Maintenance Cost and Residual Value Prediction of Heavy Construction Equipment

by

Yi Zong

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Construction Engineering and Management

Department of Civil and Environmental Engineering

University of Alberta

© Yi Zong, 2017

Abstract

Equipment cost represents a large expenditure for construction companies. Making economic decisions such as when to replace or rebuild is a complicated and difficult task that often relies on the experience of an equipment manager. This research analyzed the historical maintenance cost of heavy construction equipment to simulate and predict the maintenance cost of different types of machines. Based on historical maintenance data of 15 fleets, including 250 heavy machinery units from a construction company, regression models were developed for each type of machine to compare the relationship between the maintenance cost and machine age. A second-order polynomial expression of the Cumulative Cost Model developed by Mitchell (1998) was used to identify optimum economic decisions such as replacement and retirement. As equipment owning companies often design maintenance policies according to specific operating hour intervals, different datasets based on varying service meter reading (SMR) intervals were created to provide equipment managers with different equations, thereby providing a guideline to help the company revise maintenance policies for each type of machine. Statistical analyses were conducted for each dataset and it was found that the best regression model performance was obtained at 500 and 1000 SMR intervals. Residual plots indicate that the models can be improved by including other variables despite the high R^2 values.

Besides, the residual value of heavy construction equipment is of great significance for equipment owning companies. Many factors such as the manufacturer, model, machine age, operating hours, and even macroeconomic indicators might have direct or indirect impacts on

the price of machines in an auction market. In current practice, machine-owning companies use rule-of-thumb opinions or single-linear functions to make predictions, providing a very rough estimate to decision makers. This study considers the current state of knowledge on residual value estimation for used heavy construction equipment and introduces two effective data mining methods, k nearest neighbor (KNN) and random forest (RF) with comparisons to a single regression tree. The proposed methods are exemplified based on a dataset of articulated trucks. Equipment specifications are considered as predictive features, and a feature selection algorithm is implemented to provide a rank of different factors. Distinct models are built after multiple runs and cross validation. Compared to the single regression tree method which has been studied by other researchers, the KNN and RF methods demonstrated better performances in terms of accuracy as well as running time.

Acknowledgement

During my two years of graduate study, I not only learned a lot as a graduate student, but had opportunities to work with and learn from professionals. I would like to give my sincere thanks to those who helped and guided me in my research and life to complete my M.Sc degree in Construction Engineering and Management.

First, I would like to thank my supervisor Dr. Yasser Mohamed. He provided me with a precious opportunity to work at a professional construction company, where I could deal with construction problems in reality. Inspired by Dr. Mohamed's advice, I also learnt data mining technology and python programming. I consider myself so lucky to have had such a helpful and supportive supervisor. I would also like to give my thanks to those professionals including Oscar Gomez, Lee Channing, Coco Lu and Louwrens De Klerk in the construction company. When I had problems with data collection and interpretation, I could always get practical advice from them.

Last but not least, I would like to thank my supportive family and friends. I had such a wonderful life here in Edmonton and feel so comfortable of being a part of your community.

Table of Contents

List of Tables

List of Figures

Abbreviations

CC	Correlation Coefficient
CCM	Cumulative Cost Model
FACT	Financial Analysis and Control Tool
KNN	K Nearest Neighbor
MAE	Mean Absolute Error
PM	Preventive Maintenance
RAE	Relative Absolute Error
RF	Random Forest
RMSE	Root Mean Absolute Error
RRSE	Root Relative Squared Error
SMR	Service Meter Reading (Unit: hours)
WEKA	Waikato Environment for Knowledge Analysis

Chapter 1: Introduction

Equipment management is a difficult and complex process that influences almost every aspect of a company's operations. For a construction company to be financially healthy and competitive, it is important for it to place a priority on equipment acquisition and disposal. Decisions about acquisition and disposal not only require knowledgeable and experienced equipment management, but also good awareness of the equipment market. In the area of heavy construction equipment management, the greatest concerns are maintenance cost and residual value.

1.1 Maintenance Cost

Equipment management is an important yet difficult task for contractors as well as equipment-owning companies, especially for those engaged in extensive equipment use. Maintenance commences after equipment is purchased, and maintenance costs account for most of the cost over the life span of equipment. Maintenance, as defined by Geraerds (1983), is "all activities aimed at keeping an item in, or restoring it to, the physical state considered necessary for the fulfillment of its production function." Maintenance of heavy equipment includes a number of activities such as preventive maintenance, running repairs, fuel and tires, etc. Peurifoy (2006) pointed out that the cost of repairs normally comprises the largest component of machine cost, and it generally accounts for approximately 40% of the machine cost over its life span. Repair

costs associated with labor and parts, which are difficult to estimate, make up between 15% to 20% of equipment budget (Vorster, 2009). Furthermore, maintenance costs can vary depending on work conditions, operator skills, and a company's maintenance strategies. It is challenging, therefore, to estimate the cost of owning and operating equipment. Modeling maintenance costs can simulate and reveal the dynamic trend of equipment value, laying a foundation upon which economic equipment managing decisions can be made.

Vorster (1980) proposed the Cumulative Cost Model (CCM), to describe the dynamic behavior of equipment maintenance costs. The CCM provides both a numeric and graphical illustration of equipment costs, thereby aiding in the planning of economic decisions for equipment managers. Based on data consisting of 270 construction machines, Mitchell Jr (1998) evaluated 19 different linear and transformed non-linear models, and noted that a second-order polynomial expression best described equipment value. The maintenance cost research of this thesis focuses on heavy construction and mining machines, including heavy rigid frame trucks (up to 320 T), excavators (up to 360T), and shovels (up to 800T).

1.2 Residual Value

It is expected that the global construction industry will continue to grow over the next decade. The total revenue of the global construction, farm and heavy machinery market expanded on a compound annual growth of 6.9% between 2010 and 2014 (Global Construction, Farm & Heavy Machinery Industry Profile, 2014). Between 2011 and 2014, the value of construction in United

States has been rising at a rate of 7.3% annually, which causes an increase in domestic demand of construction machinery (Outlaw & Young, 2014). According to Perspectives and Economics (2011), construction in emerging markets is expected to double within a decade, becoming a $6.7 trillion business by 2020. While construction management is a combined task involving many resources, equipment is vital to the success of construction projects. In practice, contractors have to spend a large portion of expenditure on owning and operating costs of heavy construction equipment. Financially, equipment-intensive projects always present great potential risks in terms of routine repairs, major rebuilds, and unexpected accidents.

Among the financial challenges of owning and operating construction equipment, depreciation is one of the most significant, especially when purchasing and selling used machines. Many equipment managers, however, circumvent residual market value by being overly conservative and assuming a zero residual value of machines when calculating depreciation (Vorster, 2004). The residual value of a piece of heavy equipment can be defined as the price that can be achieved by "disposing of a used machine in a fair transaction between an equally well informed buyer and seller in the overall market with its particular economic situation" (Lucko, Vorster, & Anderson-Cook, 2007). Depreciation models in the accounting field such