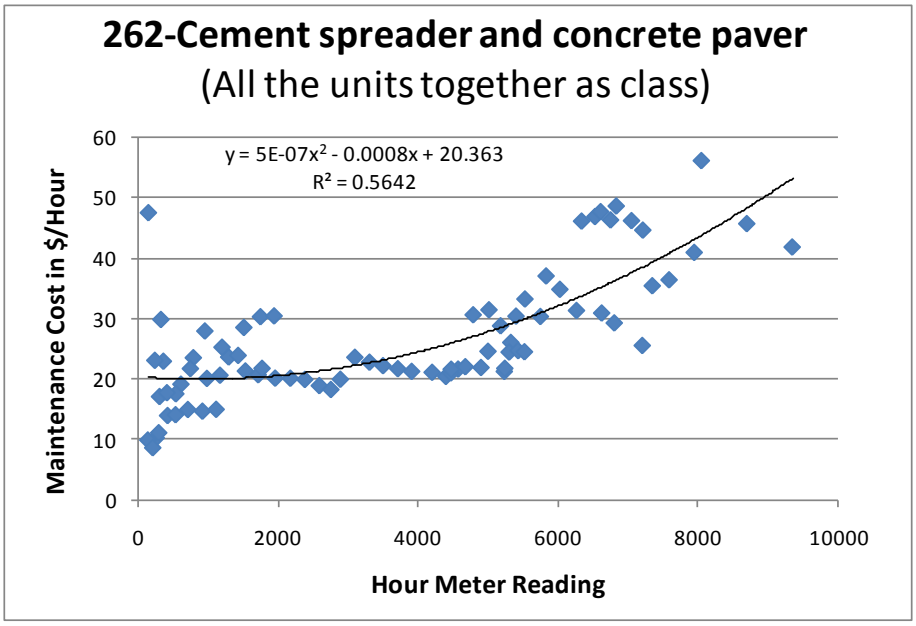
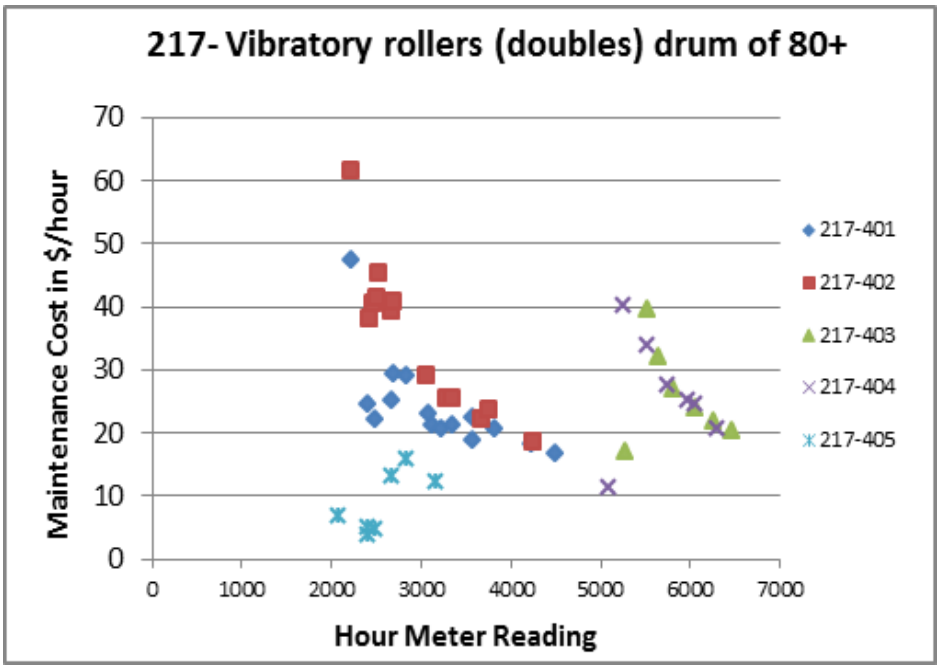


Figure 13: Cost-per-hour trend analysis for equipment class 262, cement spreader and concrete paver



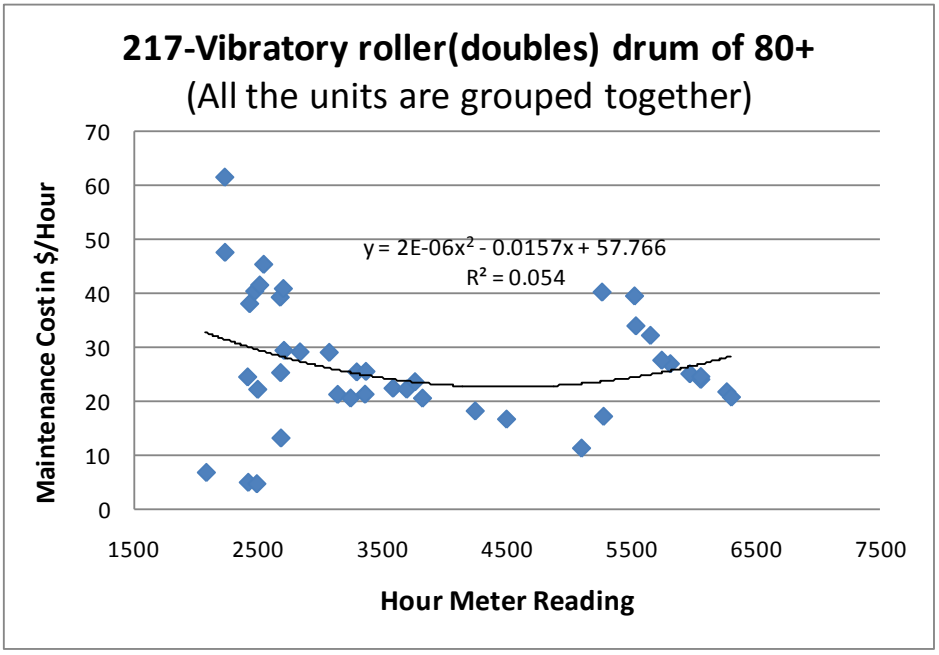


Figure 14: Cost-per-hour trend analysis for equipment class 217, vibratory roller (doubles) drum of 80+ (Bayzid et al., 2013)

4.4 Cumulative Cost Modeling

Life-to-date (LTD) and part-cost-based **(**PCB) regression analysis are two approaches of CCM. For these two approaches, a satisfactory amount of

equipment units in an equipment class is required. However, most of the equipment classes have 5 to 15 equipment units and very few equipment classes contain more than 2 equipment units with enough amounts of 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 compatible for these analyses.

Sample data for the LTD analysis of equipment class 240, graders (150 to 225 hp), is shown in Table 4. Only 3 equipment units were found which are suitable for the LTD analysis. The age or hour meter readings were divided by 1000 to obtain A and B values in Equation 5 in the literature review. Cumulative maintenance cost values were collected for every 1000 hour meter readings. For equipment class 240, maintenance cost data were collected from 1000 hr to 8000 hr (every 1000 hour value) for all of the equipment units so that data could be evenly distributed throughout the range of ages. Plotting the dataset for equipment class 240 and the corresponding regression analysis are shown in Figure 15. From this figure it can be found that the value of A is 1021.7 and B is 4342.1. The value of goodness of fit (R^2) is 0.9388, which indicates a good curve fit for this dataset.

Table 4: Sample data for LTD analysis (equipment class 240, graders (150 to 225 hp))

LTD		
Equipment Unit No	**Age or Hour Meter Reading (1000 hour)**	**LTD Maintenance Cost($)**
240-404	1.028	1263.975
240-404	2.115	29,127.00
240-404	3.159	50,493.00
240-406	1.049	2,436.00
240-406	2.024	13,203.00
240-406	3	19,347.00
240-407	1.162	5,301.00
240-407	2.019	15,657.00
240-407	3.134	30,196.50

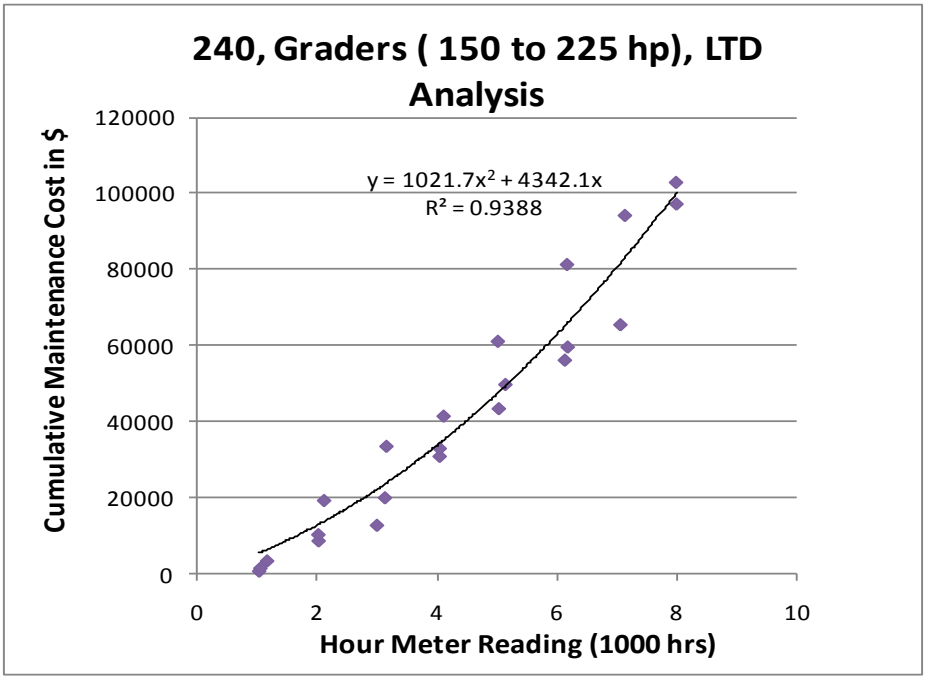


Figure 15: LTD analysis for equipment class 240, graders (150 to 225 hp)

In PCB regression analysis, an average maintenance cost value has to be calculated. For every piece of equipment in 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 5. Then these values were plotted with the linear regression analysis that is presented in Figure 16. A best-fit line was determined, but as the dataset for this PCB analysis is too small and scattered, it did not fit well. That's why the value of R^2 is only 0.5361. The values of C and D in Equation 7 were found in Figure 16 to calculate the values of A and B in Equation 5. From Figure 16 the values of C and D were found to be 2611 and 2797.2. Then, from Equation 8 and 9, the values of A and B are calculated. These are 2611 and 1398.6.

Table 5: Sample data for PCB analysis (equipment class 240, graders (150 to 225 hp))

PCB						
Equipment Unit No	**2009-2010 Maintenance Cost ($)**	**Start 2009 meter reading**	**End 2010 meter reading**	**Step**	**Midpoint (1000 hr)**	**2009-2010 Maintenance Cost($)/ 1000 hour**
240-404	44,961.00	5986	7845	1859	6.9155	24,185.58
240-405	89,481.00	10936	13102	2166	12.019	41,311.63
240-406	50,323.50	5150	7481	2331	6.3155	21,588.80
240-407	41,670.00	3134	5236	2102	4.185	19,823.97
240-481	54,099.00	11880	12625	745	12.2525	72,616.10
240-709	61,864.50	3924	5336	1412	4.63	43,813.38

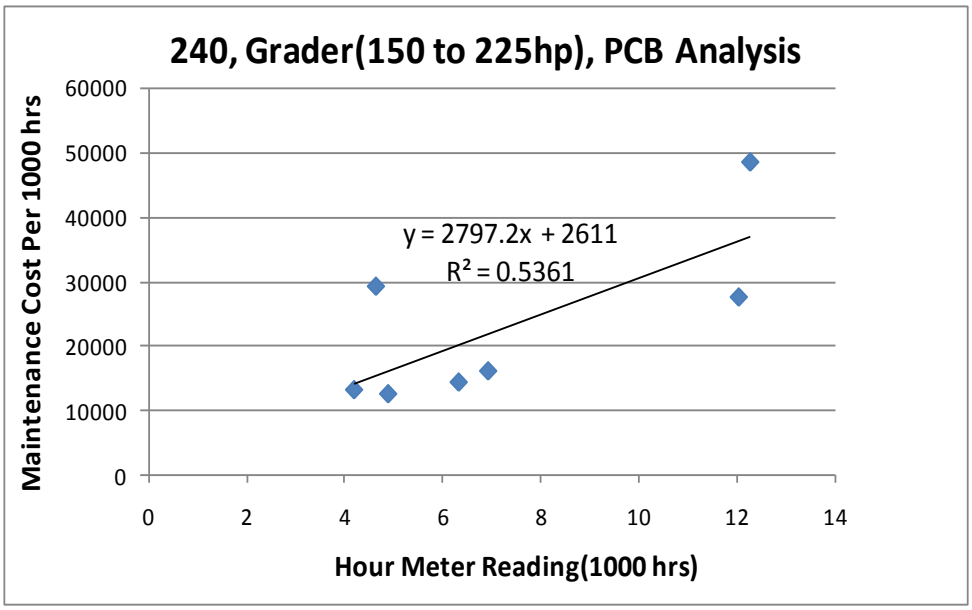


Figure 16: PCB analysis for equipment class 240, graders (150 to 225 hp)

After obtaining the values of A and B, equations from PCB and LTD analyses can be plotted in one figure. These two regression trend lines are compared visually in Figure 17. Although the two methods used different equipment unit's data, the cumulative cost equations for the two methods are quite similar; this is visible in Figure 17. From these two equations, cumulative maintenance costs can be predicted. As an example, to predict the cumulative maintenance cost for 16000

hours, the LTD and PCB equations provide values of $331,028.80 and $399,817.60, which are not too far from one another.

The trend line of PCB and LTD analyses and the comparison between them for equipment class 262, cement spreader and concrete paver, is presented in Appendix 2.

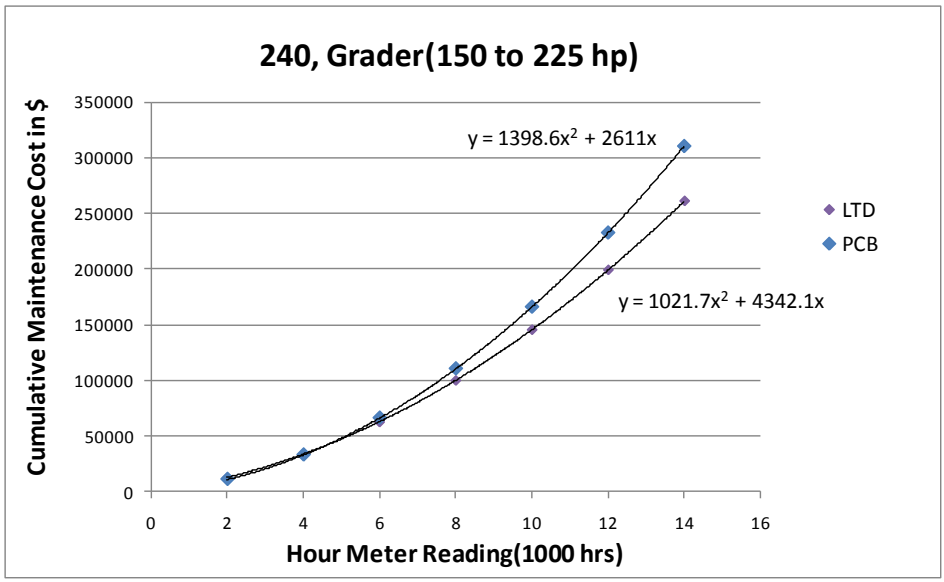


Figure 17: Comparison of LTD and PCB analysis for equipment class 240, graders (150 to 225 hp)

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 appropriate for this database because:

- Within most of the equipment classes very few equipment units have sufficient instances of data.

- Only a few equipment classes have more than 2-3 equipment units that have maintenance cost data from a zero hour meter reading.
- It was very hard to find any particular one or two year's 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 mandatory for PCB analysis.

4.5 Data Mining Analysis

During general trend analysis it was found that there are only 15 equipment classes between equipment class numbers 200 and 299 which have adequate realistic datasets for this data mining analysis.

Testing and Training Option

One of the steps in performing data mining analysis is to decide on data splitting for training and testing. Two common ways are percentage split and cross validation. Percentage split is for 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 for k times (Duan et al., 2003). The objective of this prediction analysis is to predict the upcoming year's maintenance cost, so there is no justification for using cross-validation, as it does not help to train and test data to predict for upcoming years. In cross validation there is no way to separate just the last one or two year's data for testing. On the other hand, with percentage split, one can test the model by only last year's data and train the model using the

previous year's data. So by using the percentage split option, a good prediction model for upcoming years could be built. In this analysis, for each class of equipment the last one or two years of data were taken for testing the model and the rest of the data were used for training the model.

Attributes for Analysis

Four attributes were used for predicting maintenance cost/hour. The attributes are

a) Manufacturer of the equipment (Manufacturer)

b) Working year of equipment (Year)

c) Hour meter reading of equipment (Hour Meter)

d) Purchase price of the equipment (Purchase Price)

In this data mining part, first the most frequently used second order nonlinear regression analysis was conducted using MS Excel. After that, seven other algorithms were analyzed using WEKA software.

4.5.1 Second Order Nonlinear Regression Analysis

In this section, the prediction analysis of equipment maintenance cost using a second order nonlinear regression analysis is illustrated. Here, as an example of a training set, the partial data of equipment class 222 (Wheel Loaders, 4cy) is presented in Table 6. Also last year's data as a testing set is shown in Table 7. Training data were used to build the second order polynomial equation by MS Excel graphical analysis (Figure 18) and then the equation was evaluated and validated by test data.

Table 6: Sample of training set data for equipment class 222, wheel loaders (4cy)

Manufacturer	Year	Hour Meter Reading	Purchase Price($)	Maintenance Cost ($)/Hour
Komatsu	2001	11093	127,500.00	9.54
Komatsu	2002	12081	127,500.00	12.75
Komatsu	2003	12694	127,500.00	47.14
Komatsu	2004	13840	127,500.00	33.60
Komatsu	2005	14958	127,500.00	37.65
CAT	2010	1507	203,087.43	8.04
CAT	2010	1713	203,087.43	13.28
CAT	2010	2168	203,087.43	12.19

Table 7: Testing set data for equipment class 222, wheel loaders (4cy)

Manufacturer	Year	Hour Meter Reading	Purchase Price ($)	Maintenance Cost($)/Hour (Actual)	Maintenance Cost($)/Hour (Predicted from equation)
Komatsu	2012	22374	127,500.00	37.82	42.39
CAT	2012	2744	304,631.14	9.55	13.09
CAT	2012	3644	304,631.14	17.55	15.45
CAT	2012	3897	304,631.14	16.71	16.09

From the equation generated in MS Excel, graphical analysis for equipment class 222 (wheel loaders, 4cy), the maintenance cost was predicted for last one year, which is presented in the last column of testing set (Table 7).

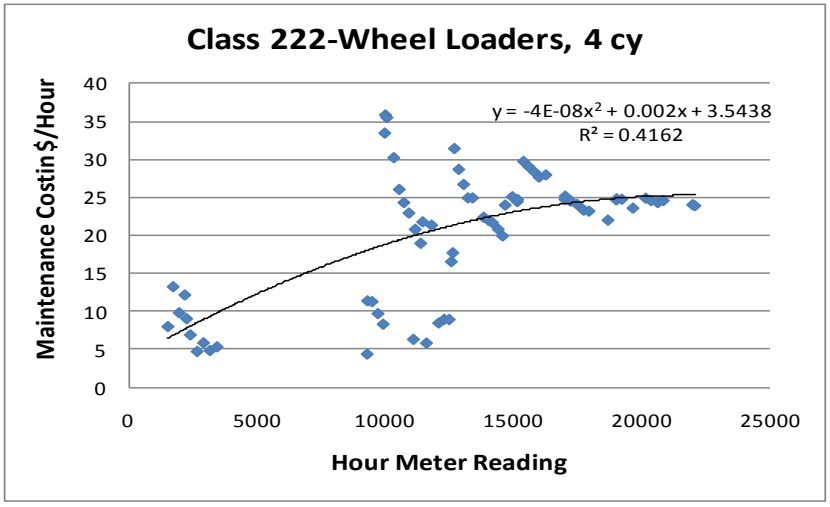


Figure 18: Second order nonlinear regression analysis by testing dataset for equipment class 222, wheel loaders (4cy)

There is also actual maintenance cost-per-hour data for each of the instances in Table 7. The graphical comparison of actual and predicted maintenance cost data is presented in Figure 19. Figure 19 shows that the predicted values of equipment class 222 are close to the actual value. So, by this graphical analysis, a decision can be taken that, for equipment class 222 the second order nonlinear regression equation may be used for prediction of the maintenance cost. All other graphs for second order nonlinear regression analysis are presented in Appendix 3.

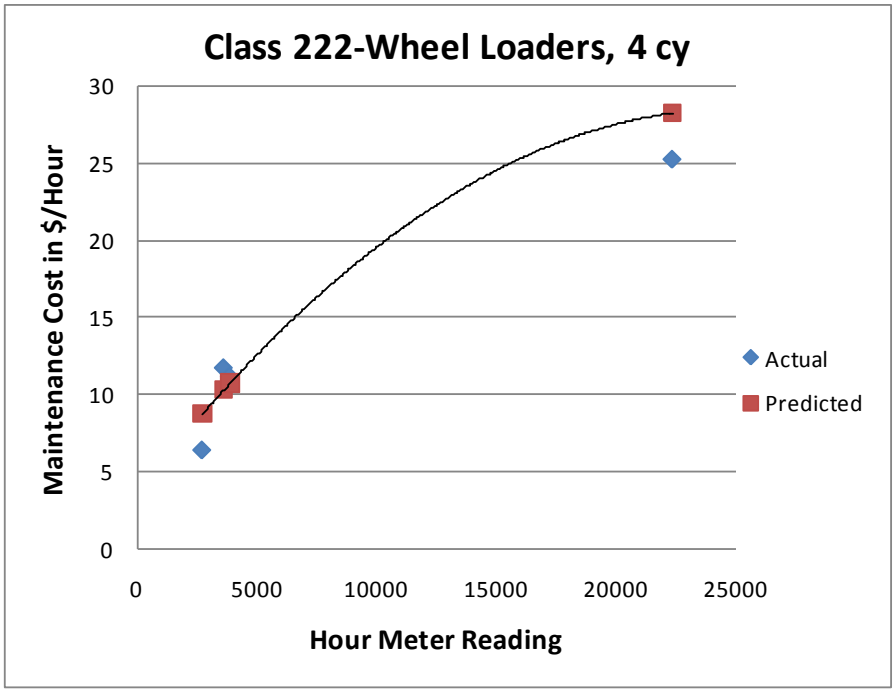


Figure 19: Comparison of actual value with prediction value for equipment class 222, wheel loaders (4cy)

From the actual and predicted values of maintenance cost per hour and by using the model evaluation and validation methods described in the literature review section 2.7, the following table, Table 8, is prepared for equipment class 222. For other equipment classes these calculations for second order nonlinear regression analysis are shown in Appendix 4 with other algorithms of WEKA analysis.

Table 8: Calculation of model evaluation and validation methods for second order nonlinear regression analysis of equipment class 222, wheel loaders (4cy)

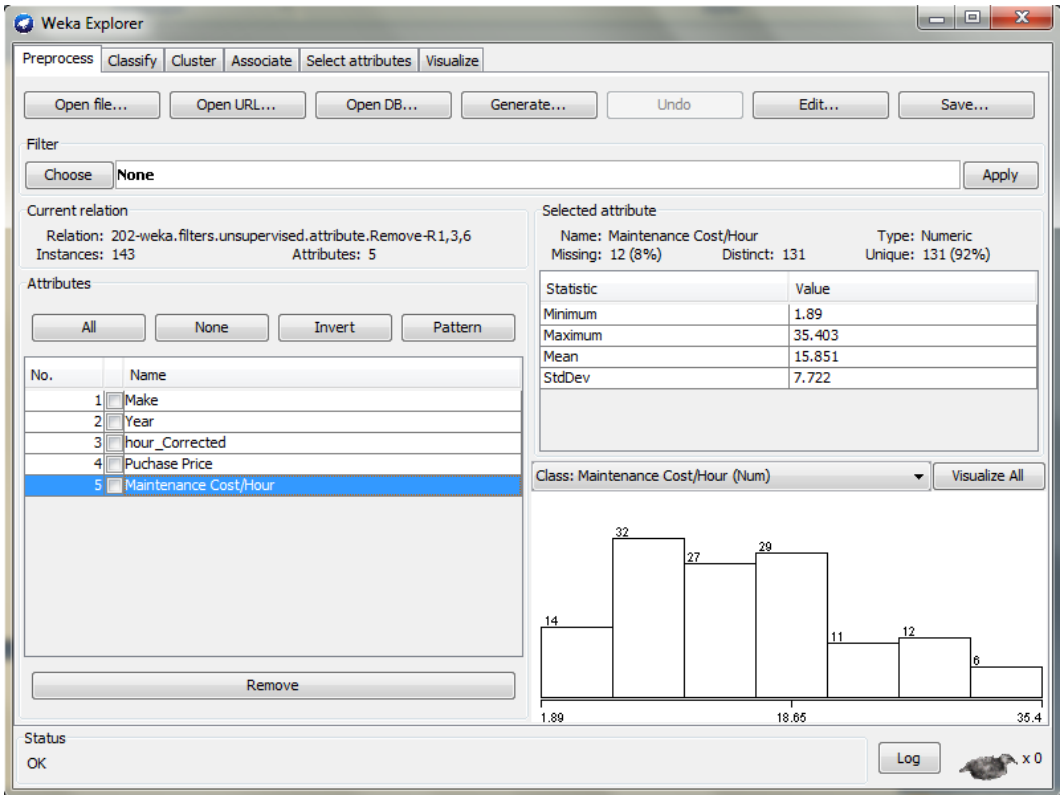
Methods	Values
Mean Absolute Error	1.80543
Root Mean Absolute Error	2.060584
Relative Absolute Error	31.10545
Root Relative Squared Error	29.37491
Correlation Coefficient	0.913711

4.5.2 WEKA Analysis

WEKA is machine learning software developed at the University of Waikato, New Zealand (Hall et al., 2009). It is popular software for data mining analysis. The main two user interfaces of WEKA are Experimenter and Explorer. A large number of algorithms exist in WEKA that can be used in Experimenter to make comparisons and determine the best algorithm for a particular type of dataset. Also, Explorer can build models for almost all of the algorithms. For this database Explorer and Experimental have been used to determine which algorithm is the best out of seven algorithms.

4.5.2.1 Explorer

In Explorer, after uploading a dataset a statistical summary can be visualized, this is effective to get a primary idea about the dataset (Figure 20). The top figure shows minimum, maximum, mean and standard deviation values of a dataset for an attribute. The bottom figure shows the bar chart of all the available attributes.



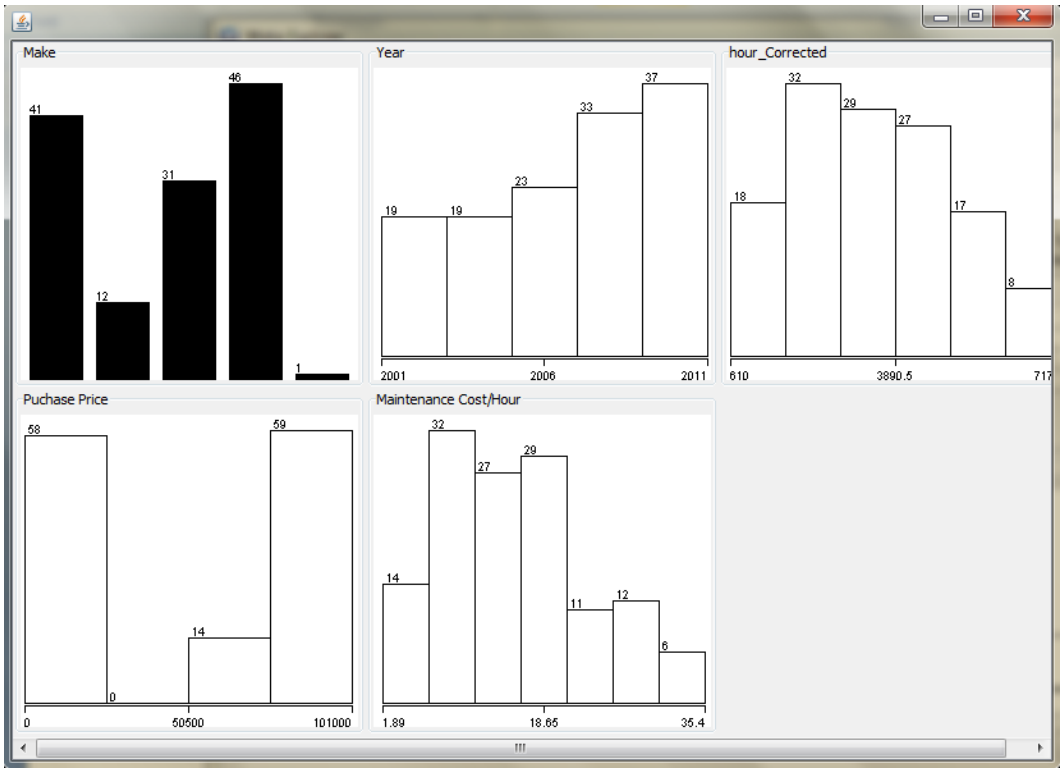


Figure 20: An example of a statistical output from the software WEKA

After loading the input data, the relationship of one attribute with other attributes can be visualized in the "Visualize" function, which is shown in Figure 21. It is possible to see all the relationships separately such as in Figure 22, where the maintenance cost/hour vs. hour meter reading is visualized for equipment class 213 (vibratory compactor (50+hp)).

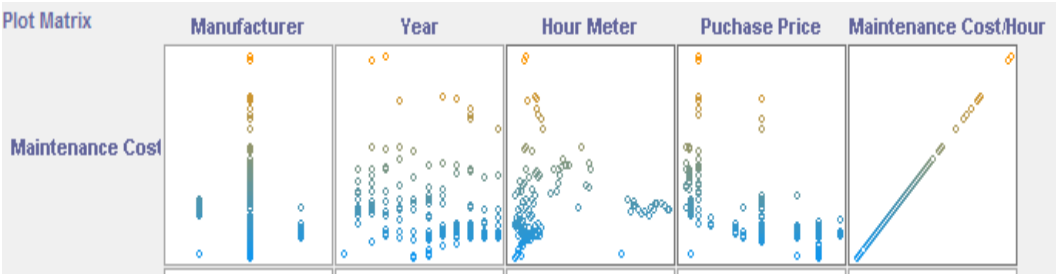


Figure 21: Relationship of maintenance cost with other attributes in the "Visualize" function for equipment class 213, vibratory compactor (50+hp)

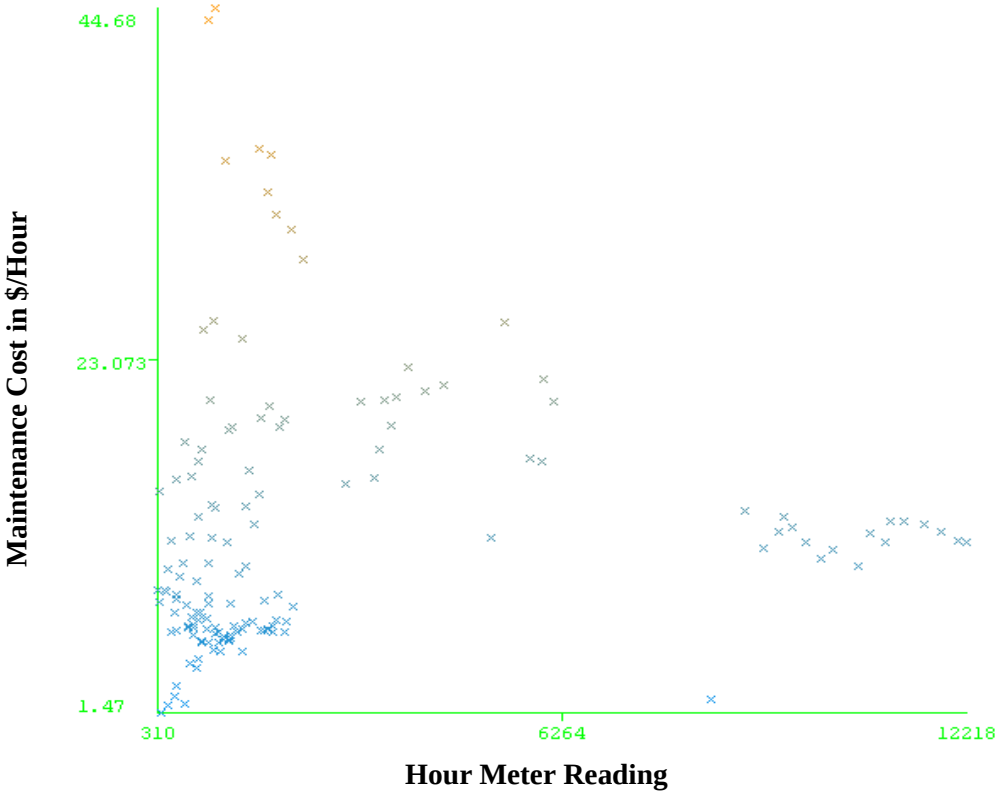


Figure 22: Relationship of maintenance cost to hour meter reading in "Visualize" function for equipment class 213, vibratory compactor (50+hp)

```
=== Run information ===

Scheme:weka.classifiers.functions.LeastMedSq -S 4 -G 0
Relation:     213
Instances:    143
Attributes:   5
              Manufacturer
              Year
              Hour Meter
              Puchase Price
              Maintenance Cost/Hour
Test mode:split 92.0% train, remainder test

=== Classifier model (full training set) ===


Linear Regression Model

Maintenance Cost/Hour =

     -6.1046 * Manufacturer=Ingersol,Dynapac +
      7.6675 * Manufacturer=Dynapac +
     -0.1188 * Year +
      0.0008 * Hour Meter +
     -0.0001 * Puchase Price +
    248.518

Time taken to build model: 0.27 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.6123
Mean absolute error                      3.2776
Root mean squared error                  5.7224
Relative absolute error                 52.7932 %
Root relative squared error             78.8396 %
Total Number of Instances               11
```

Figure 23: Explorer output of least median square classifier for equipment class 213, vibratory compactor (50+hp)

In the classifier of Explorer, seven algorithms were used to build the best model for each of the 15 equipment classes. Explorer provides correlation coefficient and different error calculations of each model by comparing testing data with the

built model. A sample of the Explorer output for the classifier is shown in Figure 23.

Classifier error can be visualized in the WEKA Explorer by comparing the actual and predicted maintenance cost/hour value which is shown in Figure 24. In Figure 24, one big cross is visible, which shows that the actual value is big while the predicted value is small, or vice versa.

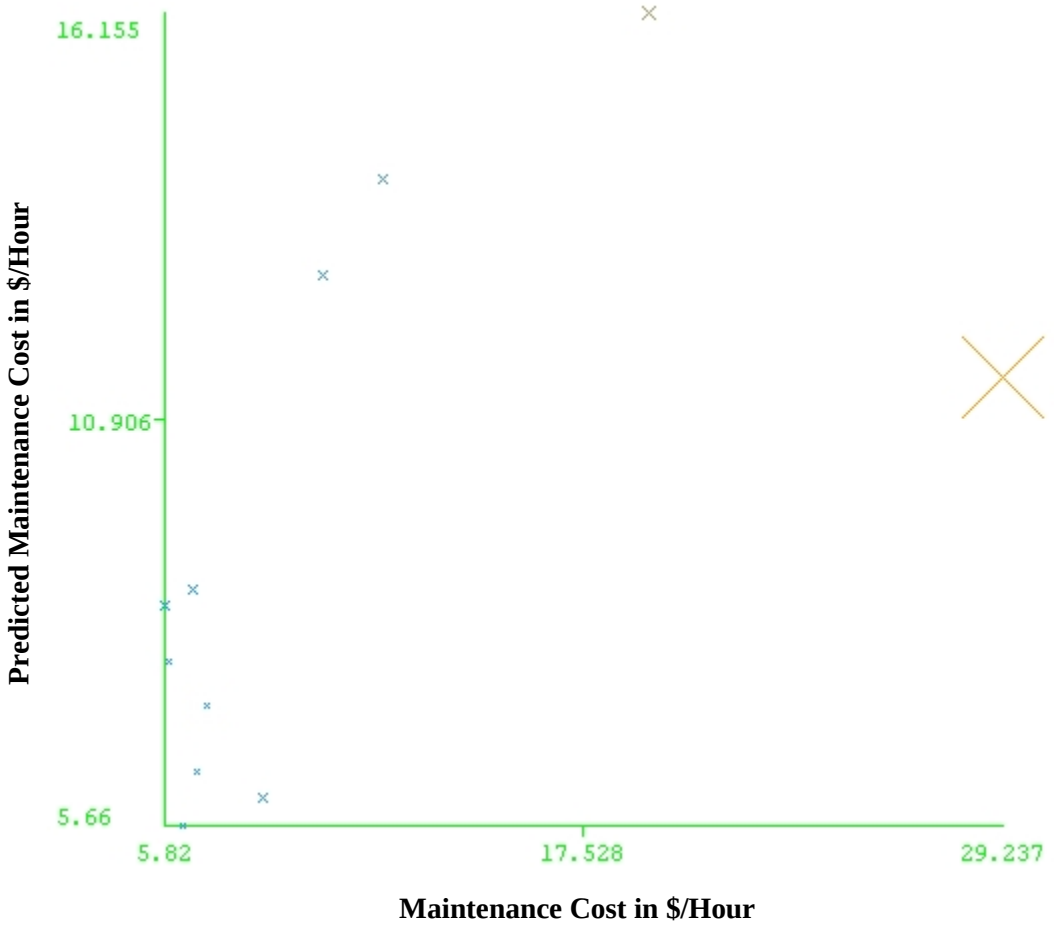


Figure 24: Error visualization of least median square algorithm for equipment class 213, vibratory compactor (50+hp)

For each of the seven algorithms, the root relative squared error was listed (Table 9) and compared to get the best algorithm for each of the available classes. Then this best model was used in Experimenter as the base model.

Table 9: Root relative squared error of equipment class 213, vibratory compactor (50+hp)

Algorithms	Root Relative Squared Error
Least Median Square	78.8
Linear Regression	80.8
Conjunctive Rule	94.2
Decision Stump	93.8
M5Rule	80.8
REP Tree	79.8
Multilayer Perceptron	64.37

4.5.2.2 Experimenter

The best algorithm determined in Explorer was used in experimenter as the base algorithm to compare it with other algorithms. Pared T-Tester was used to verify whether the base algorithm was the best one or whether there was another, better algorithm. From Table 9 it can be seen that multilayer perceptron has the least root relative squared error. Figure 25 shows that for equipment class 213 multilayer perceptron was taken as a base algorithm.

In Figure 25, "v" means victory of any classifier compared with the base one, and "*" means the classifier is worse than the base one. However, in this analysis none of the classifiers were found to be better or worse than the base one. This was shown for all of the available equipment classes. The size of the data set was too small for T-Test to work properly; T-Test requires a bigger data set. Hence, a decision about using an algorithm for a certain equipment class was made on the basis of the results from WEKA Explorer.

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Root_relative_squared_error
Datasets:   1
Resultsets: 7
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       07/11/13 2:47 PM


Dataset                 (1) function | (2) funct (3) rules (4) trees (5) rules (6) trees (7) funct
-----------------------------------------------------------------------------------------------------
213                      (1)   78.84 |   80.80     94.21     93.85     80.80     79.82     64.37
-----------------------------------------------------------------------------------------------------
                             (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) functions.LeastMedSq '-S 4 -G 0' 4288954049987652970
(2) functions.LinearRegression '-S 0 -R 1.0E-8' -3364580862046573747
(3) rules.ConjunctiveRule '-N 3 -M 2.0 -P -1 -S 1' -5938309903225087198
(4) trees.DecisionStump '' 1618384535950391
(5) rules.M5Rules '-M 4.0' -1746114858746563180
(6) trees.REPTree '-M 2 -V 0.0010 -N 3 -S 1 -L -1' -9216785998198681299
(7) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
```

Figure 25: Output of WEKA Experimenter for equipment class 213, vibratory compactor (50+hp)

4.6 Model Evaluation and Validation

The second order nonlinear regression analysis in MS Excel and seven other algorithms in WEKA were evaluated and validated by five different measures- mean absolute error, root mean squared error, relative absolute error, root relative squared error and correlation coefficient. For the equipment class, the last one or two years of data were taken as a testing set and the rest were taken as a training set. The main purpose of the evaluation was to find out which of the eight selected algorithms is best for an equipment class, but validation was used to attempt to

find an algorithm or algorithms that can be implemented for a certain equipment class.

4.6.1 Comparison of Different Algorithms

The errors and correlation coefficient have been compared for all eight algorithms for all available equipment classes, such as for class 213 the errors and correlation coefficients are presented in Table 10. From Table 10 it can be understood that multilayer perceptron is the best algorithm for this class. Also in this evaluation section an attempt was made to find which algorithm or algorithms are close to the best that can be used as a substitute for the best algorithm. For example, least median square can be used for equipment class 213 because it is close to the best algorithm.

Table 10: Comparison of correlation coefficient and different errors for equipment class 213, vibratory compactor (50+hp)

Algorithms	**Mean Absolute Error**	**Root Mean Squared Error**	**Relative Absolute Error**	**Root Relative Squared Error**	**Correlation Coefficient**
Second Order Nonlinear Regression	6.85	7.77	110.3	107.1	0.134
Least Median Square	**3.27**	**5.72**	**52.8**	**78.8**	**0.61**
Linear Regression	4.03	5.86	65.0	80.8	0.59
Conjunctive Rule	3.72	6.84	60.0	94.2	0.33
Decision Stump	3.76	6.81	60.5	93.8	0.33
M5Rule	4.04	5.86	65.0	80.8	0.59
REP Tree	3.96	5.79	63.7	79.8	0.57
Multilayer Perceptron	**3.17**	**4.6**	**51.08**	**64.37**	**0.75**

Table 11 and Table 12 show the best case as well as the worst case of the model evaluation and validation part. The best case is shown in Table 11, where for equipment class 222 (Wheel Loader, 4cy), almost all the algorithm's error percentage is very low and the correlation coefficient is almost 1. In Table 12, for equipment class 240 (grader (150 to 225 hp)), almost all types of error are more than 100 and the correlation coefficient is 0 or close to 0. From the comparison of errors in both Table 11 and Table 12 it can be seen that M5Rule is the best one. Though the M5Rule is best for equipment class 240 it should not be used for this equipment class as the percentage of error for both the relative absolute error and root relative squared error is close to 100%.

Table 11: Comparison of correlation coefficient and different errors for equipment class 222, wheel loader (4cy)

Algorithms	**Mean Absolute Error**	**Root Mean Squared Error**	**Relative Absolute Error**	**Root Relative Squared Error**	**Correlation Coefficient**
Second Order Nonlinear Regression	1.81	2.1	19.96	21.58	0.98
Least Median Square	2.42	2.67	26.8	28.1	0.96
Linear Regression	2.94	3.3	32.6	34.6	0.96
Conjunctive Rule	2.55	2.56	28.3	26.9	0.96
Decision Stump	2.5	2.56	27.6	26.9	0.96
M5Rule	**1.5**	**1.71**	**16.6**	**17.9**	**0.97**
REP Tree	2.22	2.41	24.6	25.3	0.96
Multilayer Perceptron	2.88	3.37	31.89	35.32	0.97

Table 12: Comparison of correlation coefficient and different errors for equipment class 240, grader (150 to 225 hp)

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	3.4	4.4	148.4	150.45	-0.45
Least Median square	3.31	4.52	143.58	154.6	-0.53
Linear Regression	3.85	4.58	166.96	156.5	-0.39
Conjunctive Rule	3.01	3.54	130.4	121.3	0
Decision Stump	2.84	3.38	123.4	115.6	0
M5Rule	**2.17**	**3.06**	**94.3**	**104.5**	**0.0324**
REP Tree	2.85	3.38	123.38	115.6	0
Multilayer Perceptron	7.01	9.08	212.36	215.47	0.25

5.6.2 Comparison of Cross Validation with Percentage Split

In the Cross Validation option, the total database is divided randomly into a number of folds, and then one fold is used for testing and the rest of the folds for training. This process is repeated for the same number of folds; in a case when the model is being trained for one time, that model will be able to see the data from the previous training time. Therefore, it could happen that for the equipment class which has a high error rate in the percentage split could become lower in the cross-validation, such as for equipment class 240 where all the errors are high and the correlation coefficient is low in the percentage split option (Table 12).

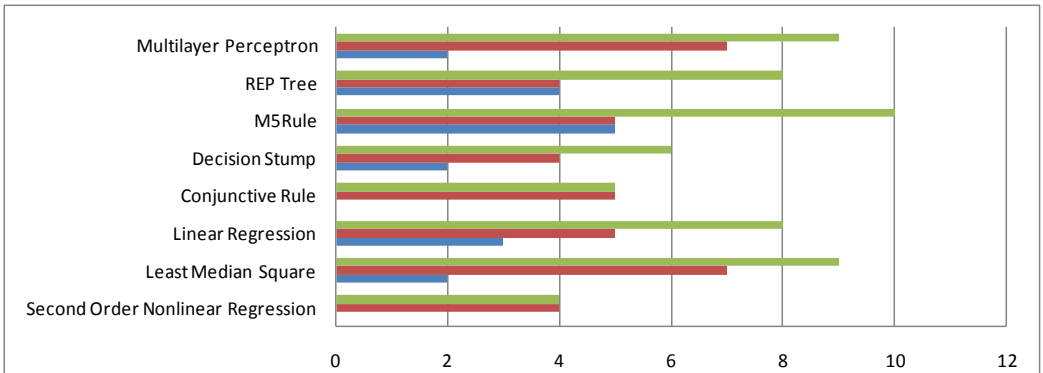
However, in the 10-fold cross validation option, almost all the errors became lower and the correlation coefficient became higher, as show in Table 13.

Table 13: Comparison of correlation coefficient and different errors for equipment class 240, grader (150 to 225 hp), by cross validation

Algorithms	**Mean Absolute Error**	**Root Mean Squared Error**	**Relative Absolute Error**	**Root Relative Squared Error**	**Correlation Coefficient**
Least Median Square	4.0	6.5	86.93	91.46	0.49
Linear Regression	3.77	5.24	81.87	73.69	0.67
Conjunctive Rule	3.91	5.14	84.95	72.32	0.69
Decision Stump	3.89	5.10	84.64	71.82	0.69
M5Rule	**2.64**	**4.28**	**57.37**	**60.31**	**0.80**
REP Tree	3.19	4.51	69.39	63.52	0.77
Multilayer Perceptron	3.6	5.04	78.15	70.96	0.72

5.6.3 Selection of Algorithms for Different Equipment Classes

Figure 26 shows the number of equipment classes in which an algorithm is best and the number of equipment classes in which the algorithm can be used as close to the best algorithm. As shown in Figure 26, the M5Rule is the best algorithm that fits most equipment classes and the second best algorithm is either the least median square or the multilayer perceptron. As the least median square and M5Rule are both regression analyses, these two algorithms can be used for most of the equipment classes.



	Second Order Nonlinear Regression	Least Median Square	Linear Regression	Conjunctive Rule	Decision Stump	M5Rule	REP Tree	Multilayer Perceptron
Total (out of 15)	4	9	8	5	6	10	8	9
Usable Algorithm	4	7	5	5	4	5	4	7
Best Algorithm	0	2	3	0	2	5	4	2

Figure 26: Number of times used as best or usable algorithm for 15 equipment classes

Though the M5Rule and least median square fit best for most of the equipment classes, linear regression and REP tree cannot be ignored, as there are certain equipment classes where there is no option but to use either of these two methods. As an example, for equipment class 253 neither the M5Rule nor the least median square algorithm can be used, because for both of the cases the relative absolute error and root relative squared error are close to 100%. However, REP tree can be used for this equipment class that possesses least relative absolute error and least root relative squared error which is shown in Table 14. A similar case is for equipment classes 220 and 262 where Linear Regression has to be used to predict the maintenance cost per hour (Table 15 and 16). Therefore, four algorithms have to be used for these 15 equipment classes to get better performance or better prediction. All of the comparisons of correlation coefficients and different errors for 15 of the concerned equipment classes are presented in Appendix 4.

Table 14: Comparison of correlation coefficient and different errors for equipment class 253, wheel tractors (backhoe)

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	5.05	6.0	124.5	121.26	-0.73
Least Median Square	6.52	9.01	160.5	162.1	-0.37
Linear Regression	4.82	6.09	118.68	123.17	0.29
Conjunctive Rule	3.31	4.28	81.61	86.5	0
Decision Stump	3.28	4.22	80.9	85.25	0
M5Rule	6.96	7.62	171.43	154.1	-0.29
REP Tree	**1.51**	**1.66**	**37.35**	**33.5**	**0.97**
Multilayer Perception	6.35	6.92	156.51	139.86	0.76

Table 15: Comparison of correlation coefficient and different errors for equipment class 220, wheel loaders (1 to 2 cy)

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	6.31	8.41	130.88	146.2	0.92
Least Median Square	5.96	7.21	123.7	125.3	0.69
Linear Regression	**2.06**	**3.05**	**42.8**	**53.1**	**0.89**
Conjunctive Rule	2.04	2.31	42.5	40.2	0.92
Decision Stump	1.97	2.31	40.94	40.27	0.92
M5Rule	4.68	5.41	97.2	94.1	0.99
REP Tree	4.16	4.53	86.4	78.8	0.99
Multilayer Perception	4.69	5.07	97.35	88.13	0.69

Table 16: Comparison of correlation coefficient and different errors for equipment class 262, cement spreader and concrete paver

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	5.19	5.94	72.59	64.92	0.76
Least Median Square	21.37	30.35	299.2	331.7	-0.29
Linear Regression	**4.21**	**4.79**	**58.9**	**52.4**	**0.83**
Conjunctive Rule	7.23	8.23	101.25	89.97	0.6
Decision Stump	7.07	8.42	99.07	92.01	0.61
M5Rule	10.48	14.59	146.81	159.47	0.63
REP Tree	5.82	6.59	81.56	71.98	0.79
Multilayer Perception	4.49	5.17	62.86	56.5	0.85

Table 17 has been prepared from these discussions and on the basis of different types of errors and correlation coefficients to show which algorithm is applicable for which equipment classes. There are some equipment classes, such as equipment classes 240 and 256, for which none of these eight algorithms can be used, because the errors are close to 100%. Also, Table 18 shows a summary of different types of errors and correlation coefficients for these best applicable algorithms of all the equipment classes.

Table 17: Suggested algorithms for different equipment classes

No	Algorithms	Equipment Class Number[1]
1	M5Rule	219, 222, 223, 243,265
2	Least Median Square	202, 205, 213,216,217
3	Linear Regression	220,262
4	REP Tree	253
5	None of These Eight Algorithms	240, 256

[1]Name of the equipment classes with respect to class number is provided in Appendix no. 5

Table 18: Summary of correlation coefficient and different errors for the suggested algorithms of each of the equipment classes

Equipment Class No	Suggested Algorithms	Mean Absolute	Root Mean Squared	Relative Absolute	Root Relative Squared Error	Correlation Coefficient
Class 202	Least Median Square	1.72	2.12	17.5	21.4	0.82
Class 205	Least Median Square	1.55	1.61	47.3	35.6	0.9
Class 213	Least Median Square	3.27	5.72	52.8	78.8	0.61
Class 216	Least Median Square	3.35	4.61	29.5	37.5	0.66
Class 217	Least Median Square	4.78	5.10	47.6	48.6	0.86
Class 219	M5Rule	2.82	3.49	64.3	68.8	0.97
Class 220	Linear Regression	2.06	3.05	42.8	53.1	0.89
Class 222	M5Rule	1.5	1.71	16.6	17.9	0.97
Class 223	M5Rule	5.75	6.1	42.9	42.6	0.95
Class 240	None	-	-	-	-	-
Class 243	M5Rule	8.46	11.09	67.16	69.9	0.29
Class 253	REP Tree	1.51	1.66	37.35	33.5	0.97
Class 256	None	-	-	-	-	-
Class 262	Linear Regression	4.21	4.79	58.9	52.4	0.83
Class 265	M5Rule	3.26	4.11	48.9	49.9	0.96

4.6.4 Model Building from the Selected Algorithms

From the chosen algorithms, models for each of the equipment classes were constructed from Explorer output; for example, for equipment class 253, REP tree is the best algorithm. The model is presented in Figure 27.

```
REPTree
============

Make = Ford : 27.42 (6/63.23) [2/317.29]
Make = Not found
|   hour_Corrected < 1123 : 2.49 (2/0.89) [1/0.11]
|   hour_Corrected >= 1123 : 9.48 (2/2.37) [1/0.36]
Make = JD
|   Year < 2007.5
|   |   Year < 2006.5
|   |   |   hour_Corrected < 2617 : 4.85 (5/0.14) [5/2.28]
|   |   |   hour_Corrected >= 2617 : 5.76 (6/0.06) [0/0]
|   |   Year >= 2006.5 : 8.52 (4/0.01) [3/52.91]
|   Year >= 2007.5
|   |   Puchase Price < 89215
|   |   |   Puchase Price < 77000 : 12.8 (5/8.29) [4/10.01]
|   |   |   Puchase Price >= 77000
|   |   |   |   hour_Corrected < 2915.5 : 3.88 (2/0.19) [2/2.43]
|   |   |   |   hour_Corrected >= 2915.5
|   |   |   |   |   hour_Corrected < 5370 : 7.27 (3/0.45) [2/0.19]
|   |   |   |   |   hour_Corrected >= 5370 : 10.73 (2/0.06) [2/0.31]
|   |   Puchase Price >= 89215 : 16.8 (11/2.92) [3/2.66]

Size of the tree : 20
```

Figure 27: Model for equipment class 253, wheel tractors (backhoe), from the analysis of the REP tree algorithm (WEKA output)

Similarly, for equipment class 219, the M5Rule is the best algorithm and the model is shown in Figure 28. Also the best algorithm for equipment class 213 is the least median square and the model from this algorithm is shown in Figure 29.

```
=== Classifier model (full training set) ===

M5 pruned model rules
(using smoothed linear models) :
Number of Rules : 3

Rule: 1
IF
        Make=Case > 0.5
        Year > 2007.5
THEN

Maintenance Cost/Hour =
        4.3291 * Make=Case
        + 1.8869 * Year
        - 0.0046 * hour_Corrected
        + 0 * Puchase Price
        - 3766.1173 [30/51.002%]

Rule: 2
IF
        Year <= 2010.5
        Make=Case <= 0.5
THEN

Maintenance Cost/Hour =
        2.0864 * Make=Case
        + 0.7948 * Year
        - 0.0004 * hour_Corrected
        - 1590.1416 [25/29.843%]

Rule: 3

Maintenance Cost/Hour =
        11.9204 * Make=Case
        + 1.5501 * Year
        - 0.002 * hour_Corrected
        + 0.0003 * Puchase Price
        - 3110.3841 [32/83.784%]
```

Figure 28: Model for equipment class 219, wheel loaders (0 to 1 cy), from the analysis of M5Rule (WEKA output)

```
Maintenance Cost/Hour =

     -6.1046 * Manufacturer=Ingersol,Dynapac +
      7.6675 * Manufacturer=Dynapac +
     -0.1188 * Year +
      0.0008 * Hour Meter +
     -0.0001 * Puchase Price +
    248.518
```

Figure 29: Model for equipment class 213, vibratory compactor (50+hp), from the analysis of the least median dquare algorithm (WEKA output)

4.7 Conclusion

In this chapter the total analysis of this research has been elaborately described. A basic concept was constructed about the maintenance cost of different equipment classes by trend analysis. After that, CCM was conducted but due to a lack of an adequate amount of data, CCM did not become applicable. However, data mining analysis was a good option for this database. Eight algorithms were chosen and compared to each other to find the best one for each of the available 15 equipment classes. Although attempts were made to find only one algorithm for all of the equipment classes, analysis using WEKA found that a minimum of four algorithms have to be used for better prediction of equipment maintenance costs. It is hoped that these results will be beneficial for an equipment manager at Standard General Inc. and will make it possible for the company to make any maintenance-related decisions for these equipment classes.

Chapter 5: Conclusions

5.1 Research Summary

This research was motivated to improve the equipment management system, specifically in the area of equipment maintenance. There are many different sectors in the equipment maintenance system that could be improved with proper attention. However, on the basis of Standard General Inc.'s current needs and the availability of the database, this research focused on the systematic approach to predicting maintenance cost for upcoming years. The main objective of this research was finding models/algorithms that could be used to predict maintenance cost for different equipment classes and, overall, to propose a systematic method which could be followed by future researchers or equipment managers to predict maintenance cost.

This research was initiated with an understanding about the equipment maintenance systems and different costs for equipment maintenance for different companies. Then the researcher focused on the literature about the forecasting system of maintenance cost, data mining systems, the algorithms used for data mining, and model evaluation and validation systems described in Chapter 2.

In the preliminary stage, the researcher became familiar with the M-track software system to store maintenance-related data and how these data can be collected through MS SQL server. This data collection and preparation stage consumed a large portion of time in this research. Problems related to hour meter reading were faced and then solved. The work related to the data collection and preparations was presented in Chapter 3.

After preparing the dataset about the maintenance cost of all the available equipment classes, a trend analysis of the maintenance cost was conducted. The main purpose of this trend analysis was to explore the behaviour of maintenance cost trends. For prediction analysis cumulative cost modeling (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 a lack of a sufficient amount of data instances in almost all the equipment classes, CCM is not feasible for this kind of database. In the data mining part, first the common second order nonlinear regression analysis was conducted in MS Excel. As it is a common and popular algorithm, it was compared with other WEKA algorithms. For data mining analysis, 15 equipment classes were chosen between equipment class numbers 200 and 299, who have sufficient amount of maintenance related data. In WEKA, seven algorithms were chosen for data mining analysis. After evaluating and validating these eight algorithms (including the second order nonlinear regression analysis), it was found that the M5Rule and the least median square are the two algorithms that should be used for almost all of the concerned equipment classes, but for two equipment classes linear regression and for one equipment class REP tree should be implemented. All of these analyses and the results were elaborated in Chapter 4.

5.2 Research Contributions

This research work has a number of contributions to academic areas and the equipment management field. The main contributions are discussed below:

- A systematic procedure has been proposed here about how the raw data should be collected, processed and prepared for research analysis.
- The trend analysis of maintenance cost that is described in this thesis will be helpful for the concerned company's equipment managers to get a basic idea of how the age of equipment affects maintenance cost. According to the findings from the literature, maintenance cost should increase with the age of equipment, but from the trend analysis it was found that for four out of 15 equipment classes, initially there is downward trend of equipment maintenance cost. After this trend analysis and from discussions with an expert in this field, it was found that equipment maintenance cost could fluctuate with the equipment ages.
- Because of an inadequate amount of data in this assigned database, CCM was not a useful option for predicting maintenance cost. However, the process to conduct CCM by two approaches (LTD and PCB) has been elaborately described. When there is a sufficient amount of data for different equipment classes, the same process can be utilized and at that time it can be a very useful tool for predicting maintenance cost.
- From data mining analysis, four algorithms (M5Rule, least median square, linear regression and REP tree) were found to build better models for predicting maintenance cost for 13 concerned equipment classes. These findings can help the company's equipment managers in estimating equipment maintenance cost for upcoming years.

5.3. Research Limitations

This research work has some limitations which are listed below:

- The hour meter reading was not available for almost all the maintenance cost data, but for this research work it was essential that the hour meter reading be available with all the instances of maintenance cost data. With the help of the preventive maintenance dataset where the hour meter reading was available, a database was prepared where all the maintenance cost data could be available with the corresponding hour meter reading. However, there could be some problems related to the precision of maintenance cost value due to the pre-processing of original database.
- Equipment maintenance cost has been predicted here on the basis of different attributes that were found in the M-Track maintenance work database. In the current M-Track database there is no data about the equipment working hour in the field, the ideal time of equipment and the weather condition at the field. Better prediction of maintenance cost is possible if those equipment field operation data can be utilized in data mining analysis.
- Due to limited instances of maintenance cost data for most of the equipment units within every equipment class, CCM was conducted for only two equipment classes. The analysis would be more useful if many pieces of equipment could have sufficient instances of equipment maintenance cost data.
- Data mining analysis was conducted only for 15 equipment classes of equipment between equipment class numbers 200 and 299. Due to an

insufficient amount of data, this analysis could not be done for rest of the equipment classes.

- This research work was conducted on maintenance cost data from Standard General Inc., so the research output does not represent all construction equipment in general.

5.4. Recommendations for Future Work

There are some recommendations for future researchers in the same arena. These are listed below:

- In this research work, the database was imported to MS Access and then processed in both MS Access and MS Excel using many queries and functions. However, in the future it will be better if any automatic system can be generated using the same queries, functions and steps that were taken in this research.
- Trend analysis was done in the research work to see the behaviour of maintenance cost with respect to equipment age. This is the first step in equipment replacement analysis, so replacement analysis could be done by utilizing these trend analyses in future research work.
- The general trend analysis was done in this research by grouping equipment into class rather than single unit of equipment. In future a comparative study can be done between group of equipment and equipment units. That comparative analysis will justify which method is better for general trend analysis.

- Equipment field operation data could not be used in this research work. In the future an attempt should be made to obtain this data, and then many attributes from field operational databases could be utilized in data mining analysis to better predict maintenance cost.
- In data mining analysis, eight algorithms were compared to obtain a better algorithm for each of the concerned equipment classes. However, there are many other algorithms that could be used in future research to find more accurate models.

References

Ai, I. W., & Langla, P. (1992). Introduction of one level decision tree. *Proceedings of the Ninth International Conference on Machine Learning*. Morgan Kaufmann. Retrieved 02 September, 2013 from http://citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.23.2878&type=ab

Ali, S., Tickle, K., & Pang, B. (2008). Rule based base classifier selection for bagging algorithm. *DMIN08 - International Conference on Data Mining,* 14-17 July, 2008, Las Vegas, USA, pp. 26-29.

Ankerst, M., Breunig, M. M., Kriegel, H. -., & Sander, J. (1999). In Delis A., Faloutsos C. and Chandeharizadeh S.(Eds.), *OPTICS: Ordering points to identify the clustering structure* United States, ACM ASSOCIATION FOR COMPUTING MACHINERY.

Antonie, M.L., Zaiane, .R. and Coman, A. (2001). Application of data mining techniques for medical image classification. *Proceedings of the Second International Workshop on Multimedia Data Mining, in conjunction with ACM SIGKDD conference*, San Francisco, USA.

Bayzid, S.M., Al-Hussain, M., & Mohamed, Y. (2013). Modeling the trend of maintenance cost for road construction equipment. CON-102, *4th Construction Specialty Conference, CSCE 2013 Annual Conference*.

Blaxton, A.C., Fay, M.J., Hansen, C.M., and Zuchristian, C.M. (2003). *An Analysis of USMC Heavy Construction Equipment (HCE) Requirements*. Thesis, presented to Naval Post Graduate School, at Montercy, CA.

Campbell, G.E., & Bolton, A.E. (2005). HBR validation: interpreting lessons learned from multiple academic disciplines, applied communities, and the AMBR project. In K.A. Gluck and R.W. Pew (Eds.) *Modeling human behavior with*

integrated cognitive architectures: Comparison, evaluation and validation, (pp. 365-395). New Jersey: Lawrence Erlbaum & Associates.

Caterpillar Performance Handbook. (1995). Caterpillar, Inc., Peoria, IL.

Cox, E. A. (1971). Equipment economics. *Handbook of Heavy Construction*, J. A. Havers and F. W. Stubbs, eds., McGraw-Hill, New York, NY.

Crossman, A. (n.d). Linear regression analysis. About.com.Sociology, Retrieved December 05, 2013 from http://sociology.about.com/od/Statistics/a/Linear-Regression-Analysis.htm

Decision stump (2012) In Wikipedia: The free encyclopaedia. Retrieved Sept 08, 2013, from http://en.wikipedia.org/wiki/Decision_stump

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, (2), 113

Devasena, C.L., Sumathi, T., Gomathi, V.V., & Hemalatha, M. (2011). Effectiveness evaluation of rule based classifiers for the classification of iris data set. *Bonfring International Journal of Man Machine Interface*, Vol. 1, Special Issue.

Drinking Water Source Protection (2013). Glossary. Retrieved 09 September 2013 from http://www.sourcewaterinfo.on.ca/content/spProject/glossary.php

Duan, K., Keerthi, S. S., and Poo, A. N. (2003). *Evaluation of simple performance measures for tuning SVM hyperparameters.* Neurocomputing, 51 41-59.

Fan, H. (1997). *Leveraging operational data for intelligent decision support in construction engineering management.* PhD thesis, University Of Alberta, Edmonton, Alberta.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases.* Al Magazine, 17(3), 37.

Foss, A., & Zaiane, O. R. (2002). A parameterless method for efficiently discovering clusters of arbitrary shape in large datasets. Paper presented at the *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, 179-186.

GeneXpro tools (n.d.). Analyzing GeneXpro tools models statistically: Relative absolute error. Retrieved August 21, 2013 from http://www.gepsoft.com/gxpt4kb/Chapter10/Section2/SS15.htm

Gonzalez-Villalobos, C. V. (2011). *Analysis of industrial construction activities using knowledge discovery techniques.* M.Sc. Thesis, University of Alberta, Edmonton, AB.

Gore, B. F. (2010). *Validating integrated human performance models involving time-critical complex systems.* Ph.D. Thesis, University of Toronto, Toronto, ON.

Government of Alberta. (2013). *Roads and highways.* Retrieved 2nd September 2013 from http://www.albertacanada.com/business/overview/roads-and-highways.aspx

Halpin, D. W., & Senior, B. A. (2011). *Construction management / Daniel W. Halpin, Bolivar A. Senior*. Hoboken, NJ : Wiley, c2011.

Hammad, A. M. (2009). *An integrated framework for managing labour resources data in industrial construction projects: A knowledge discovery in data (KDD) approach.* Ph.D. Thesis, University of Alberta, Edmonton, AB.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009), The WEKA Data Mining Software: An Update, SIGKDD Explorations, Volume 11, Issue 1

Higgins, J. (2006). *The Radical Statistician: A Beginners Guide to Unleashing the Power of Applied Statistics in The Real World* (5th Ed.). Jim Higgins Publishing

Holmes, G., Hall, M., & Frank, E. (1999). *Generating rule sets from model trees*. Advanced Topics in Artificial Intelligence, 1747 1-12.

Huang, C., Chen, M., and Wang, C. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications,* (4), 847

IBM (2011). Non-linear regression. IBM SPSS Statistics. Retrieved September 1, 2013 from http://pic.dhe.ibm.com/infocenter/spssstat/v20r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Fidh_nlre.htm

Kim, Y. H. (1989). *A forecasting methodology for maintenance cost of long-life equipment.* Ph.D. Thesis, The University of Alabama, Alabama, US.

Kumar, C. (2013). *Estimation and planning methodology for industrial construction scaffolding*. M.Sc. Thesis, University of Alberta, Edmonton, AB.

Liao, C. W., & Perng, Y. H. (2008). Data mining for occupational injuries in the taiwan construction industry. *Safety Science*, 46(7), 1091-1102.

Ling, Y., & Mahadevan, S. (2013). Quantitative model validation techniques: New insights. *Reliability Engineering and System Safety,* 111 (2013), 217–231

Makridakis, S. G., & Wheelwright, S. C. (1989). *Forecasting methods for management / spyros makridakis, steven C. wheelwright* New York : Wiley, c1989; 5th ed.

Mitchell, Z. W.,Jr. (1998). *A statistical analysis of construction equipment repair costs using field data and the cumulative cost model.* Ph.D. Thesis, Virginia Polytechnic Institute and State University, Virginia, US.

Mitchell, Z., Hildreth, J., & Vorster, M. (2011). Using the cumulative cost model to forecast equipment repair costs: Two different methodologies. *Journal of Construction Engineering & Management, 137*(10), 817-822.

Mount, D.M., Netanyahu, N.S., Romanik, K., Silverman, R., & Wue, A.Y. (2007). A practical approximation algorithm for the LMS line estimator. *Computational Statistics & Data Analysis*, (5), 2461.

Nakajima, S. (1988). *Introduction to TPM : Total productive maintenance*. Cambridge, Mass. [u.a.].

Nakajima, S. (1989). *TPM Development Program: Implementing Total Productive Maintenance*. Tokyo: Japan Institute for Plant Maintenance

Nichols, H. L. (1976). *Moving the Earth*. North Castle Books, Greenwich, CT.

Noriega, L. (2005). Multilayer perceptron tutorial. School of Computing. Staffordshire University. Retrieved 09 September 2013 from http://www.cs.sun.ac.za/~kroon/courses/machine_learning/lecture5/mlp.pdf

Nunnally, S. W. (2000). *Managing Construction Equipment*. Prentice-Hall, Englewood Cliffs, NJ.

OBERKAMPF, W. L., & TRUCANO, T. G. (2002). *Verification and validation in computational fluid dynamics.*

Oracle (2008). About classification. Oracle® Data Mining Concepts 11*g* Release 1(11.1), Retrieved 27 August 2013 from http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#i1005746

Panagiotidou, S., & Tagaras, G. (2007). Optimal preventive maintenance for equipment with two quality states and general failure time distributions. *European Journal Of Operational Research*, *180*(1), 329-353. doi:10.1016/j.ejor.2006.04.014.

Panchal, G., Ganatra, A., Kosta, Y. P., & Panchal, D. (2011). Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. *International Journal of Computer Theory and Engineering*, 3(2), 332-337.

Park, J., Hsiao-Rong Tyan, & Kuo, C. -. J. (2006). GA-based internet traffic classification technique for QoS provisioning. *2006 International Conference on Intelligent Information Hiding & Multimedia*, 251.

Peng, K. (2012). *Equipment management in the post-maintenance era [electronic resource] : A new alternative to total productive maintenance (TPM) / kern peng* Boca Raton, Fla. : CRC Press, 2012.

Poveda, C. A. (2008). *Predicting and evaluating construction trades foremen performance: Fuzzy logic approach.* M.Sc. Thesis, University of Alberta, Edmonton, AB.

Peurifoy, R. L., & Schexnayder, C. J. (2002). *Construction Planning, Equipment,and Methods, 6th ed.* McGraw-Hill, New York, NY.

Quinlan, J. (1992). Learning with continuous classes. *Proceedings AI'92, Singapore: World Scientific*, 343-348.

Rapp, R. & George, B. (1998). Maintenance management concepts in construction equipment curricula. *Journal of Construction Education Summer 1998*, 3(2), 102-117.

Sahoo, S., & Jha, M. K. (2013). Groundwater-level prediction using multiple linear regression and artificial neural network techniques: A comparative assessment. *Hydrogeology Journal*, (8), 1865.

Sammut, C. (2011). *Encyclopedia of machine learning [electronic resource]*. New York: Springer, c2011.

Saraee, M. M., Waheed, A. A., Javed, S. S., & Nigam, J. J. (2005). Application of Data Mining in Medical Domain: Case of Cardiology Sickness Level. *Metmbs - International Conference*, 337-340.

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, 7(1), 12-24.

Sheu, C., & Krajewski, L. J. (1994). A decision model for corrective maintenance management. *International Journal of Production Research, 32*(6), 1365.

Soibelman, L., & Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering, 16*(1), 39-48.

Tan, H. C., Carrillo, P. M., Anumba, C. J., Bouchlaghem, N., Kamara, J. M., & Udeaja, C. E. (2007). Development of a methodology for live capture and reuse of project knowledge in construction.(author abstract). *Journal of Management in Engineering*, (1), 18.

Trochim, W.M.K. (2006). Introduction to Evaluation. Research Methods, Knowledge Base. Retrieved 09 September 2013 from http://www.socialresearchmethods.net/kb/intreval.php

Viet, A., Fourichon, C., Jacob, C., Guihenneuc-Jouyaux, C., & Seegers, H. (2006). Approach for qualitative validation using aggregated data for a stochastic simulation model of the spread of the bovine viral-diarrhoea virus in a dairy cattle herd. *Acta Biotheoretica, 54*(3), 207-217.

Vorster, M. (1980). *A systems approach to the management of civil engineering construction equipment*. Ph.D. Thesis, University of Stellenbosch, Stellenbosch, South Africa.

Vorster, M (2009). *Construction equipment economics.* Pen, Christiansburg, VA.

Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.Working paper, Hamilton, New Zealand: University of Waikato, Department of Computer Science.

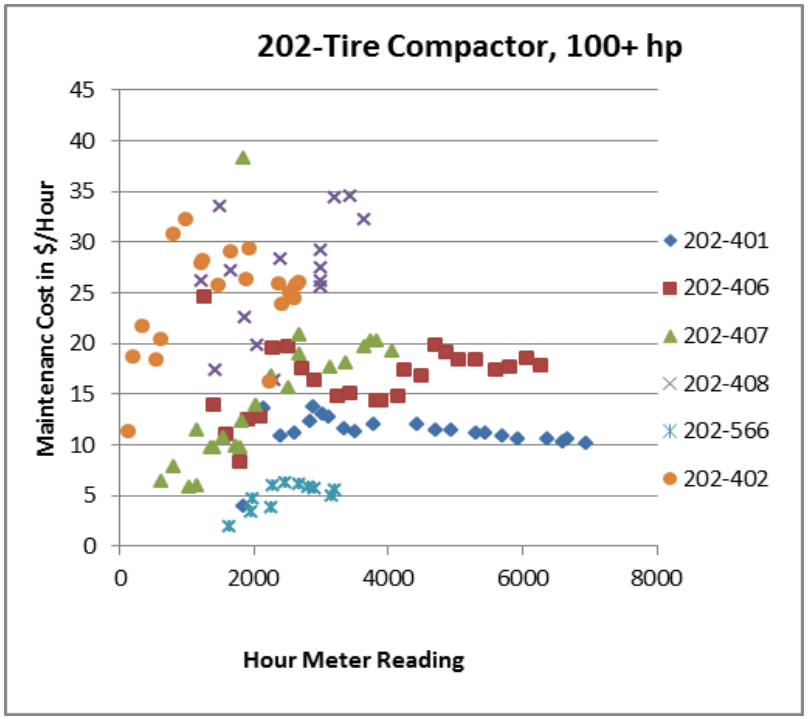
Weisstein, Eric W. (n.d.). Mean-Value Theorem. From *MathWorld*--A Wolfram Web Resource. Retrieved 18 august 2013 from http://mathworld.wolfram.com/Mean-ValueTheorem.html

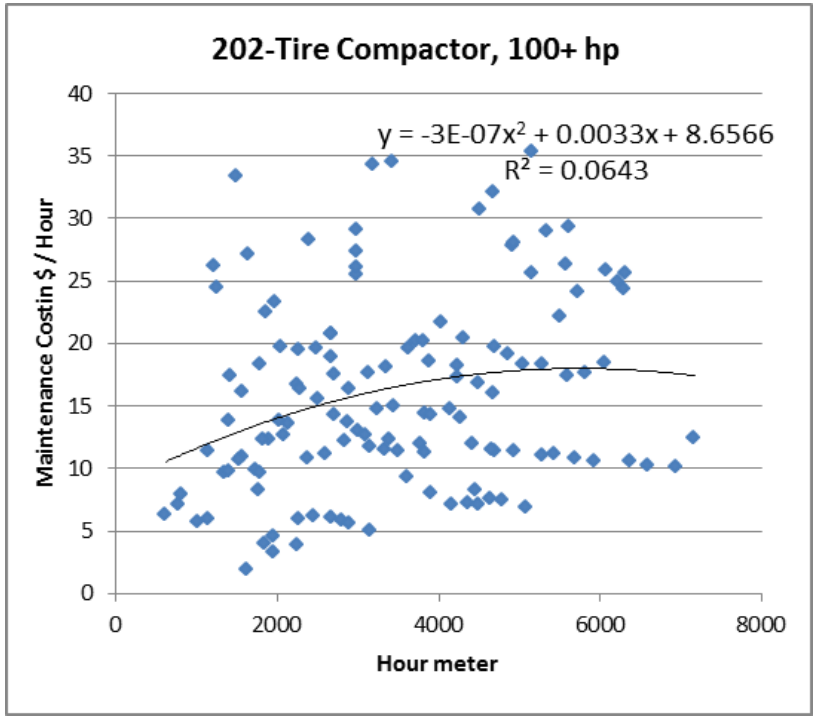
Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79-82.

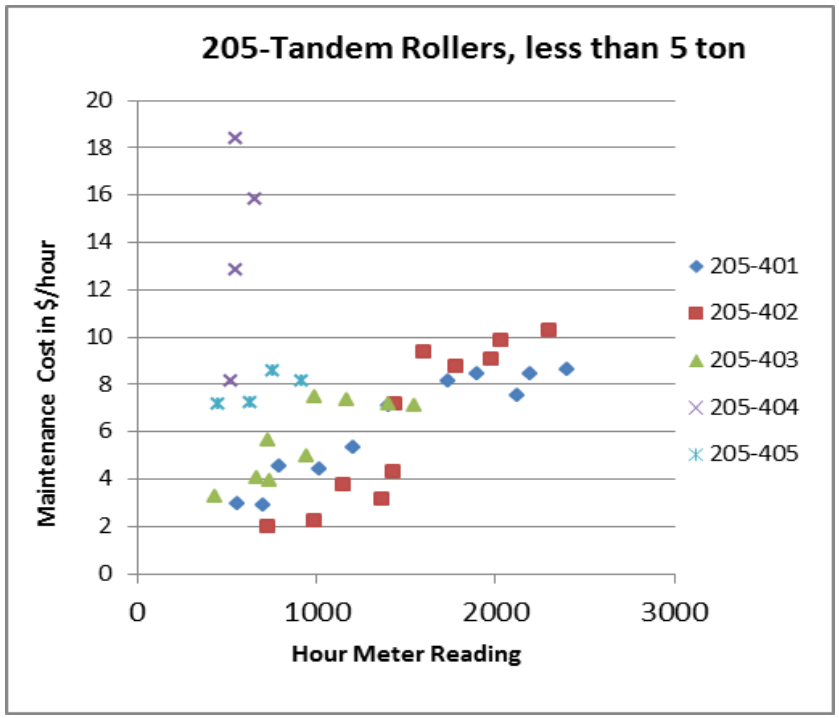
Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques, second edition*. San Francisco, CA: Morgan Kauffman Publishers.

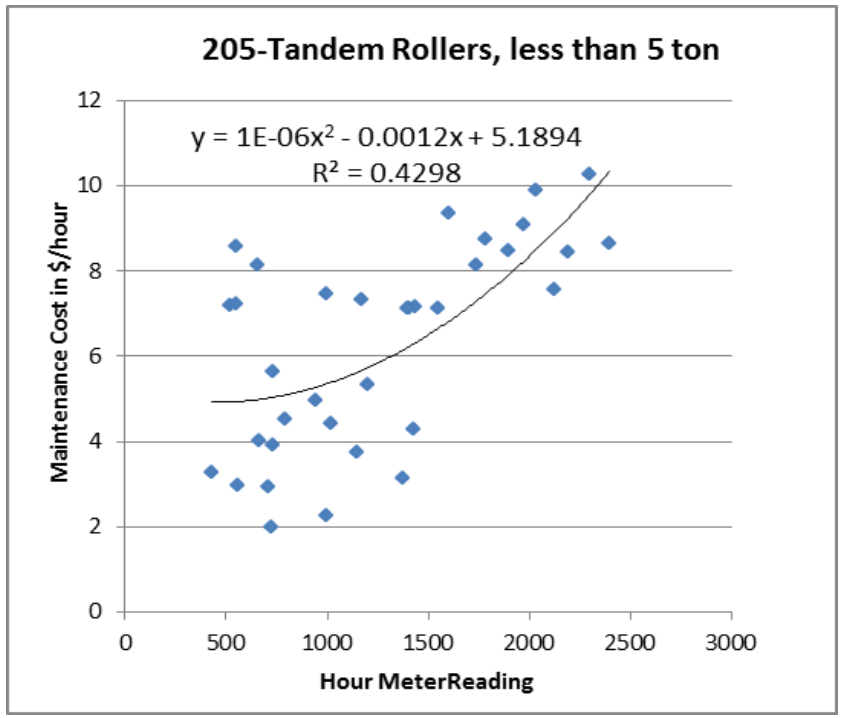
Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining practical machine learning tools and techniques* (3rd ed.). Amsterdam, Boston, Heidelberg, London, New York, Ocford, Paris, San Diego, San Francisico, Singaporem, Sydney, Tokyo: Morgan Kaufmann Publishers.

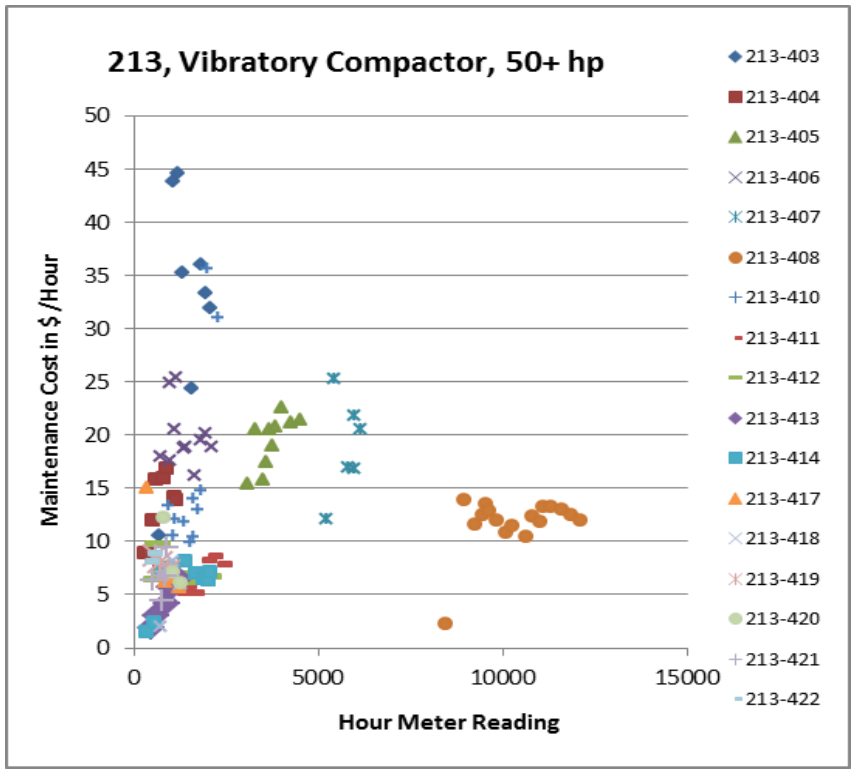
Appendix 1- Cost per hour trend analysis for all available equipment classes between class number 200 and 299

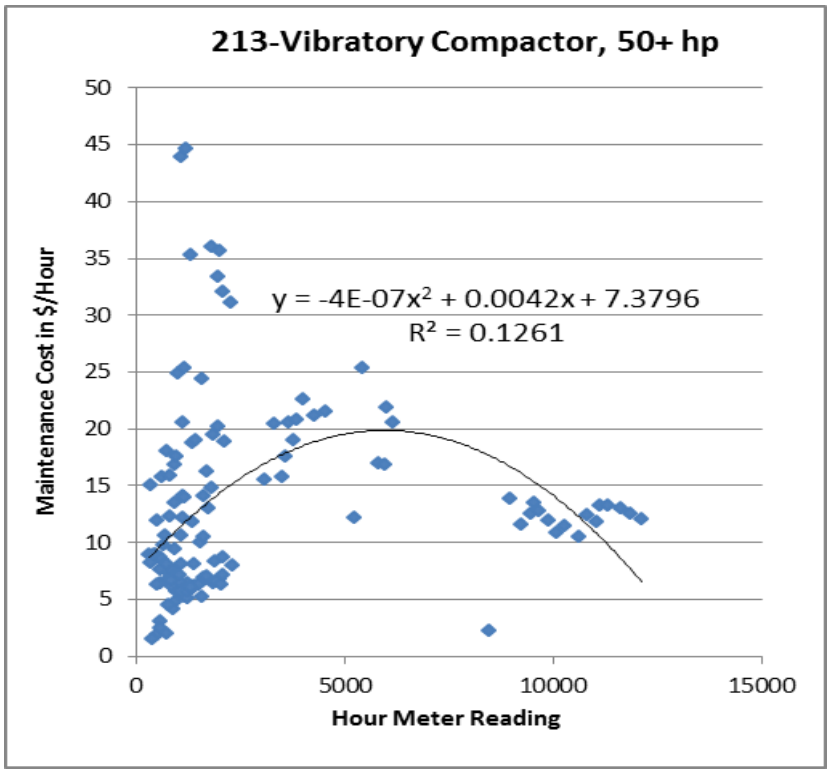


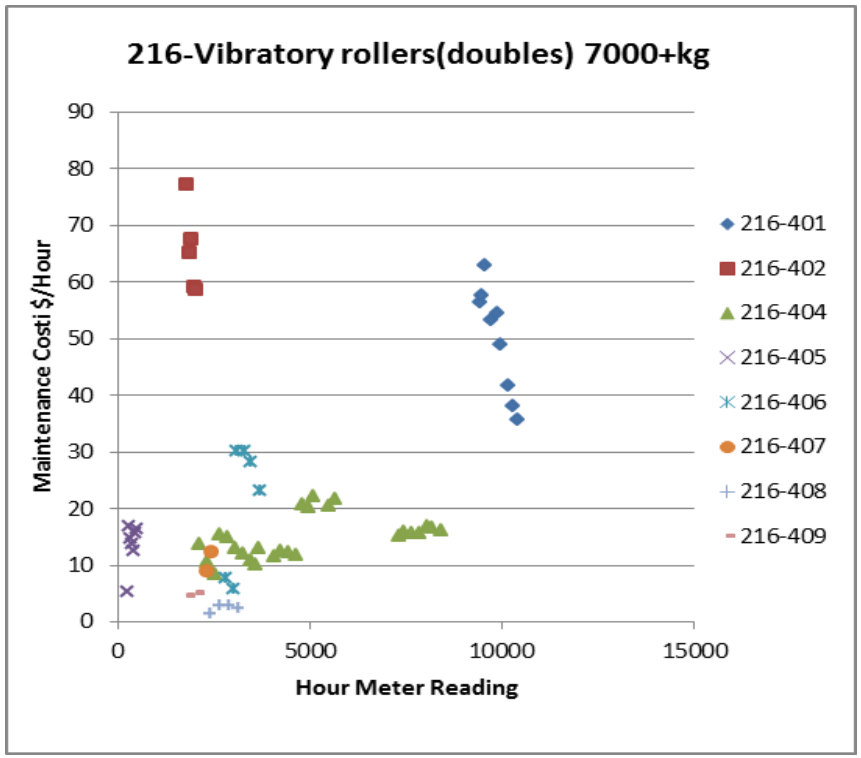


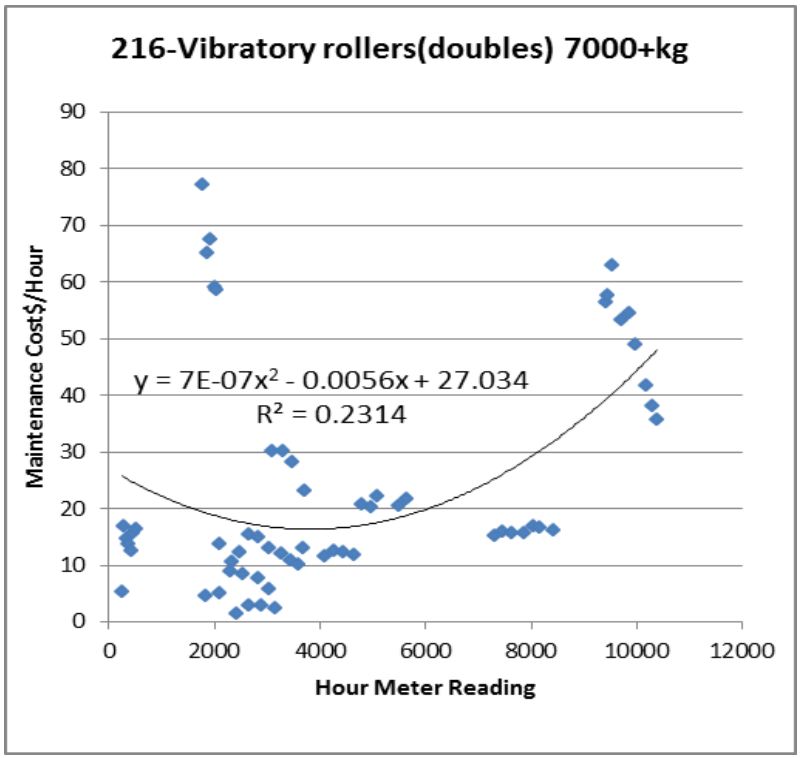

205-Tandem Rollers, less than 5 ton
Maintenance Cost in $/hour
Hour Meter Reading
0
2
4
6
8
10
12
14
16
18
20
0
1000
2000
3000
205-401
205-402
205-403
205-404
205-405

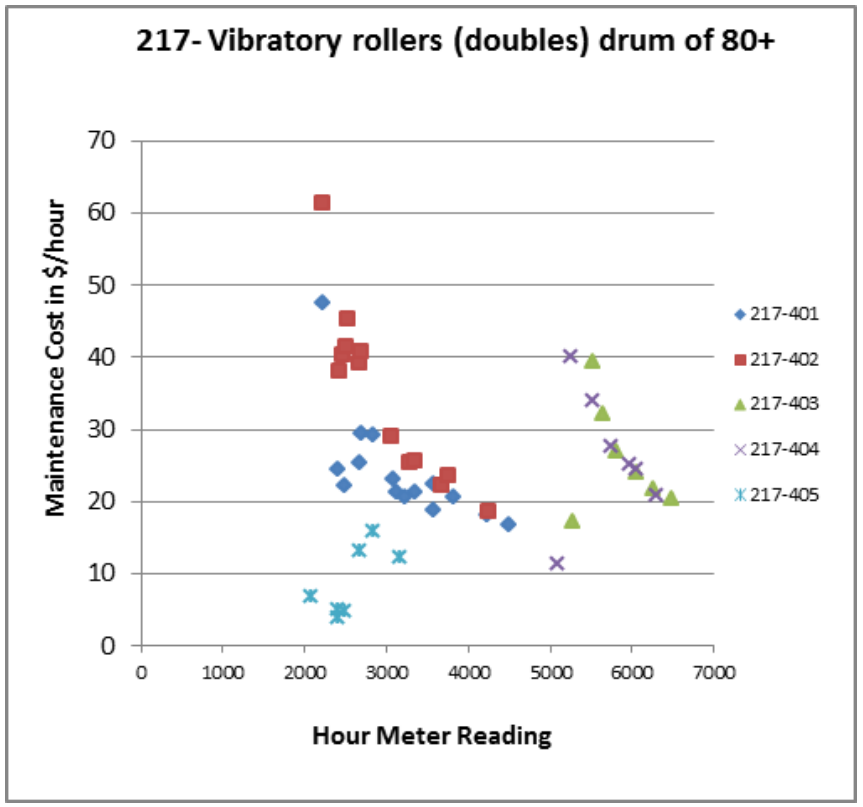

205-Tandem Rollers, less than 5 ton
y = 1E-06x² - 0.0012x + 5.1894
R² = 0.4298
Maintenance Cost in $/hour
Hour MeterReading
0
2
4
6
8
10
12
0
500
1000
1500
2000
2500
3000

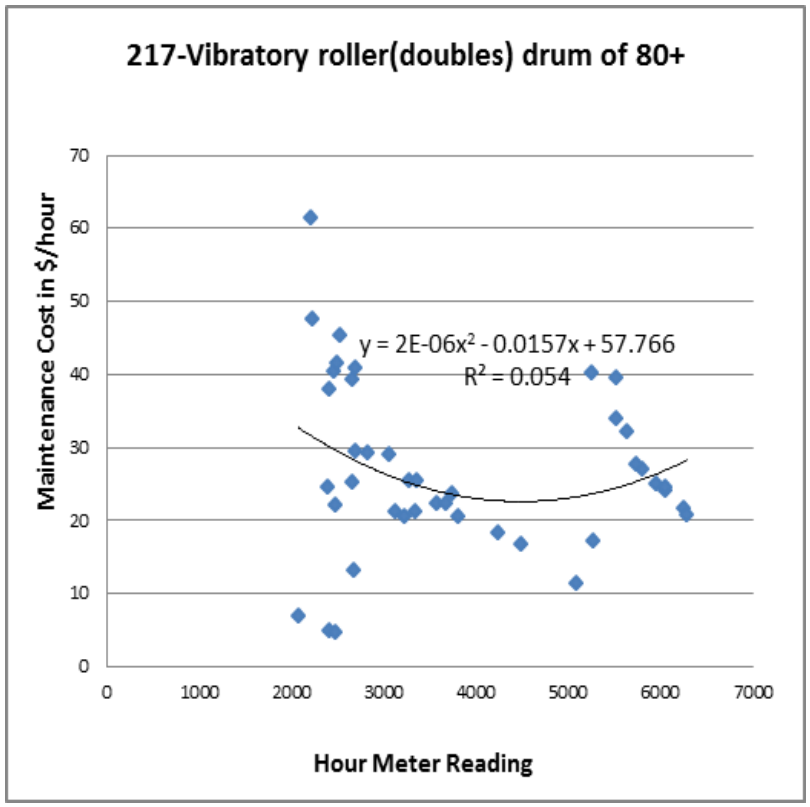

213, Vibratory Compactor, 50+ hp
Maintenance Cost in $ /Hour
Hour Meter Reading
0
5
10
15
20
25
30
35
40
45
50
0
5000
10000
15000
213-403
213-404
213-405
213-406
213-407
213-408
213-410
213-411
213-412
213-413
213-414
213-417
213-418
213-419
213-420
213-421
213-422

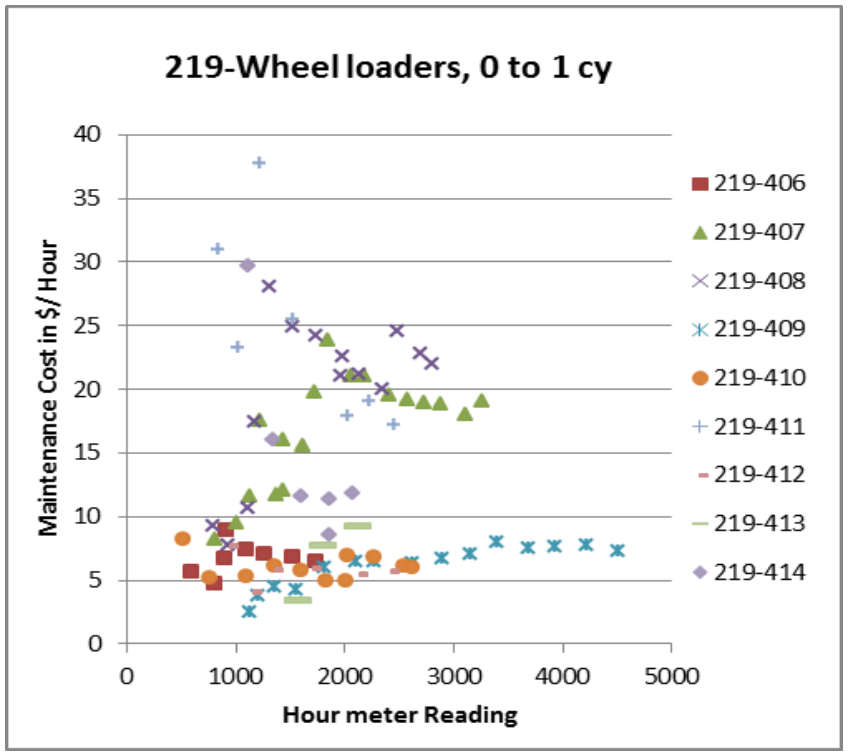

213-Vibratory Compactor, 50+ hp
Maintenance Cost in $/Hour
Hour Meter Reading
y = -4E-07x² + 0.0042x + 7.3796
R² = 0.1261
0
5
10
15
20
25
30
35
40
45
50
0
5000
10000
15000

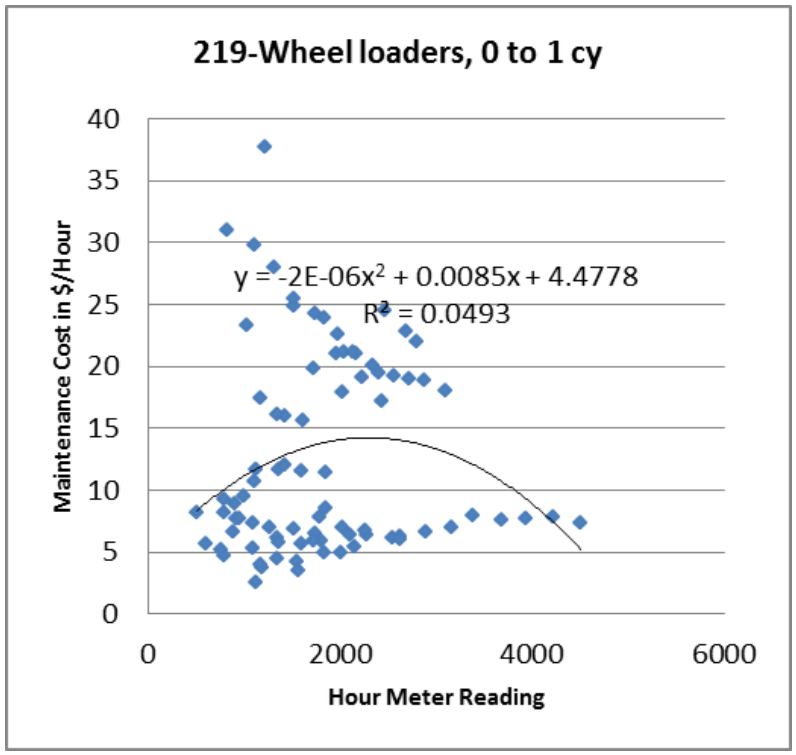

216-Vibratory rollers(doubles) 7000+kg
Maintenance Cost$/Hour
Hour Meter Reading
0
10
20
30
40
50
60
70
80
90
0
5000
10000
15000
216-401
216-402
216-404
216-405
216-406
216-407
216-408
216-409

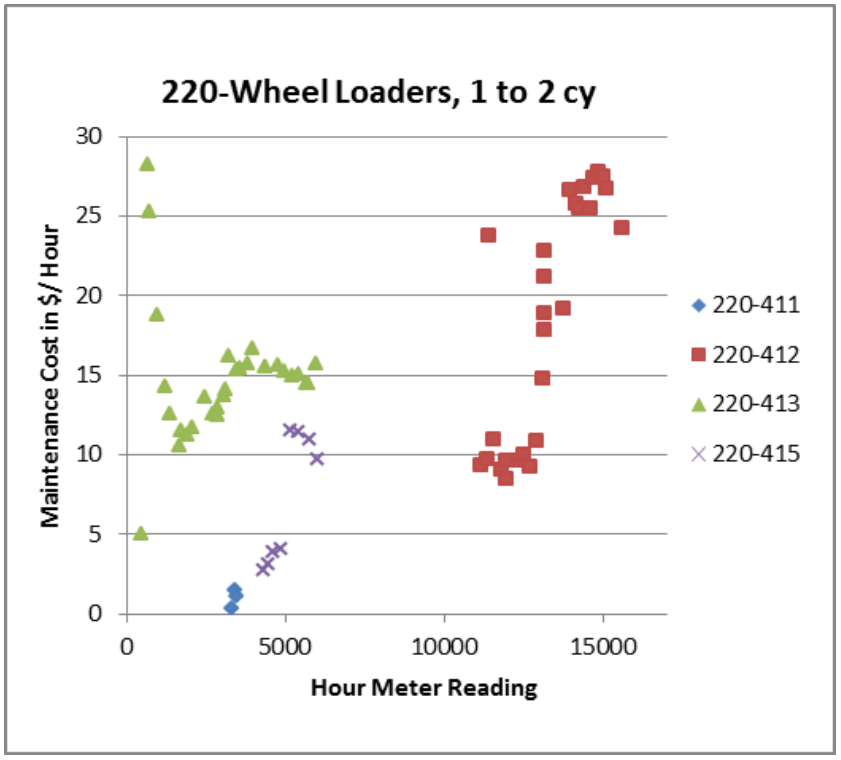

216-Vibratory rollers(doubles) 7000+kg
Maintenance Cost$/Hour
Hour Meter Reading
y = 7E-07x² - 0.0056x + 27.034
R² = 0.2314
0
10
20
30
40
50
60
70
80
90
0
2000
4000
6000
8000
10000
12000

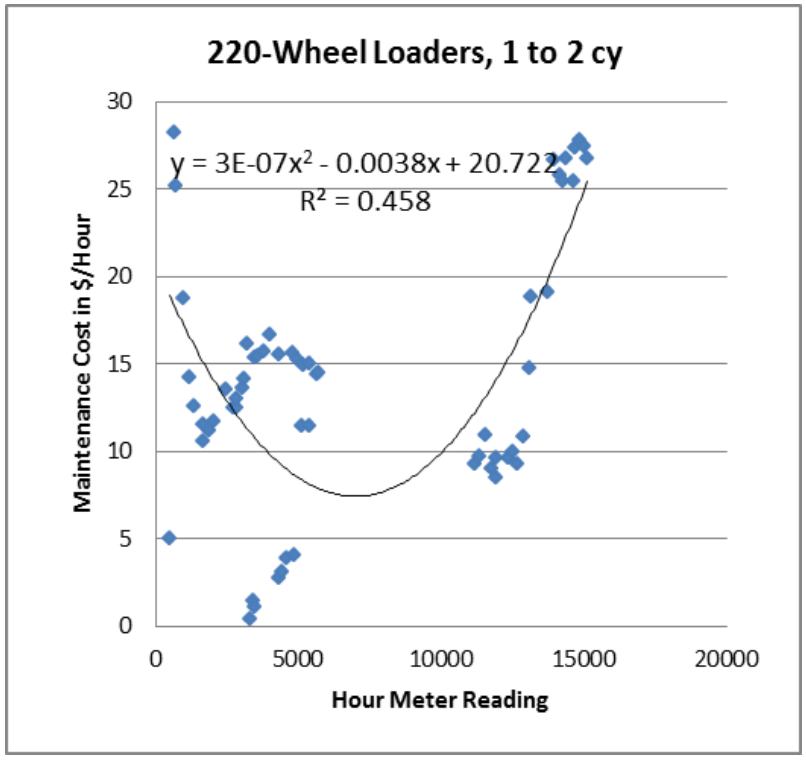

217- Vibratory rollers (doubles) drum of 80+
Maintenance Cost in $/hour
70
60
50
40
30
20
10
0
0
1000
2000
3000
4000
5000
6000
7000
Hour Meter Reading
217-401
217-402
217-403
217-404
217-405

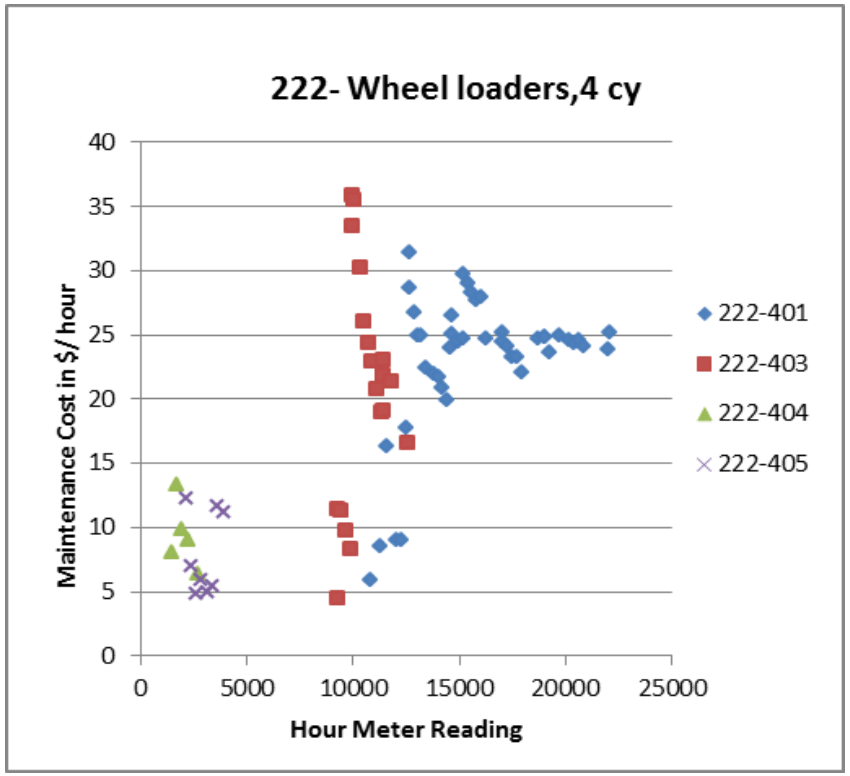

217-Vibratory roller(doubles) drum of 80+
Maintenance Cost in $/hour
70
60
50
40
30
20
10
0
0
1000
2000
3000
4000
5000
6000
7000
Hour Meter Reading
y = 2E-06x² - 0.0157x + 57.766
R² = 0.054

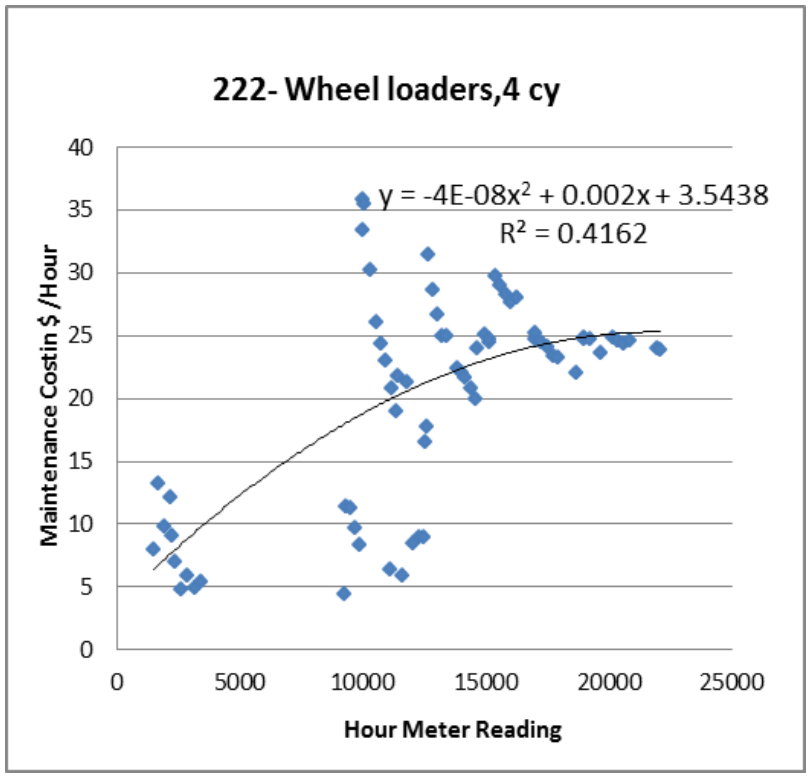

219-Wheel loaders, 0 to 1 cy
Maintenance Cost in $/ Hour
Hour meter Reading
0
5
10
15
20
25
30
35
40
0
1000
2000
3000
4000
5000
219-406
219-407
219-408
219-409
219-410
219-411
219-412
219-413
219-414

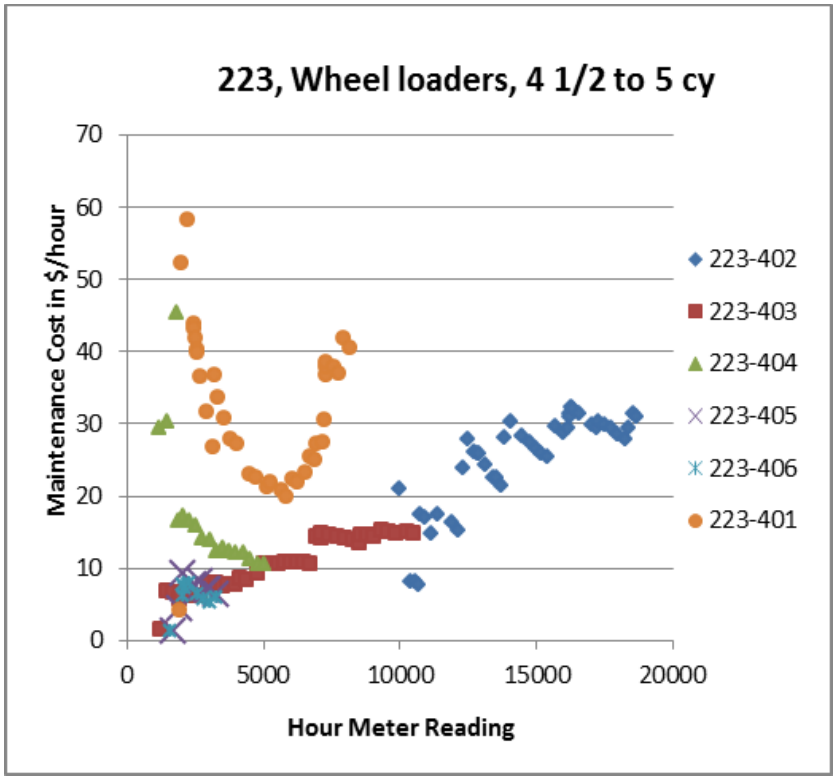

219-Wheel loaders, 0 to 1 cy
Maintenance Cost in $/Hour
Hour Meter Reading
y = -2E-06x² + 0.0085x + 4.4778
R² = 0.0493
0
5
10
15
20
25
30
35
40
0
2000
4000
6000

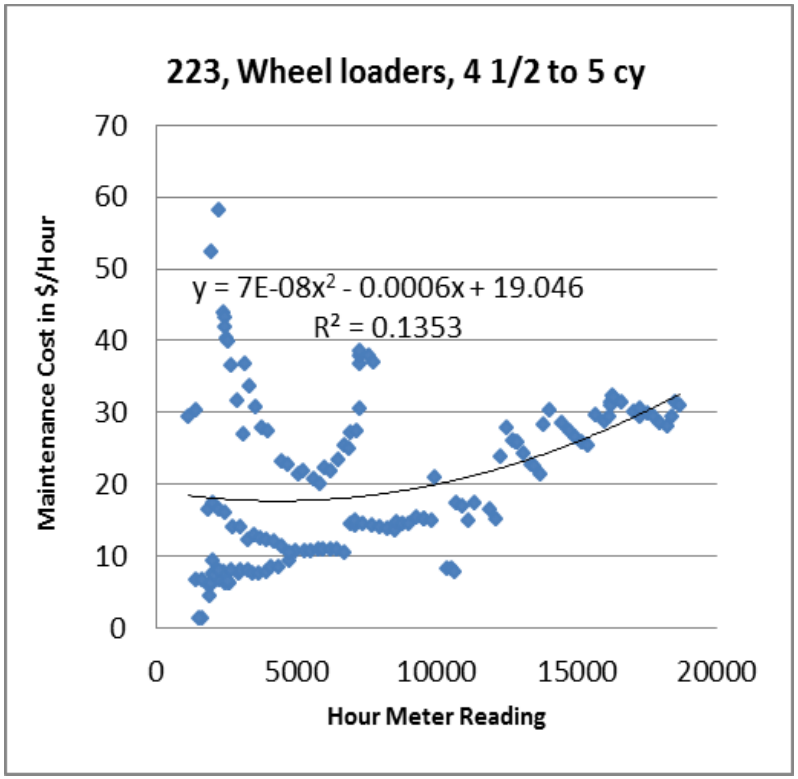

220-Wheel Loaders, 1 to 2 cy
Maintenance Cost in $/ Hour
Hour Meter Reading
0
5
10
15
20
25
30
0
5000
10000
15000
220-411
220-412
220-413
220-415

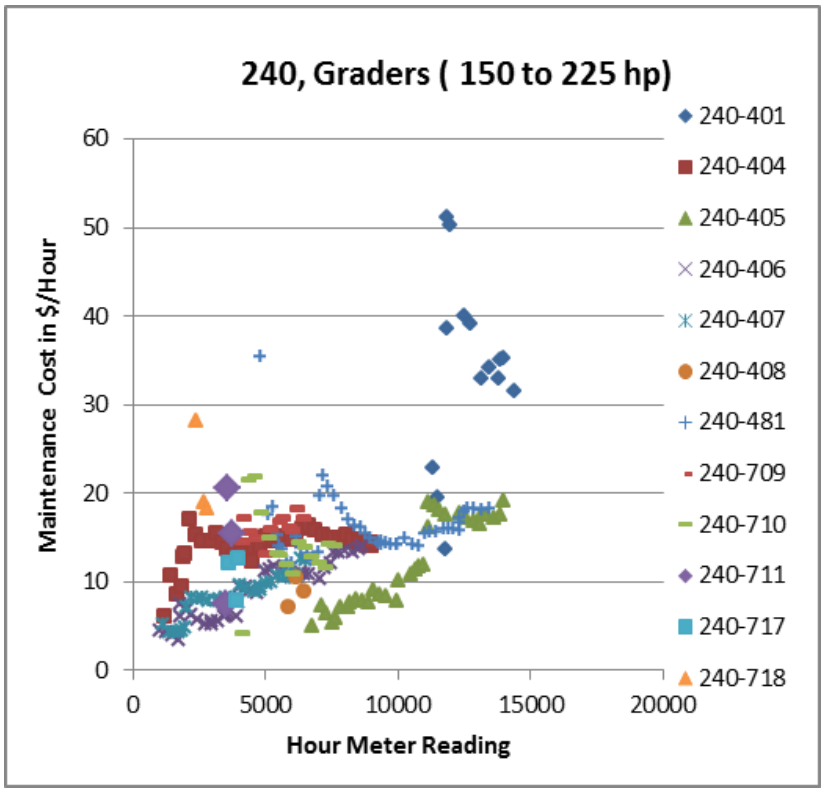

220-Wheel Loaders, 1 to 2 cy
y = 3E-07x² - 0.0038x + 20.722
R² = 0.458
Maintenance Cost in $/Hour
Hour Meter Reading
0
5
10
15
20
25
30
0
5000
10000
15000
20000

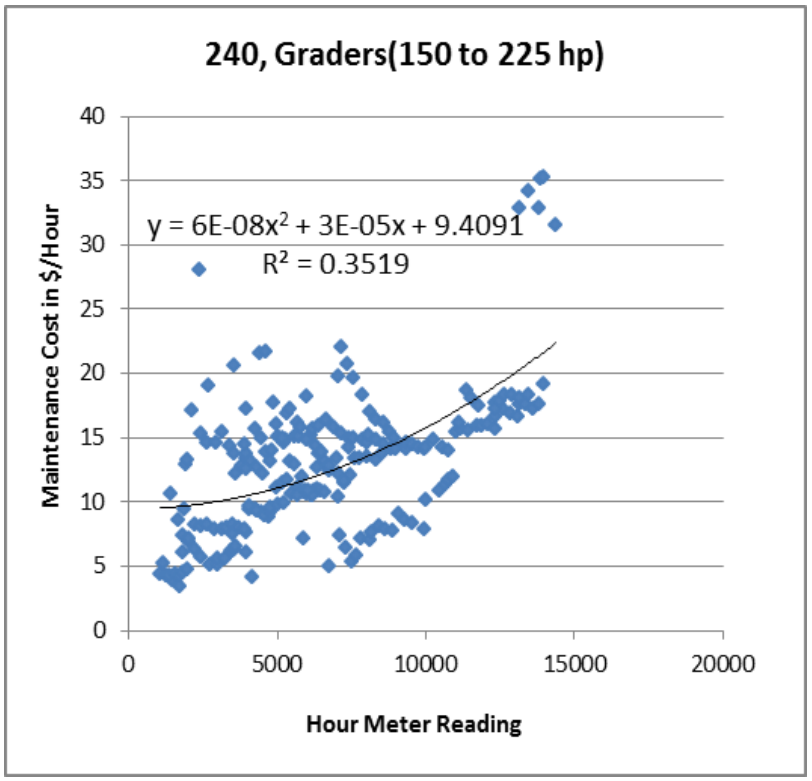

222- Wheel loaders,4 cy
Maintenance Cost in $/ hour
Hour Meter Reading
0
5
10
15
20
25
30
35
40
0
5000
10000
15000
20000
25000
222-401
222-403
222-404
222-405

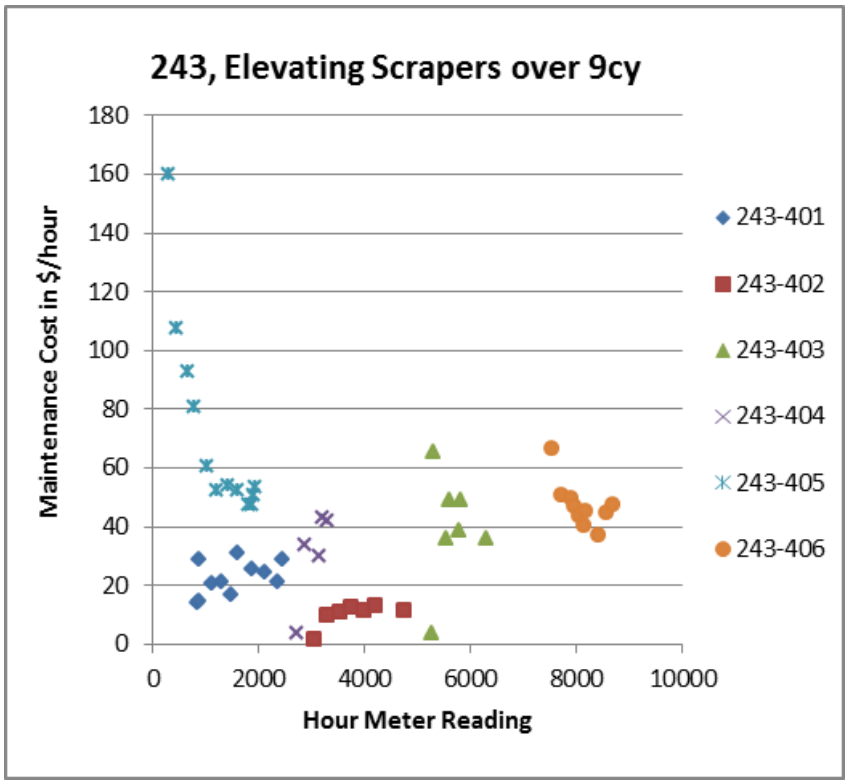

222- Wheel loaders,4 cy
Maintenance Costin $ /Hour
Hour Meter Reading
y = -4E-08x² + 0.002x + 3.5438
R² = 0.4162
0
5
10
15
20
25
30
35
40
0
5000
10000
15000
20000
25000

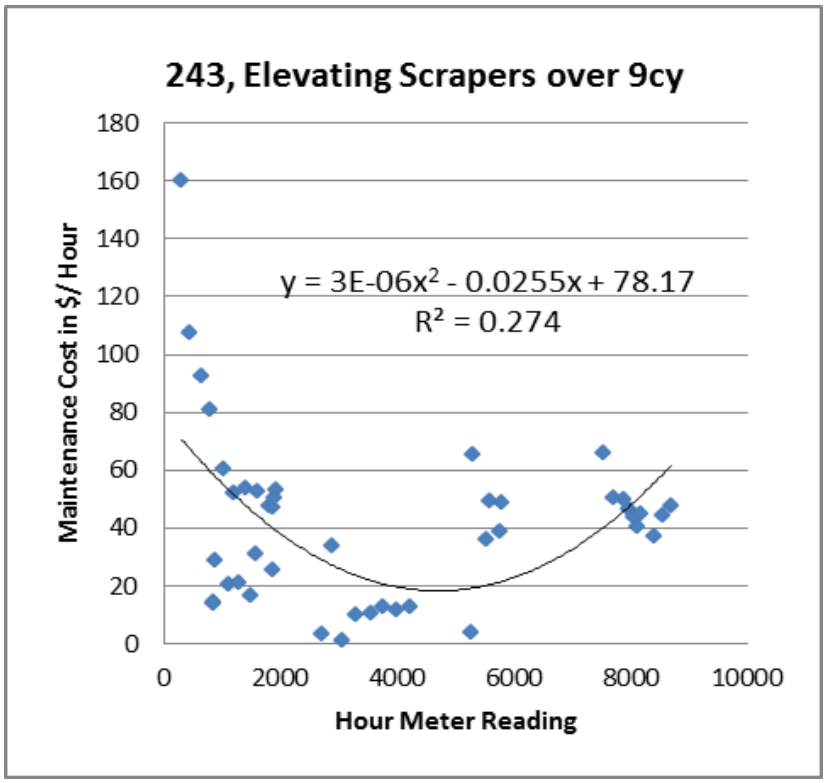

223, Wheel loaders, 4 1/2 to 5 cy
Maintenance Cost in $/hour
Hour Meter Reading
0
10
20
30
40
50
60
70
0
5000
10000
15000
20000
223-402
223-403
223-404
223-405
223-406
223-401

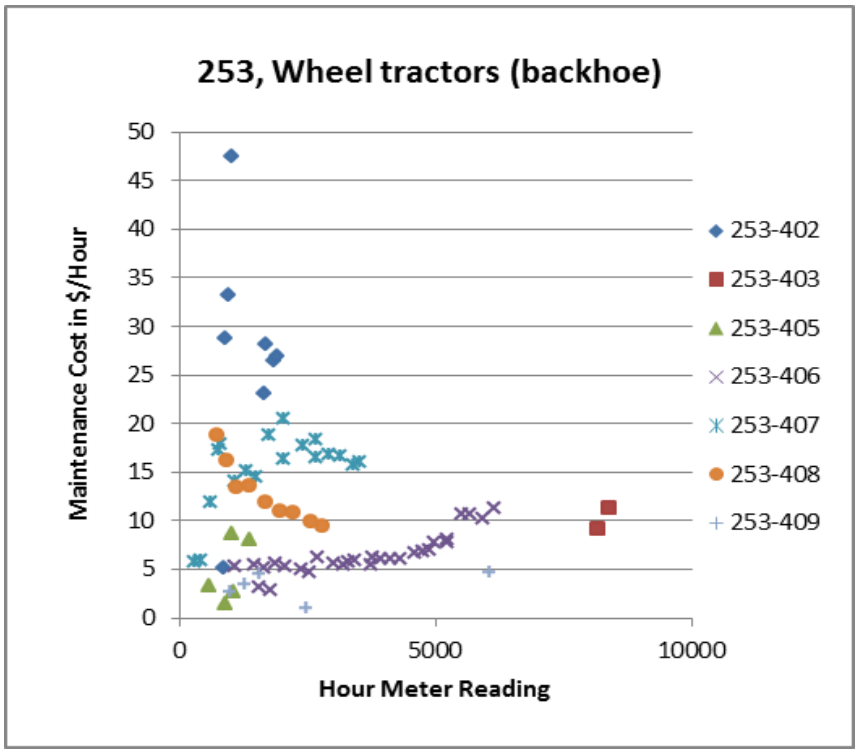

223, Wheel loaders, 4 1/2 to 5 cy
Maintenance Cost in $/Hour
Hour Meter Reading
y = 7E-08x² - 0.0006x + 19.046
R² = 0.1353
0
10
20
30
40
50
60
70
0
5000
10000
15000
20000

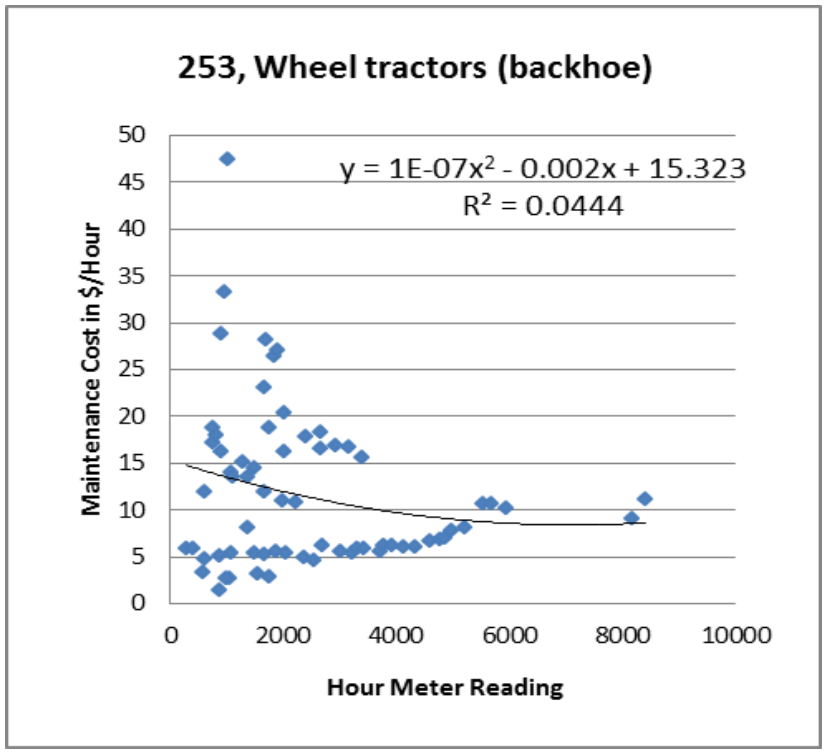

240, Graders (150 to 225 hp)
Maintenance Cost in $/Hour
Hour Meter Reading
0
10
20
30
40
50
60
0
5000
10000
15000
20000
240-401
240-404
240-405
240-406
240-407
240-408
240-481
240-709
240-710
240-711
240-717
240-718

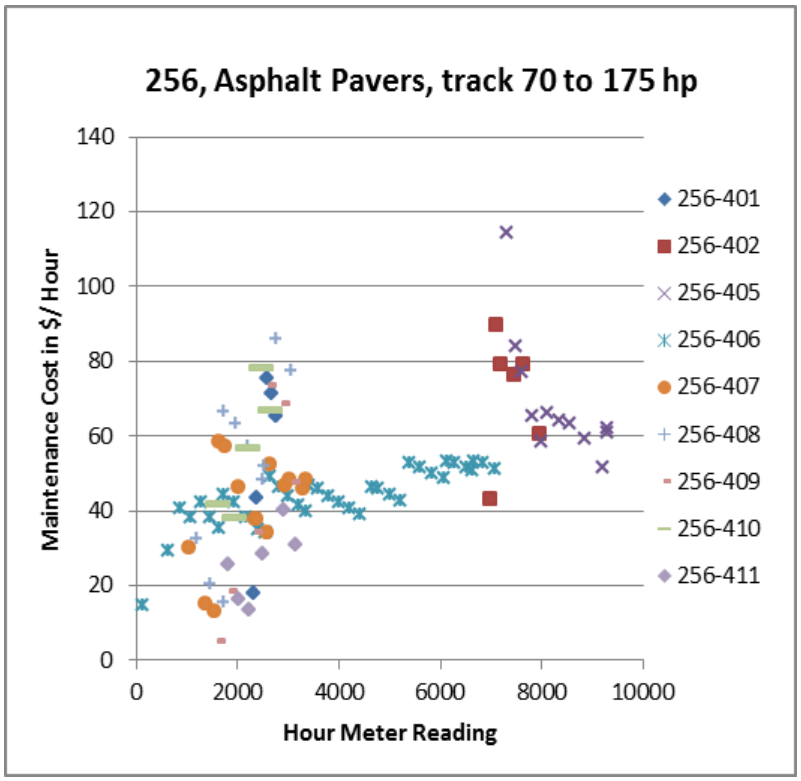

240, Graders(150 to 225 hp)
y = 6E-08x² + 3E-05x + 9.4091
R² = 0.3519
Maintenance Cost in $/Hour
Hour Meter Reading
0
5
10
15
20
25
30
35
40
0
5000
10000
15000
20000

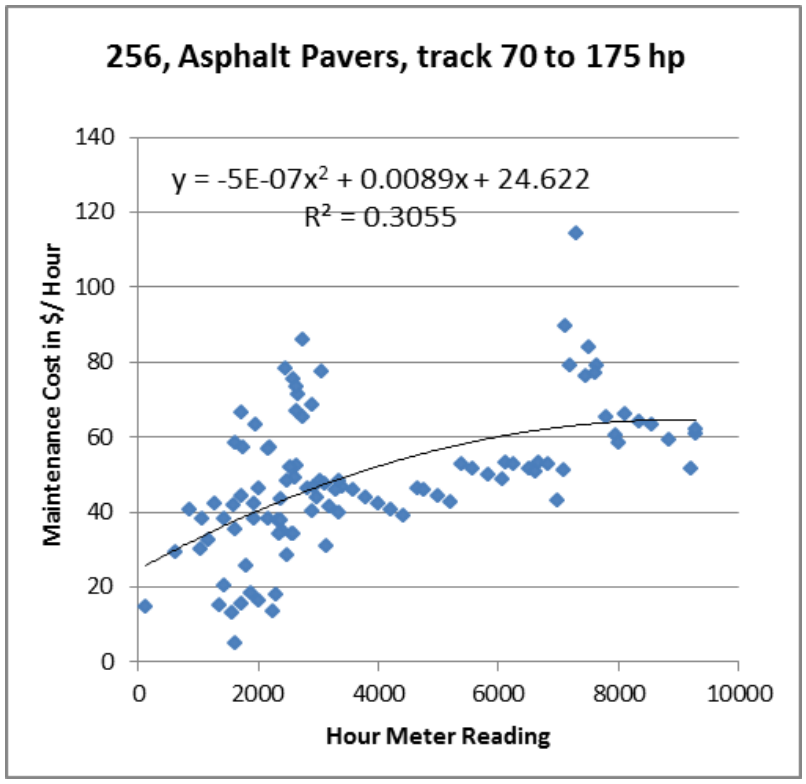

243, Elevating Scrapers over 9cy
Maintenance Cost in $/hour
180
160
140
120
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
Hour Meter Reading
243-401
243-402
243-403
243-404
243-405
243-406

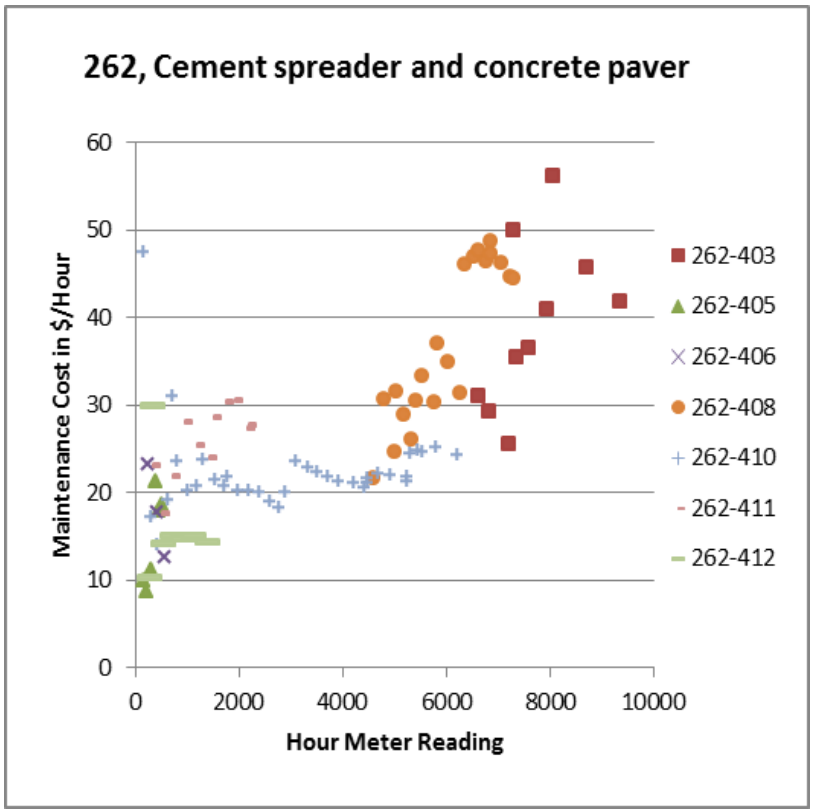

243, Elevating Scrapers over 9cy
Maintenance Cost in $/ Hour
180
160
140
120
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
Hour Meter Reading
y = 3E-06x² - 0.0255x + 78.17
R² = 0.274

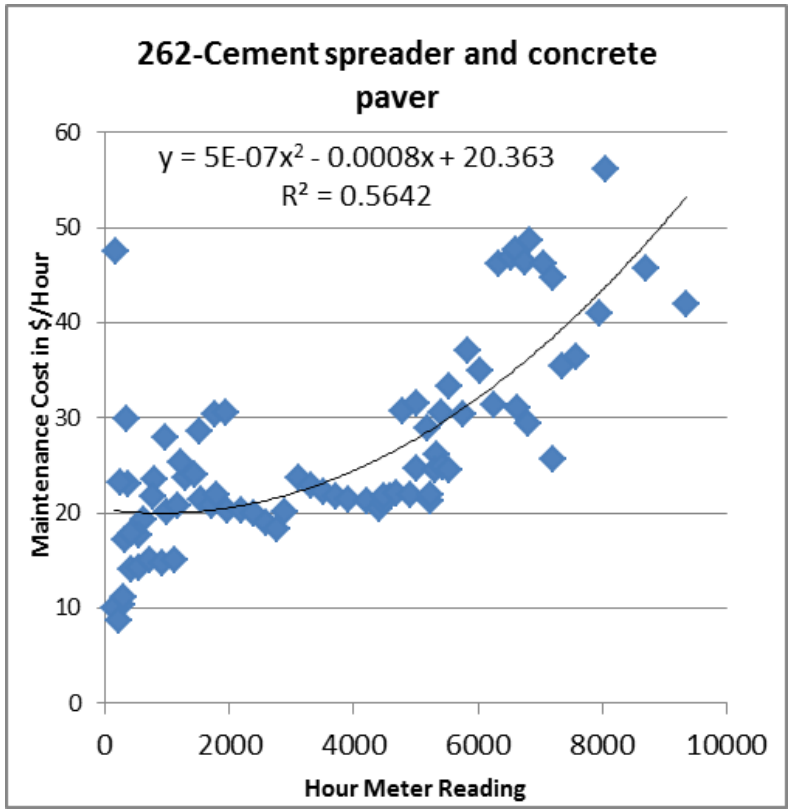

253, Wheel tractors (backhoe)
Maintenance Cost in $/Hour
Hour Meter Reading
0
5000
10000
0
5
10
15
20
25
30
35
40
45
50
253-402
253-403
253-405
253-406
253-407
253-408
253-409

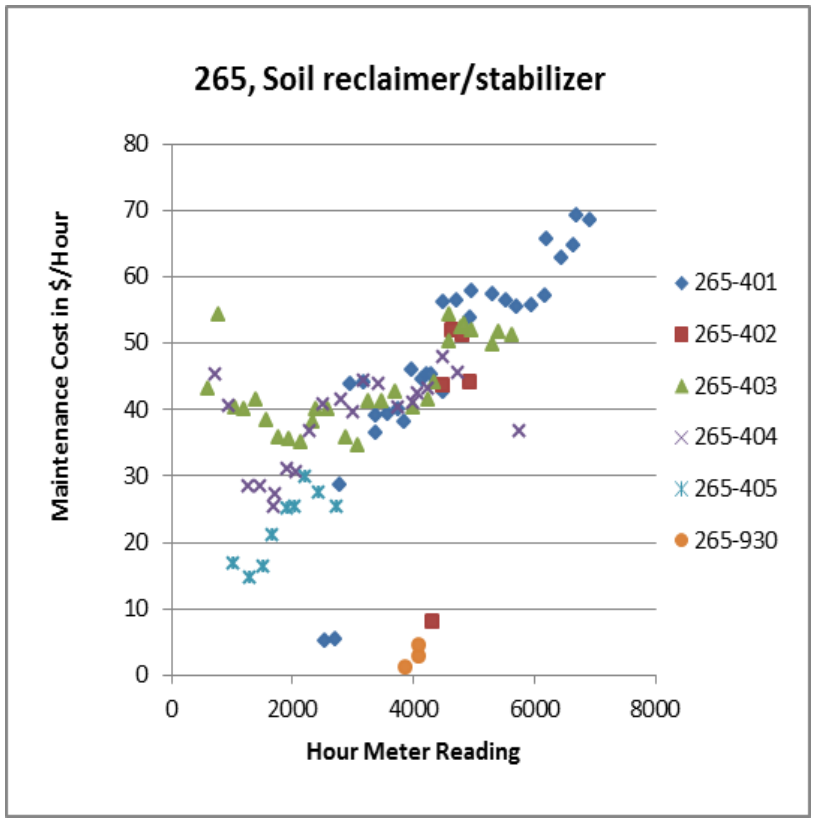

253, Wheel tractors (backhoe)
y = 1E-07x² - 0.002x + 15.323
R² = 0.0444
Maintenance Cost in $/Hour
Hour Meter Reading
0
2000
4000
6000
8000
10000
0
5
10
15
20
25
30
35
40
45
50

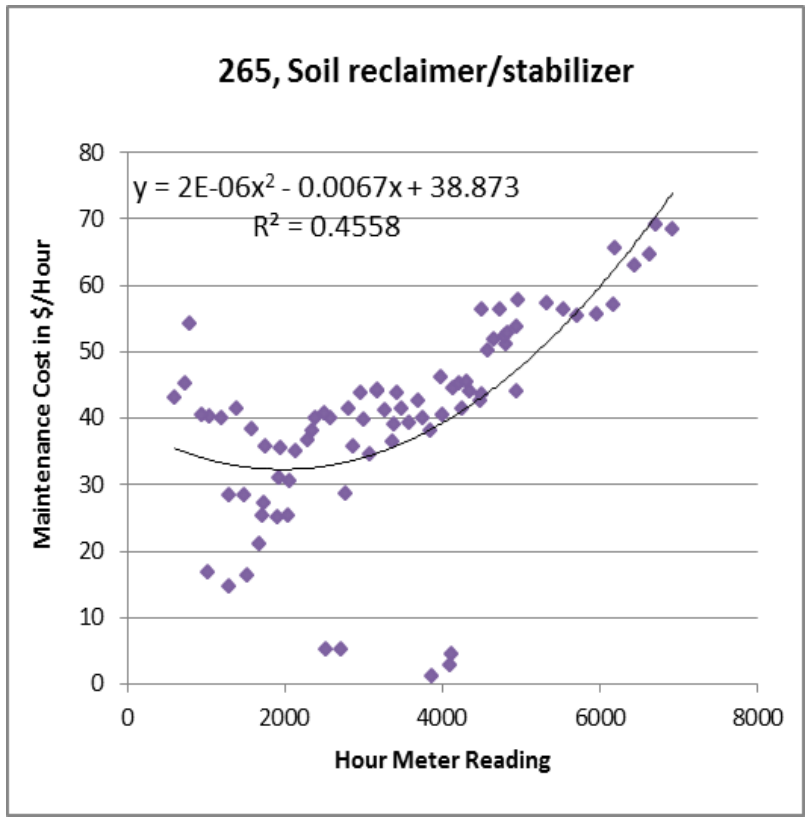

256, Asphalt Pavers, track 70 to 175 hp
Maintenance Cost in $/ Hour
Hour Meter Reading
0
20
40
60
80
100
120
140
0
2000
4000
6000
8000
10000
256-401
256-402
256-405
256-406
256-407
256-408
256-409
256-410
256-411


256, Asphalt Pavers, track 70 to 175 hp
y = -5E-07x² + 0.0089x + 24.622
R² = 0.3055
Maintenance Cost in $/ Hour
Hour Meter Reading
0
20
40
60
80
100
120
140
0
2000
4000
6000
8000
10000

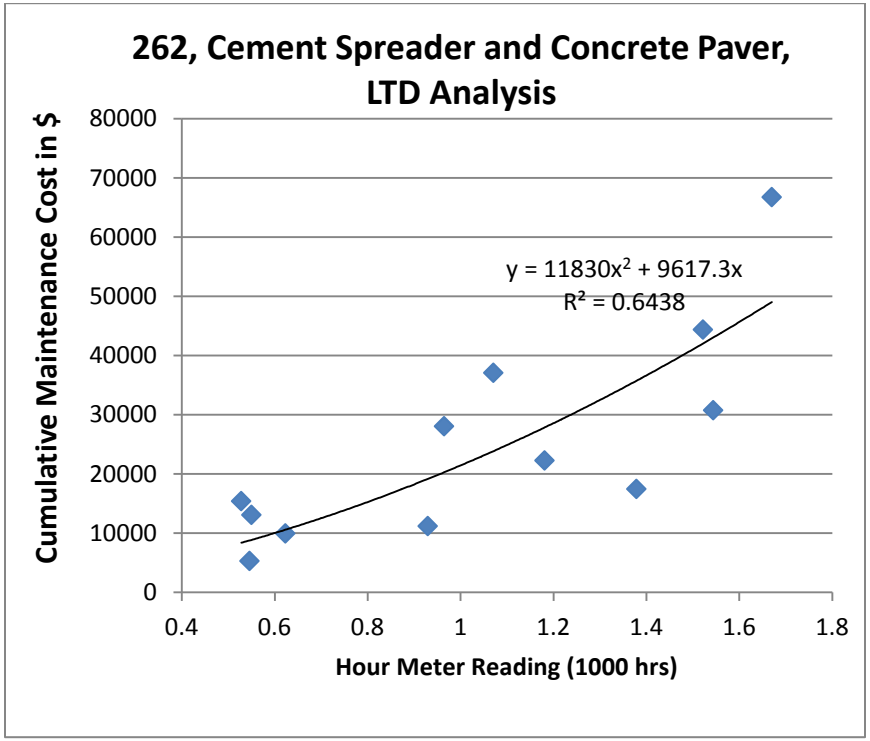

262, Cement spreader and concrete paver
Maintenance Cost in $/Hour
Hour Meter Reading
0
10
20
30
40
50
60
0
2000
4000
6000
8000
10000
262-403
262-405
262-406
262-408
262-410
262-411
262-412

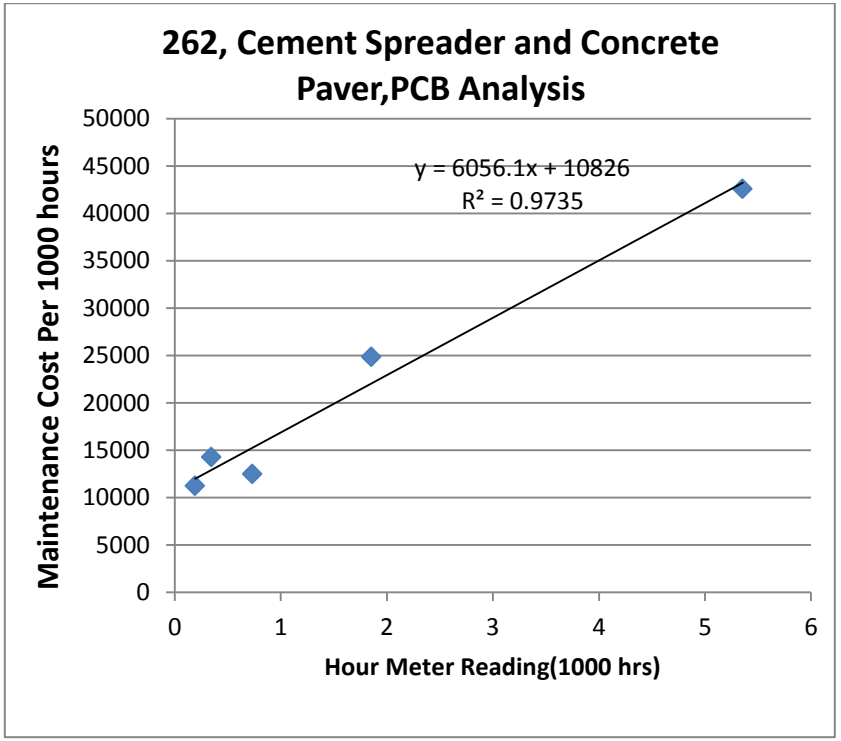

262-Cement spreader and concrete paver
y = 5E-07x² - 0.0008x + 20.363
R² = 0.5642
Maintenance Cost in $/Hour
Hour Meter Reading
0
10
20
30
40
50
60
0
2000
4000
6000
8000
10000

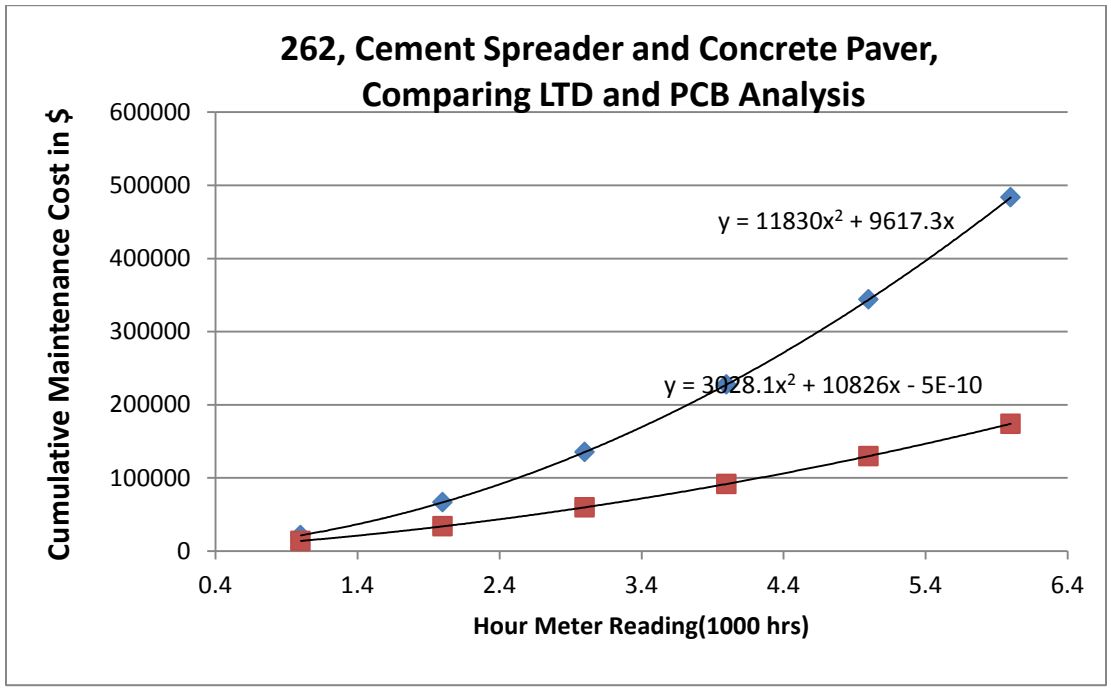

265, Soil reclaimer/stabilizer
Maintenance Cost in $/Hour
Hour Meter Reading
0
10
20
30
40
50
60
70
80
0
2000
4000
6000
8000
265-401
265-402
265-403
265-404
265-405
265-930


265, Soil reclaimer/stabilizer
y = 2E-06x² - 0.0067x + 38.873
R² = 0.4558
Maintenance Cost in $/Hour
Hour Meter Reading
0
10
20
30
40
50
60
70
80
0
2000
4000
6000
8000

Appendix 2- Cumulative cost modeling for equipment class 262 (cement spreader and concrete paver)

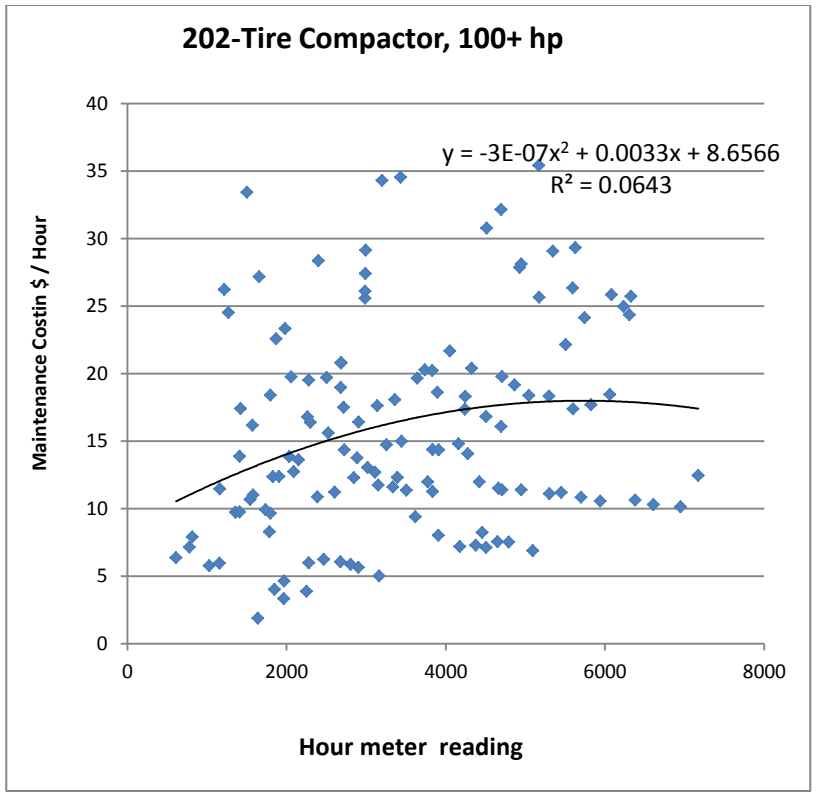


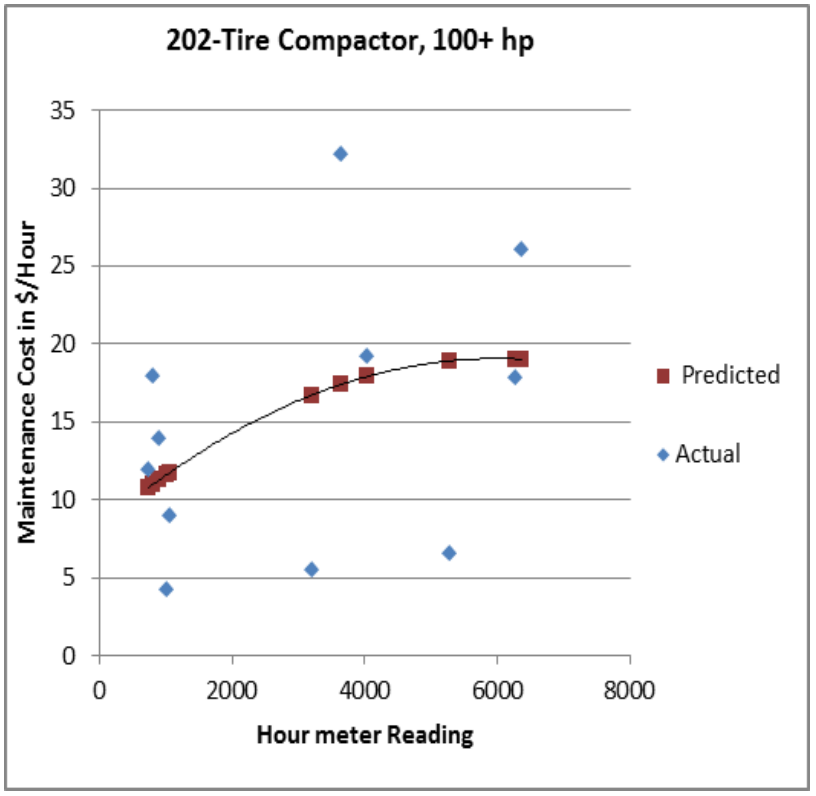


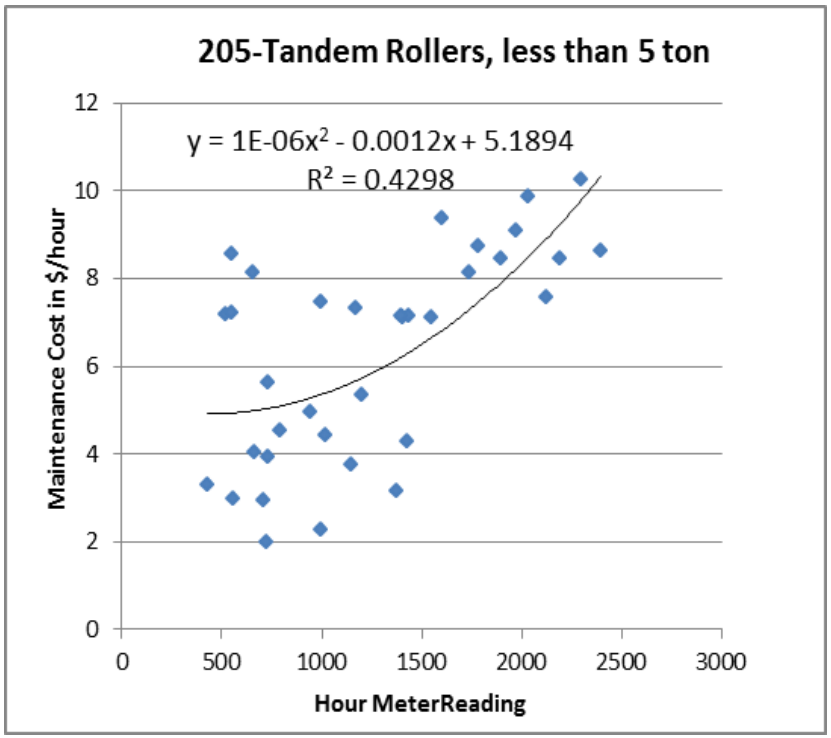

262, Cement Spreader and Concrete Paver, Comparing LTD and PCB Analysis
Cumulative Maintenance Cost in $
600000
500000
400000
300000
200000
100000
0
0.4
1.4
2.4
3.4
4.4
5.4
6.4
Hour Meter Reading(1000 hrs)
y = 11830x² + 9617.3x
y = 3028.1x² + 10826x - 5E-10

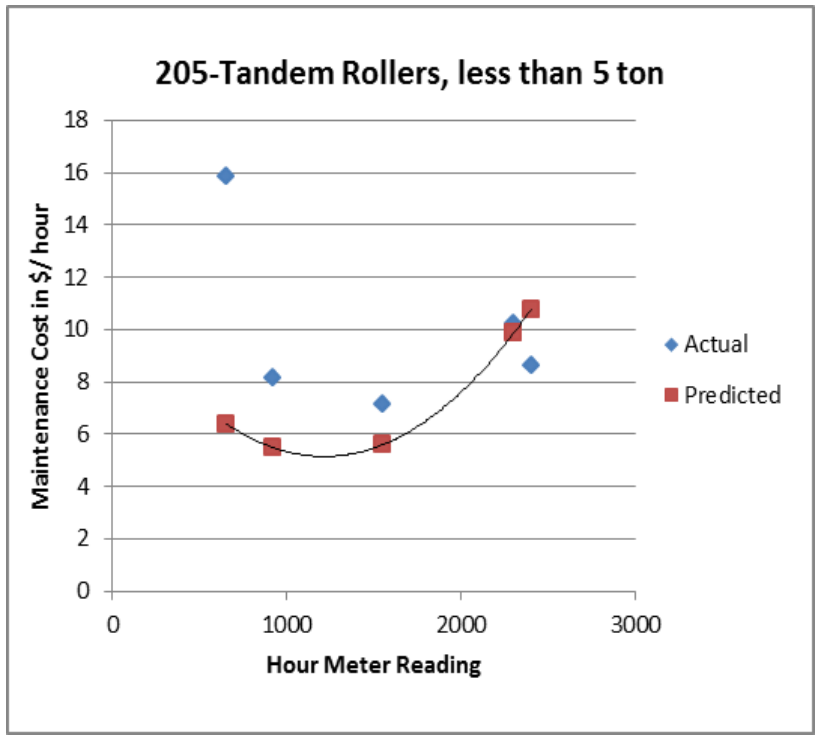
Appendix 3- Second order nonlinear regression analysis for all available equipment classes between class number 200 and 299

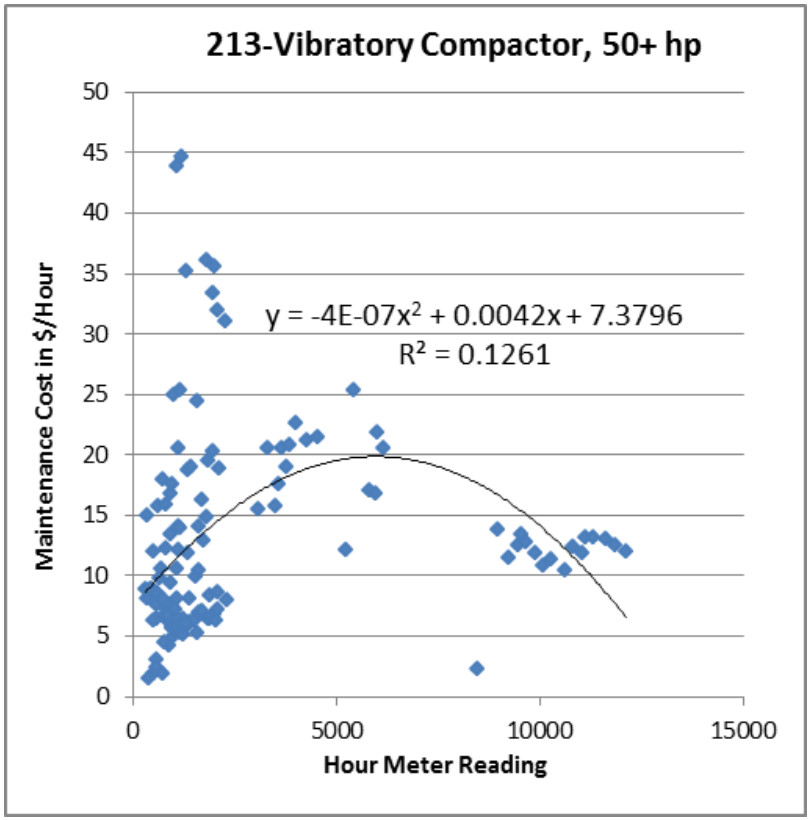
- Left Graph: Forming equation of second order nonlinear regression for all available classes between class number 200 and 299
- Right Graph: Comparison of predicted values with actual values from second order nonlinear regression analysis for all available classes between class number 200 and 299

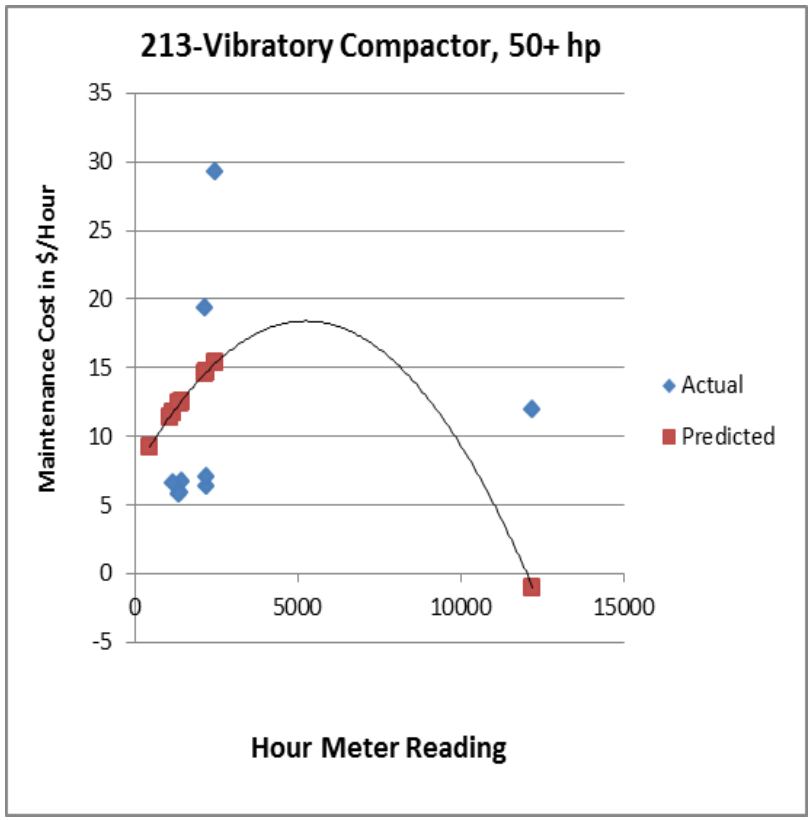

202-Tire Compactor, 100+ hp
y = -3E-07x² + 0.0033x + 8.6566
R² = 0.0643
Maintenance Costin $ / Hour
Hour meter reading
0
5
10
15
20
25
30
35
40
0
2000
4000
6000
8000

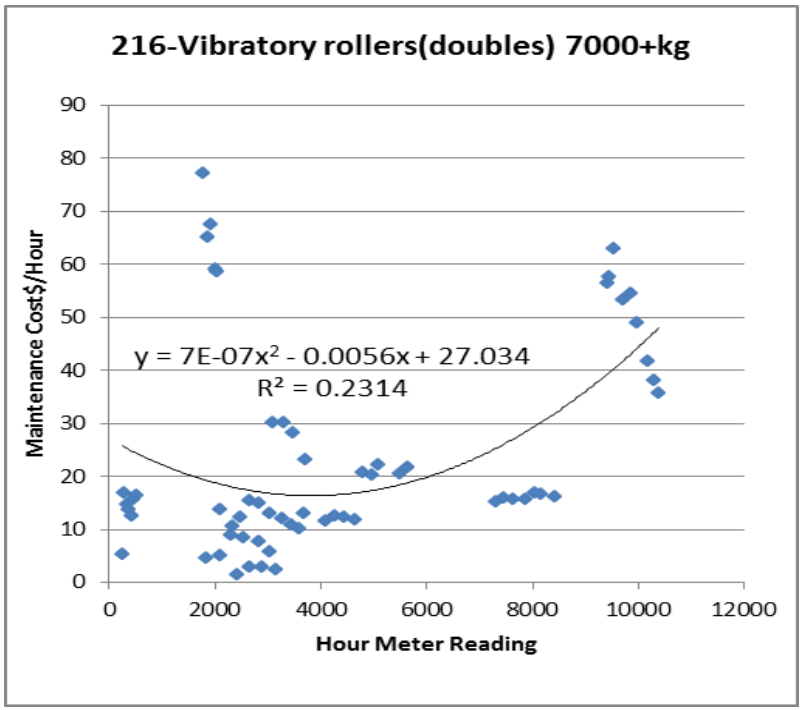

202-Tire Compactor, 100+ hp
Maintenance Cost in $/Hour
Hour meter Reading
Predicted
Actual
0
5
10
15
20
25
30
35
0
2000
4000
6000
8000

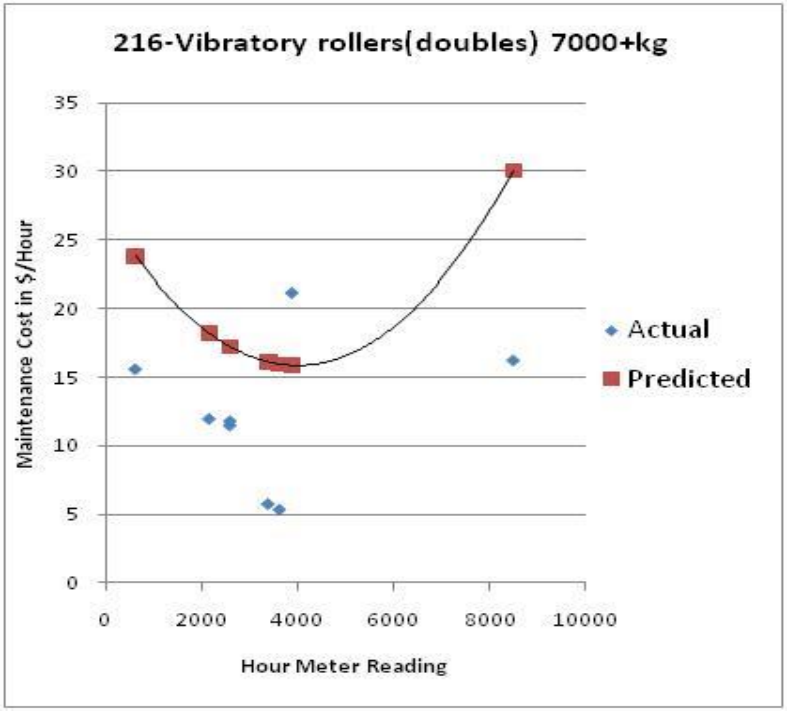

205-Tandem Rollers, less than 5 ton
y = 1E-06x² - 0.0012x + 5.1894
R² = 0.4298
Maintenance Cost in $/hour
Hour MeterReading
0
2
4
6
8
10
12
0
500
1000
1500
2000
2500
3000

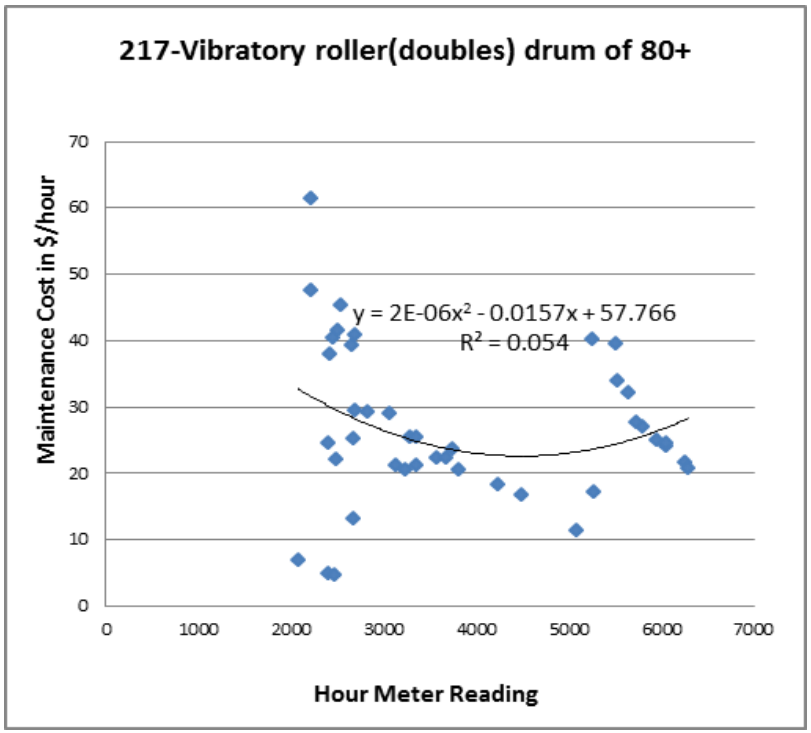

205-Tandem Rollers, less than 5 ton
Maintenance Cost in $/ hour
Hour Meter Reading
Actual
Predicted
0
2
4
6
8
10
12
14
16
18
0
1000
2000
3000

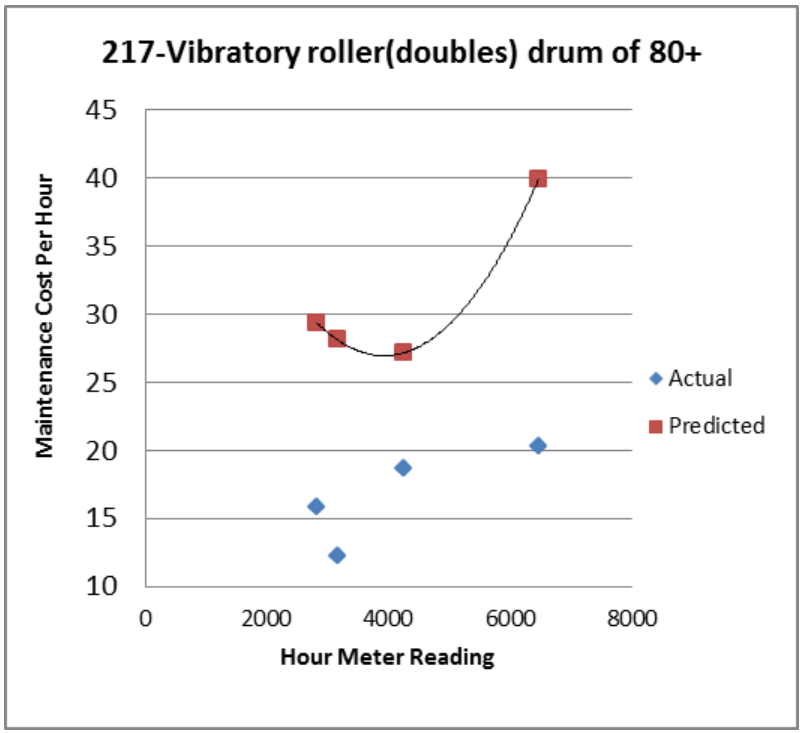

213-Vibratory Compactor, 50+ hp
y = -4E-07x² + 0.0042x + 7.3796
R² = 0.1261
Maintenance Cost in $/Hour
Hour Meter Reading
0
5
10
15
20
25
30
35
40
45
50
0
5000
10000
15000

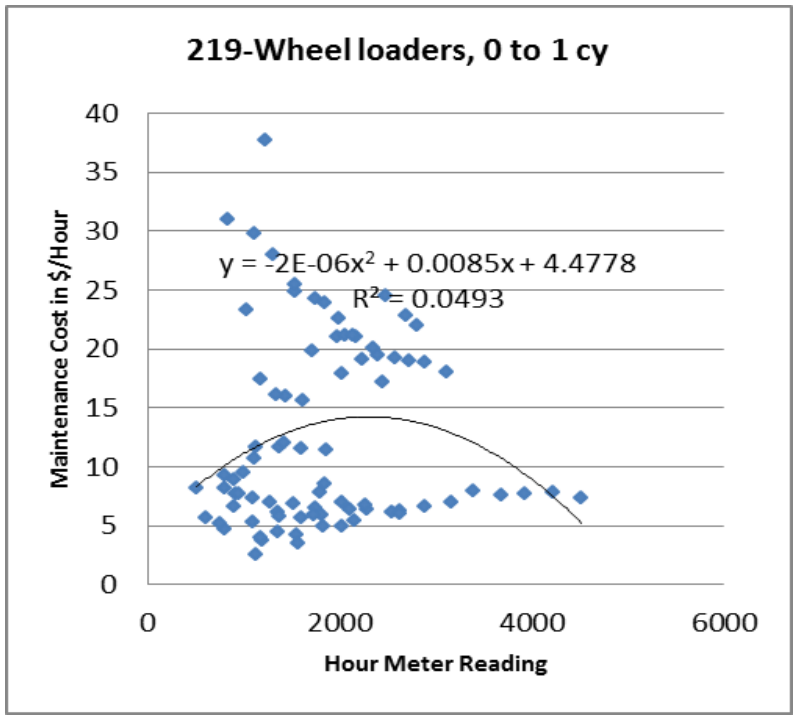

213-Vibratory Compactor, 50+ hp
Maintenance Cost in $/Hour
Hour Meter Reading
Actual
Predicted
-5
0
5
10
15
20
25
30
35
0
5000
10000
15000

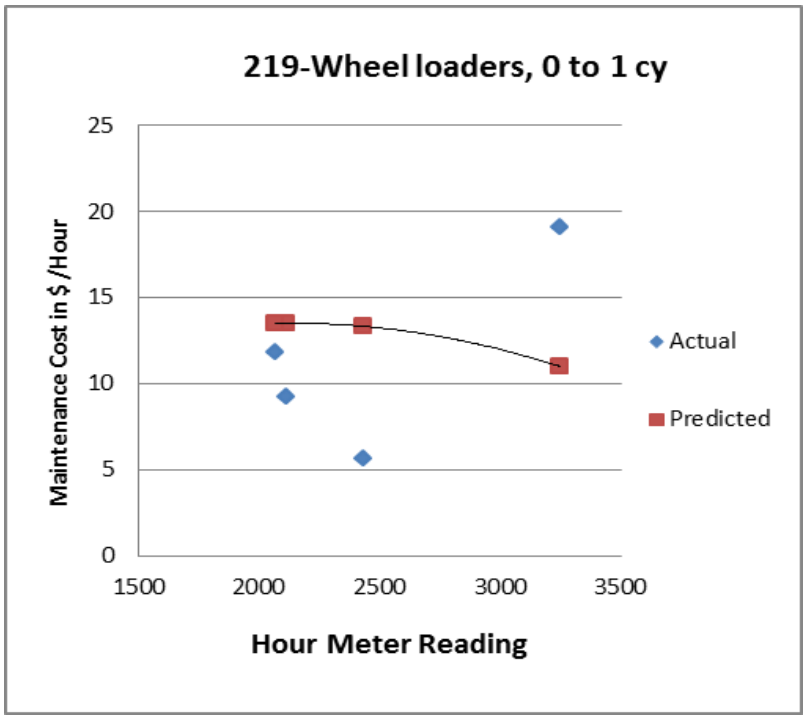

216-Vibratory rollers(doubles) 7000+kg
Maintenance Cost$/Hour
y = 7E-07x² - 0.0056x + 27.034
R² = 0.2314
Hour Meter Reading
0
10
20
30
40
50
60
70
80
90
0
2000
4000
6000
8000
10000
12000

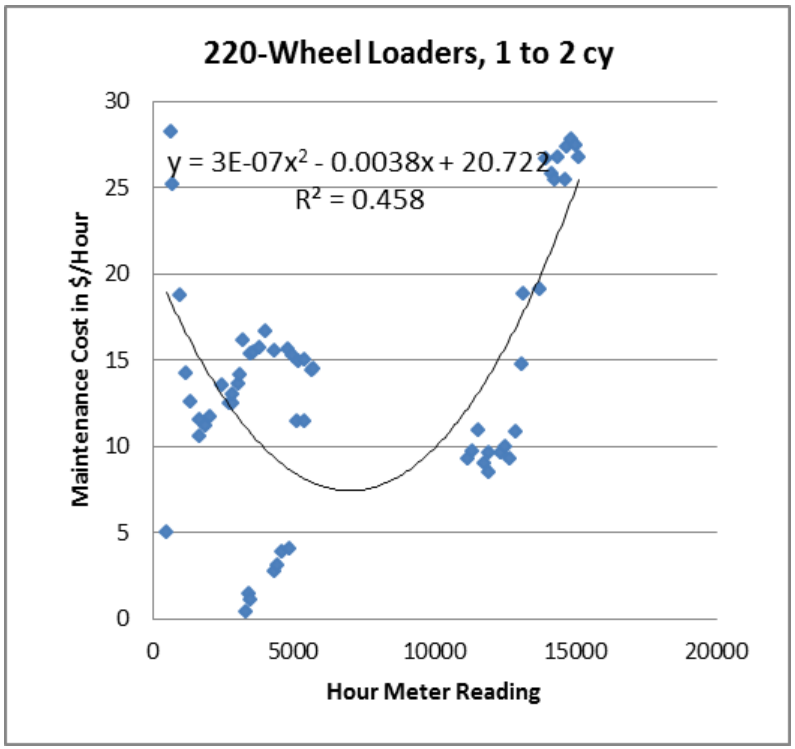

216-Vibratory rollers(doubles) 7000+kg
Maintenance Cost in $/Hour
Hour Meter Reading
Actual
Predicted
0
5
10
15
20
25
30
35
0
2000
4000
6000
8000
10000

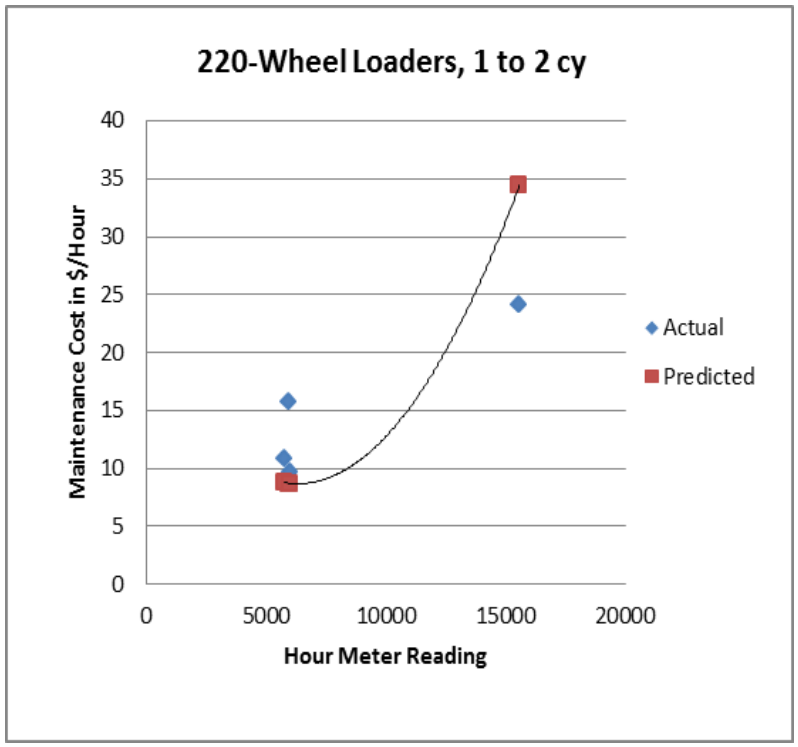

217-Vibratory roller(doubles) drum of 80+
Maintenance Cost in $/hour
y = 2E-06x² - 0.0157x + 57.766
R² = 0.054
Hour Meter Reading

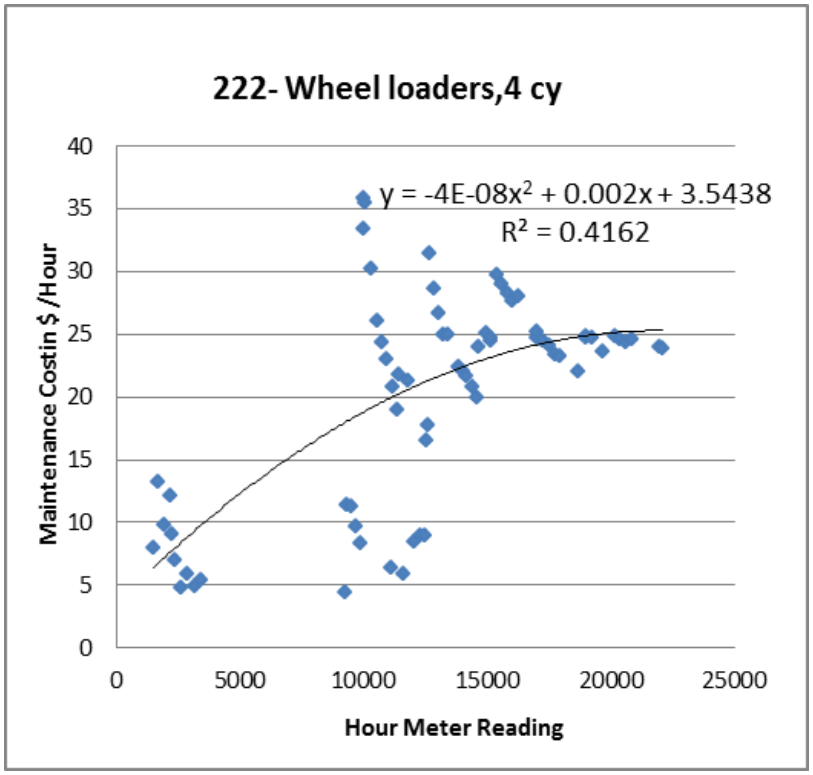

217-Vibratory roller(doubles) drum of 80+
Maintenance Cost Per Hour
Actual
Predicted
Hour Meter Reading

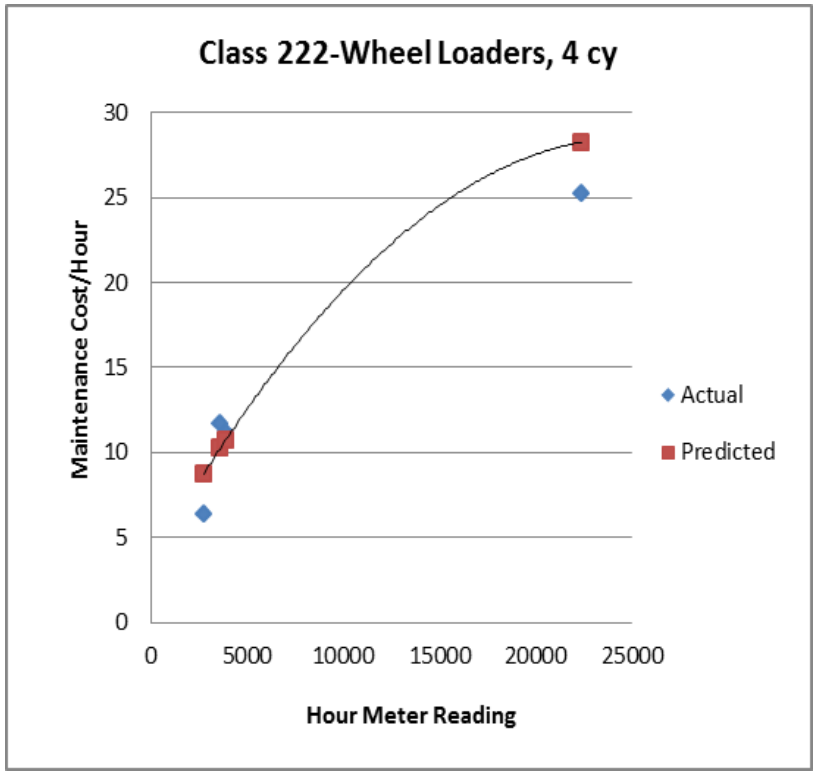

219-Wheel loaders, 0 to 1 cy
Maintenance Cost in $/Hour
y = -2E-06x² + 0.0085x + 4.4778
R² = 0.0493
Hour Meter Reading
0
5
10
15
20
25
30
35
40
0
2000
4000
6000

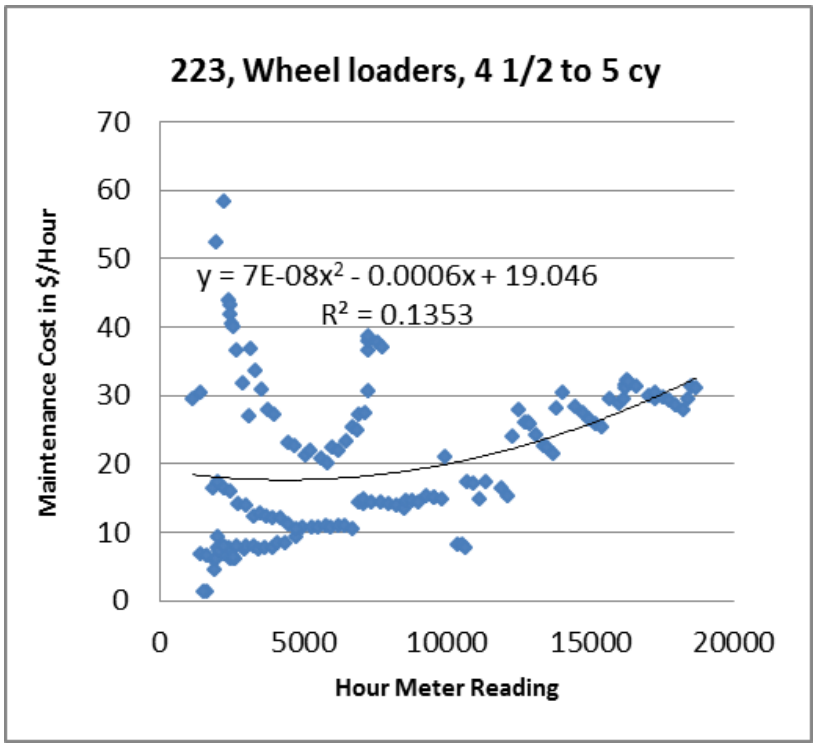

219-Wheel loaders, 0 to 1 cy
Maintenance Cost in $ /Hour
Hour Meter Reading
Actual
Predicted
0
5
10
15
20
25
1500
2000
2500
3000
3500

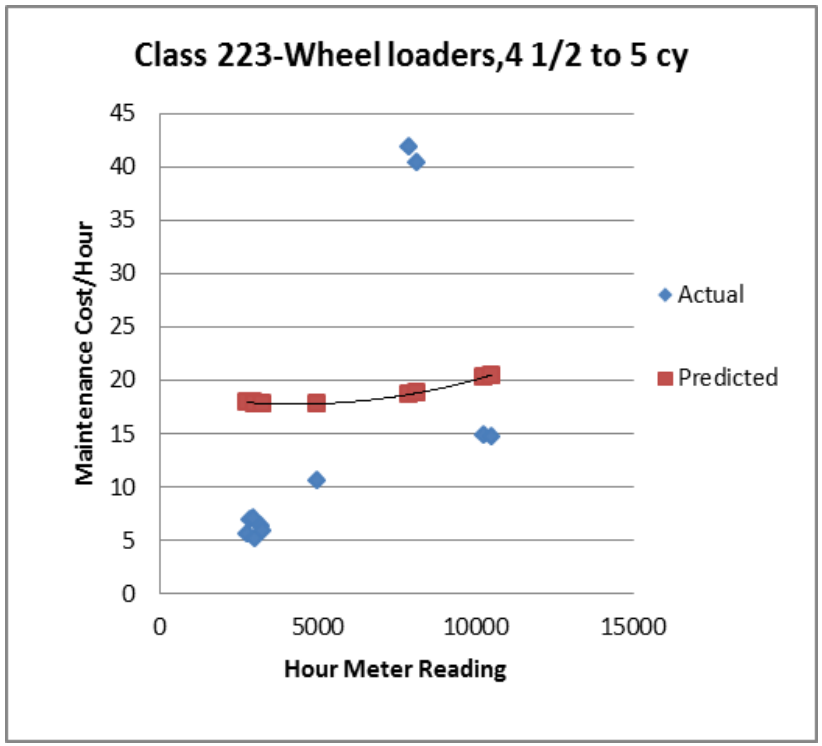

220-Wheel Loaders, 1 to 2 cy
y = 3E-07x² - 0.0038x + 20.722
R² = 0.458
Maintenance Cost in $/Hour
Hour Meter Reading
0
5
10
15
20
25
30
0
5000
10000
15000
20000

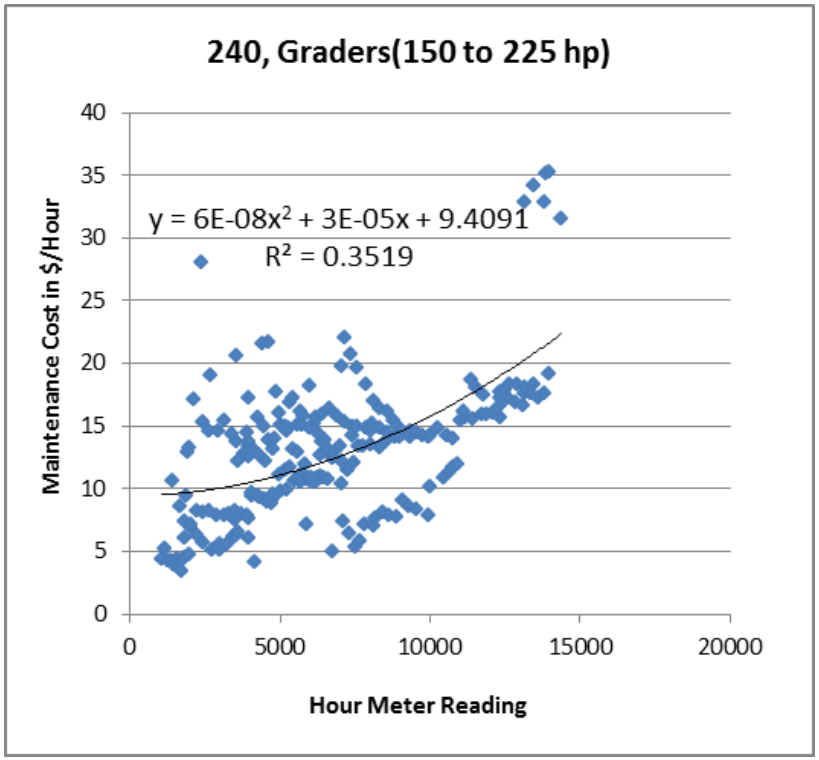

220-Wheel Loaders, 1 to 2 cy
Maintenance Cost in $/Hour
Hour Meter Reading
Actual
Predicted
0
5
10
15
20
25
30
35
40
0
5000
10000
15000
20000

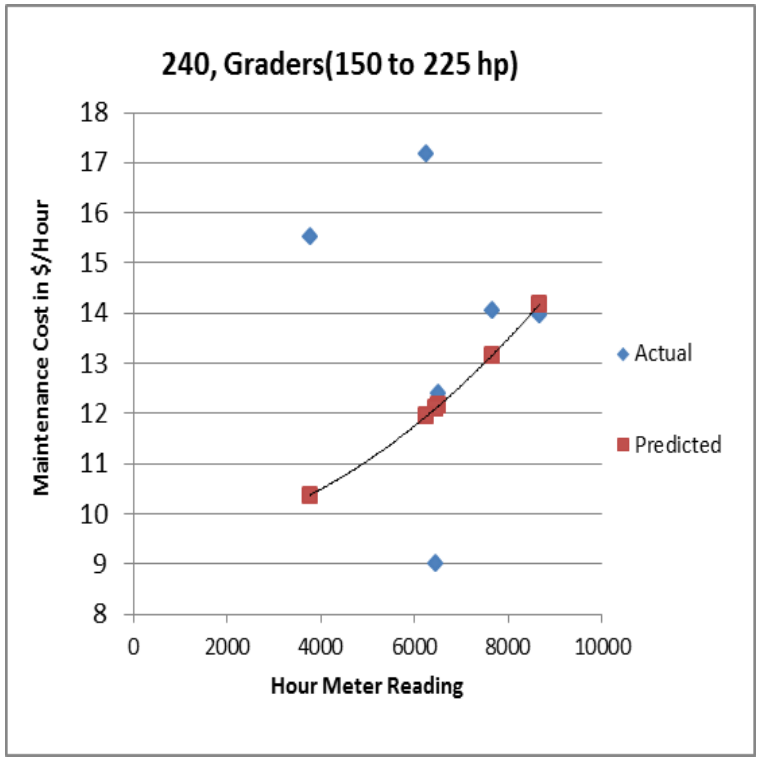

222- Wheel loaders,4 cy
y = -4E-08x² + 0.002x + 3.5438
R² = 0.4162
Maintenance Costin $ /Hour
Hour Meter Reading
0
5
10
15
20
25
30
35
40
0
5000
10000
15000
20000
25000

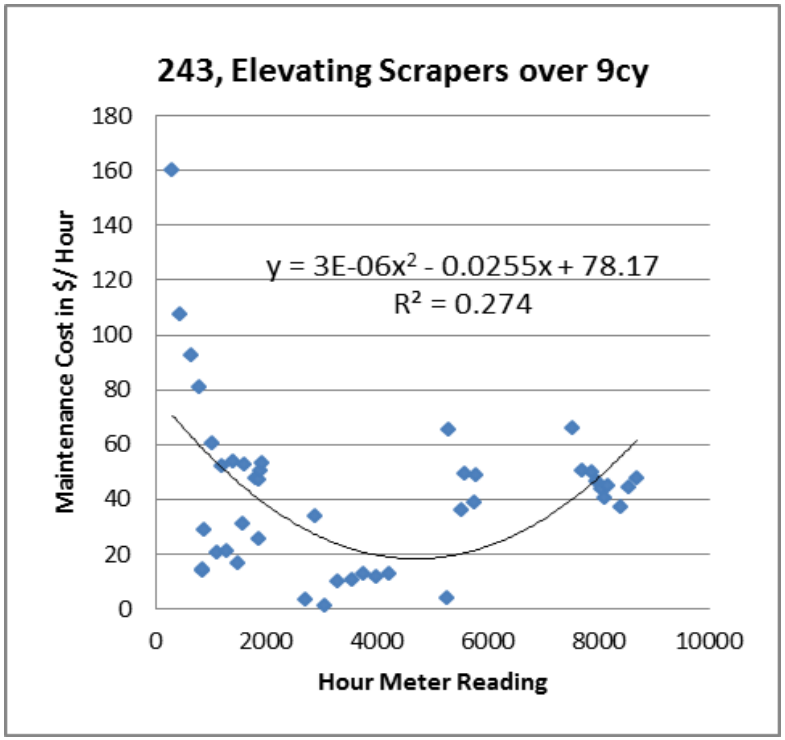

Class 222-Wheel Loaders, 4 cy
Maintenance Cost/Hour
Hour Meter Reading
Actual
Predicted
0
5
10
15
20
25
30
0
5000
10000
15000
20000
25000

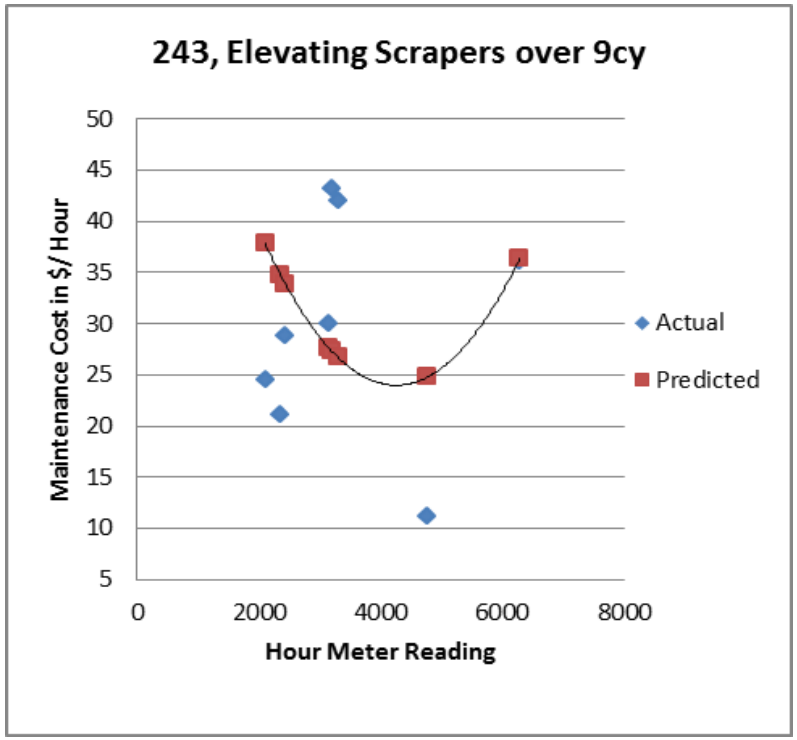

223, Wheel loaders, 4 1/2 to 5 cy
Maintenance Cost in $/Hour
70
60
50
40
30
20
10
0
y = 7E-08x² - 0.0006x + 19.046
R² = 0.1353
0
5000
10000
15000
20000
Hour Meter Reading

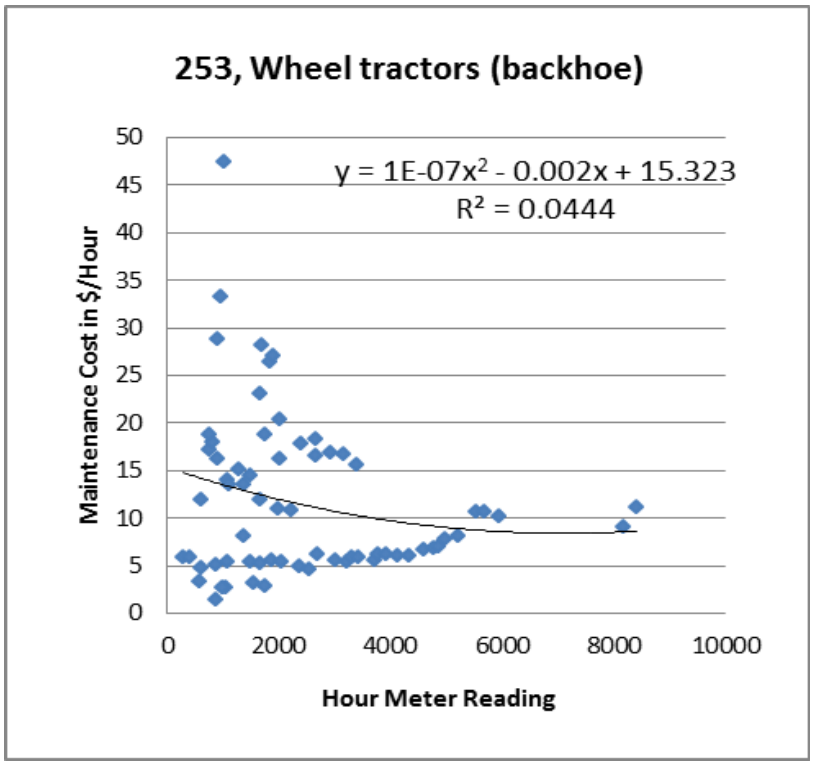

Class 223-Wheel loaders,4 1/2 to 5 cy
Maintenance Cost/Hour
45
40
35
30
25
20
15
10
5
0
0
5000
10000
15000
Hour Meter Reading
Actual
Predicted

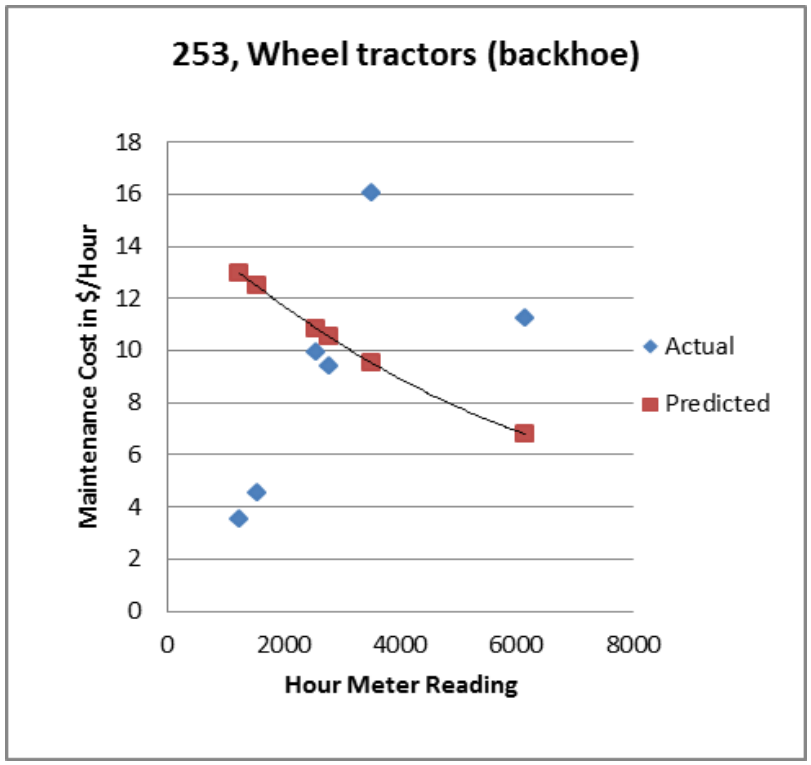

240, Graders(150 to 225 hp)
y = 6E-08x² + 3E-05x + 9.4091
R² = 0.3519
Maintenance Cost in $/Hour
Hour Meter Reading
0
5
10
15
20
25
30
35
40
0
5000
10000
15000
20000

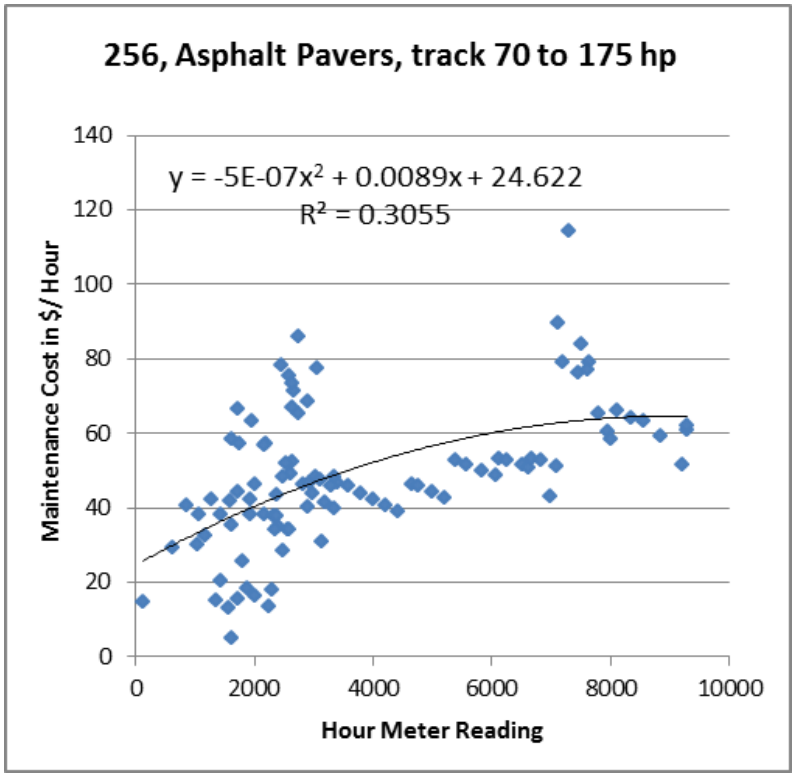

240, Graders(150 to 225 hp)
Maintenance Cost in $/Hour
Hour Meter Reading
Actual
Predicted
8
9
10
11
12
13
14
15
16
17
18
0
2000
4000
6000
8000
10000

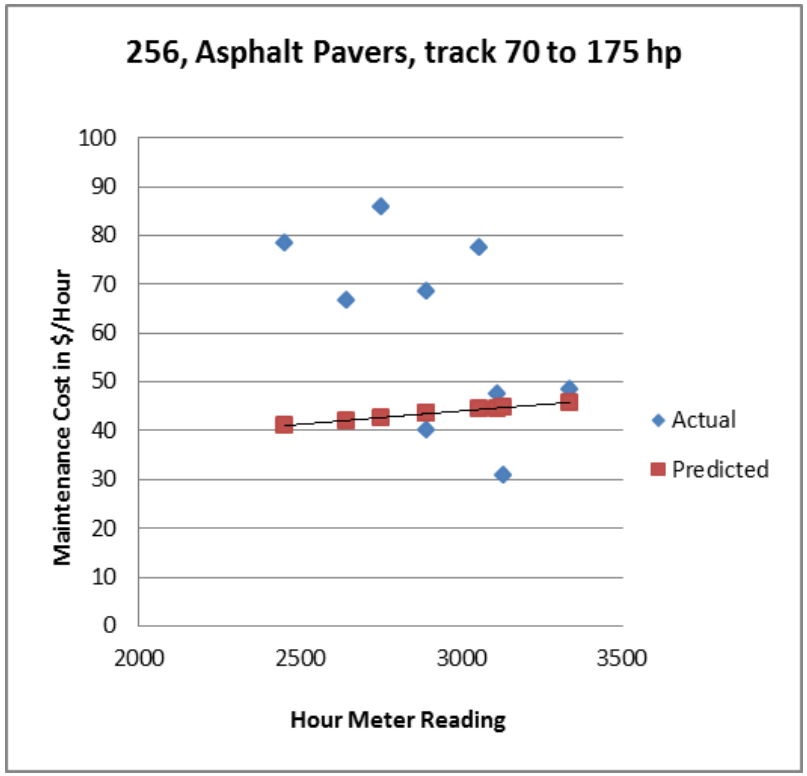

243, Elevating Scrapers over 9cy
Maintenance Cost in $/ Hour
180
160
140
120
100
80
60
40
20
0
y = 3E-06x² - 0.0255x + 78.17
R² = 0.274
0
2000
4000
6000
8000
10000
Hour Meter Reading

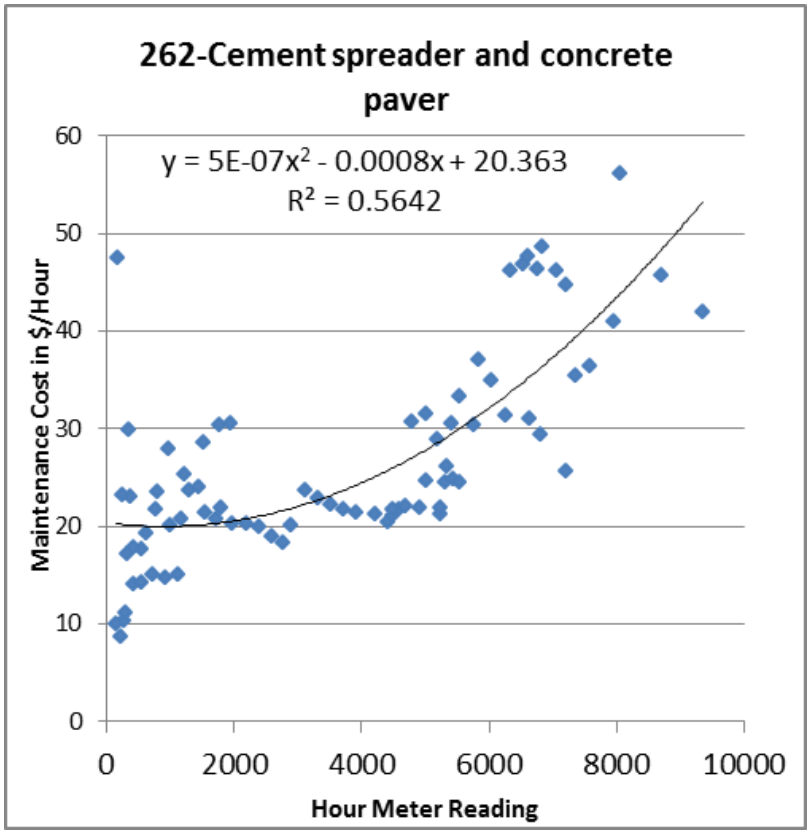

243, Elevating Scrapers over 9cy
Maintenance Cost in $/ Hour
50
45
40
35
30
25
20
15
10
5
0
2000
4000
6000
8000
Hour Meter Reading
Actual
Predicted

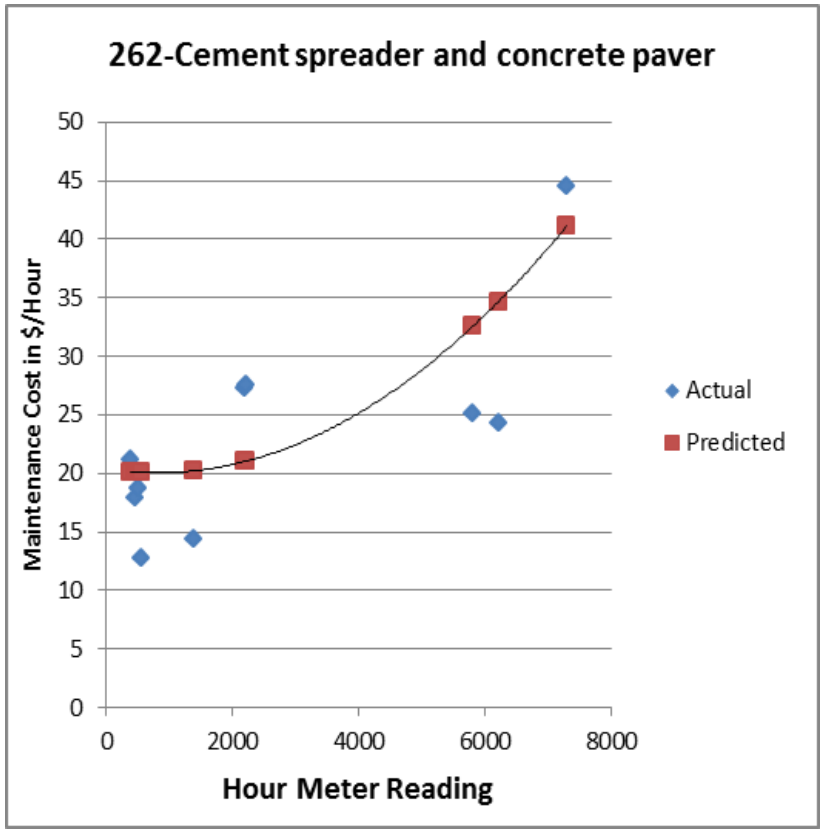

253, Wheel tractors (backhoe)
y = 1E-07x² - 0.002x + 15.323
R² = 0.0444
Maintenance Cost in $/Hour
Hour Meter Reading
0
5
10
15
20
25
30
35
40
45
50
0
2000
4000
6000
8000
10000

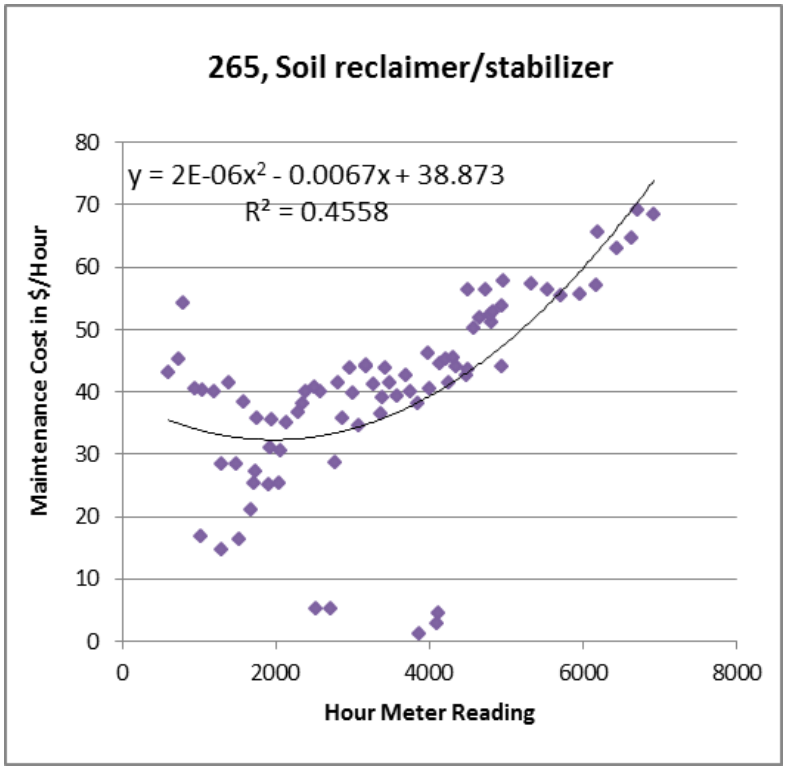

253, Wheel tractors (backhoe)
Maintenance Cost in $/Hour
Hour Meter Reading
Actual
Predicted
0
2
4
6
8
10
12
14
16
18
0
2000
4000
6000
8000

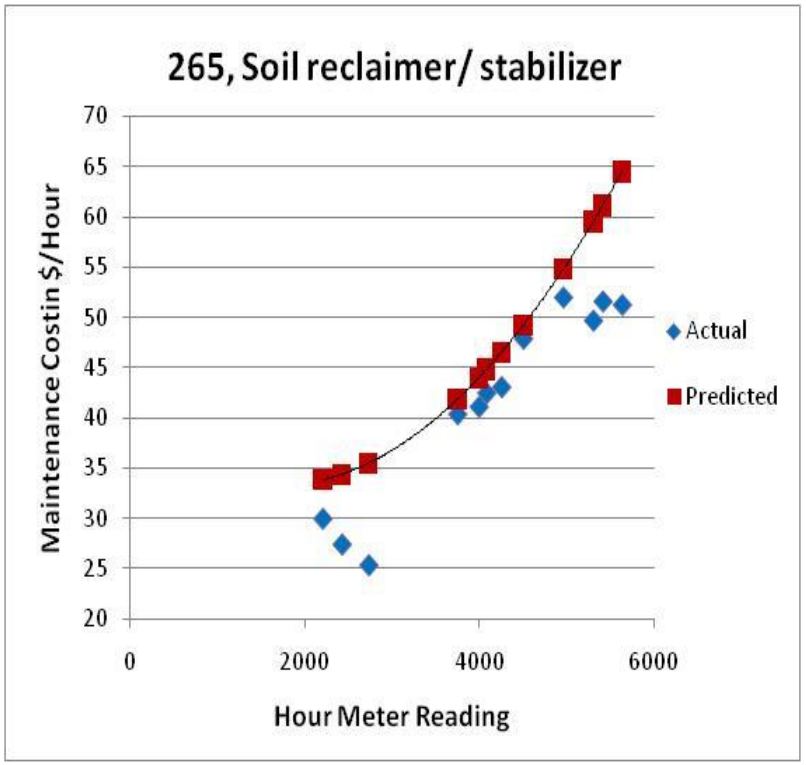

256, Asphalt Pavers, track 70 to 175 hp
y = -5E-07x² + 0.0089x + 24.622
R² = 0.3055
Maintenance Cost in $/ Hour
Hour Meter Reading
0
20
40
60
80
100
120
140
0
2000
4000
6000
8000
10000


256, Asphalt Pavers, track 70 to 175 hp
Maintenance Cost in $/Hour
Hour Meter Reading
Actual
Predicted
0
10
20
30
40
50
60
70
80
90
100
2000
2500
3000
3500


262-Cement spreader and concrete paver
y = 5E-07x² - 0.0008x + 20.363
R² = 0.5642
Maintenance Cost in $/Hour
Hour Meter Reading
0
10
20
30
40
50
60
0
2000
4000
6000
8000
10000


262-Cement spreader and concrete paver
Maintenance Cost in $/Hour
Hour Meter Reading
Actual
Predicted
0
5
10
15
20
25
30
35
40
45
50
0
2000
4000
6000
8000


265, Soil reclaimer/stabilizer
y = 2E-06x² - 0.0067x + 38.873
R² = 0.4558
Maintenance Cost in $/Hour
Hour Meter Reading
0
10
20
30
40
50
60
70
80
0
2000
4000
6000
8000


265, Soil reclaimer/ stabilizer
Maintenance Costin $/Hour
Hour Meter Reading
Actual
Predicted
20
25
30
35
40
45
50
55
60
65
70
0
2000
4000
6000

Appendix 4-Comparison or errors and correlation coefficient for all available equipment classes between class numbers 200 and 299

Class 202-Tire Compactor, 100+ hp

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative absolute error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	6.26	7.80	88.3	92.27	0.37
Least Median Square	1.72	2.12	17.5	21.4	0.82
Linear Regression	3.01	3.13	30.6	31.7	.81
Conjunctive Rule	6.84	6.9	69.5	69.8	0
Decision Stump	6.82	6.88	17.5	21.5	0.82
M5Rule	3.65	3.75	37.0	37.9	0.80
REP Tree	3.79	4.41	38.5	44.6	0.23
Multilayer Perception	4.28	5.89	48.36	60.2	0.82

Class 205-Tandem Roller, less than 5 ton

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	3.23	4.56	98.37	100.89	-0.03
Least Median Square	1.55	1.61	47.3	35.6	0.9
Linear Regression	2.04	2.31	62.2	51.2	0.91
Conjunctive Rule	2.27	3.62	69.2	80.1	0
Decision Stump	2.24	3.24	68.4	71.7	0
M5Rule	2.04	2.31	62.1	51.2	0.91
REP Tree	1.04	1.365	31.9	30.2	0.99
Multilayer Perception	2.84	3.82	86.26	84.64	0.94

Class 213-Vibratory Compactor, 50+ hp

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	6.85	7.77	110.3	107.1	0.134
Least Median Square	3.27	5.72	52.8	78.8	0.61
Linear Regression	4.03	5.86	65.0	80.8	0.59
Conjunctive Rule	3.72	6.84	60.0	94.2	0.33
Decision Stump	3.76	6.81	60.5	93.8	0.33
M5Rule	4.04	5.86	65.0	80.8	0.59
REP Tree	3.96	5.79	63.7	79.8	0.57
Multilayer Perception	3.17	4.6	51.08	64.37	0.75

Class 216- Vibratory Rollers (doubles), 7000+ kg

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	8.23	8.73	74.86	72.43	0.39
Least Median Square	3.35	4.61	29.5	37.5	0.66
Linear Regression	4.49	6.06	39.6	49.2	0.53
Conjunctive Rule	9.24	15.47	81.4	125.6	0.04
Decision Stump	5.37	6.39	47.3	51.9	0.04
M5Rule	3.96	4.59	34.9	37.3	0.65
REP Tree	5.02	6.2	44.2	50.4	0.78
Multilayer Perception	5.3	6.56	47.04	53.36	0.69

Class 217-Vibratory Rollers (doubles) drum of 80+

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	14.37	14.92	143.00	142.12	0.62
Least Median Square	4.78	5.10	47.6	48.6	0.86
Linear Regression	4.47	5.22	44.5	49.7	0.85
Conjunctive Rule	6.96	7.95	75.7	69.3	0.89
Decision Stump	8.01	8.25	79.7	78.6	0.89
M5Rule	5.12	6.13	51.01	58.4	0.89
REP Tree	8.18	9.24	81.4	88.1	0.8
Multilayer Perception	6.85	8.18	61.69	71.03	0.92

Class 219-Wheel Loaders, 0 to 1 cy

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	5.45	6.05	124.00	119.42	-0.865
Least Median Square	4.00	5.05	91.1	99.5	0.78
Linear Regression	4.09	4.19	93.1	82.6	0.98
Conjunctive Rule	2.76	3.21	62.9	63.4	0.9
Decision Stump	2.56	2.82	58.3	55.5	0.9
M5Rule	2.82	3.49	64.3	68.8	0.97
REP Tree	2.41	2.76	54.9	54.3	0.89
Multilayer Perception	2.95	3.38	67.12	66.68	0.88

Class 220-Wheel Loaders, 1 to 2 cy

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	6.31	8.41	130.88	146.2	0.92
Least Median Square	5.96	7.21	123.7	125.3	0.69
Linear Regression	2.06	3.05	42.8	53.1	0.89
Conjunctive Rule	2.04	2.31	42.5	40.2	0.92
Decision Stump	1.97	2.31	40.94	40.27	0.92
M5Rule	4.68	5.41	97.2	94.1	0.99
REP Tree	4.16	4.53	86.4	78.8	0.99
Multilayer Perception	4.69	5.07	97.35	88.13	0.69

Class 222-Wheel Loaders, 4 cy

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	1.81	2.1	19.96	21.58	0.98
Least Median Square	2.42	2.67	26.8	28.1	0.96
Linear Regression	2.94	3.3	32.6	34.6	0.96
Conjunctive Rule	2.55	2.56	28.3	26.9	0.96
Decision Stump	2.5	2.56	27.6	26.9	0.96
M5Rule	1.5	1.71	16.6	17.9	0.97
REP Tree	2.22	2.41	24.6	25.3	0.96
Multilayer Perception	2.88	3.37	31.89	35.32	0.97

Class 223-Wheel Loaders, 4 ½ to 5 cy

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	12.1	13.27	90.2	93.17	0.36
Least Median Square	5.56	6.98	41.5	49.0	0.98
Linear Regression	3.7	4.95	27.6	34.8	0.99
Conjunctive Rule	5.79	6.82	43.2	47.9	0.97
Decision Stump	5.81	6.87	43.3	48.3	0.97
M5Rule	5.75	6.1	42.9	42.6	0.95
REP Tree	4.29	4.52	32.01	31.7	0.97
Multilayer Perception	4.04	4.83	30.06	33.94	0.97

Class 240- Graders (150 to 225 hp)

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	3.4	4.4	148.4	150.45	-0.45
Least Median Square	3.31	4.52	143.58	154.6	-0.53
Linear Regression	3.85	4.58	166.96	156.5	-0.39
Conjunctive Rule	3.01	3.54	130.4	121.3	0
Decision Stump	2.84	3.38	123.4	115.6	0
M5Rule	2.17	3.06	94.3	104.5	0.0324
REP Tree	2.85	3.38	123.38	115.6	0
Multilayer Perception	7.01	9.08	212.36	215.47	0.25

Class 243- Elevating Scrapers over 9 cy

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	9.95	11.54	78.94	72.76	-0.06
Least Median Square	19.79	22.29	157.0	140.5	.01
Linear Regression	16.5	18.87	130.95	118.98	0.48
Conjunctive Rule	9.7	11.96	77.0	75.45	0
Decision Stump	9.45	11.45	74.97	72.2	0
M5Rule	8.46	11.09	67.16	69.9	0.29
REP Tree	13.69	15.89	108.65	100.2	0.24
Multilayer Perception	13.2	15.1	104.7	95.13	0.72

Class 253-Wheel Tractors (backhoe)

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	5.05	6.0	124.5	121.26	-0.73
Least Median Square	6.52	9.01	160.5	162.1	-0.37
Linear Regression	4.82	6.09	118.68	123.17	0.29
Conjunctive Rule	3.31	4.28	81.61	86.5	0
Decision Stump	3.28	4.22	80.9	85.25	0
M5Rule	6.96	7.62	171.43	154.1	-0.29
REP Tree	1.51	1.66	37.35	33.5	0.97
Multilayer Perception	6.35	6.92	156.51	139.86	0.76

Class 256-Asphalt Pavers, track 70 to 175 hp

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	20.66	25.42	112.75	113.71	-0.61
Least Median Square	16.01	18.34	87.4	82.03	0.01
Linear Regression	21.98	26.91	119.92	120.37	-0.61
Conjunctive Rule	26.69	29.99	145.67	134.16	-0.67
Decision Stump	21.25	25.8	115.96	115.4	0
M5Rule	21.98	26.91	119.93	120.37	-0.61
REP Tree	18.32	22.35	100	100	0
Multilayer Perception	15.15	17.90	82.66	80.07	0.82

Class 262- Cement Spreader and Concrete Paver

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	5.19	5.94	72.59	64.92	0.76
Least Median Square	21.37	30.35	299.2	331.7	-0.29
Linear Regression	4.21	4.79	58.9	52.4	0.83
Conjunctive Rule	7.23	8.23	101.25	89.97	0.6
Decision Stump	7.07	8.42	99.07	92.01	0.61
M5Rule	10.48	14.59	146.81	159.47	0.63
REP Tree	5.82	6.59	81.56	71.98	0.79
Multilayer Perception	4.49	5.17	62.86	56.5	0.85

Class 265- Soil Reclaimer/Stabilizer

Algorithms	Mean Absolute Error	Root Mean Squared Error	Relative absolute Error	Root Relative Squared Error	Correlation Coefficient
Second Order Nonlinear Regression	5.72	6.85	72.91	75.09	0.92
Least Median Square	4.71	5.44	60.1	59.7	0.95
Linear Regression	6.18	6.82	78.7	74.7	0.95
Conjunctive Rule	7.81	8.12	99.4	89.05	0.79
Decision Stump	6.95	7.22	88.6	79.15	0.79
M5Rule	3.26	4.11	48.9	49.9	0.96
REP Tree	6.43	7.27	81.9	79.7	0.89
Multilayer Perception	4.53	5.08	57.73	55.76	0.93

Appendix 5- List of utilized equipment classes in data mining analysis

NO	Equipment Class Number	Name of the Equipment Classes
1	Class 202	Tire Compactor, 100+ hp
2	Class 205	Tandem Roller, less than 5 ton
3	Class 213	Vibratory Compactor, 50+ hp
4	Class 216	Vibratory Rollers (doubles), 7000+ kg
5	Class 217	Vibratory Rollers (doubles) drum of 80+
6	Class 219	Wheel Loaders, 0 to 1 cy
7	Class 220	Wheel Loaders, 1 to 2 cy
8	Class 222	Wheel Loaders, 4 cy
9	Class 223	Wheel Loaders, 4 ½ to 5 cy
10	Class 240	Graders (150 to 225 hp)
11	Class 243	Elevating Scrapers Over 9 cy
12	Class 253	Wheel Tractors (backhoe)
13	Class 256	Asphalt Pavers, Track 70 to 175 hp
14	Class 262	Cement Spreader and Concrete Paver
15	Class 265	Soil Reclaimer/Stabilizer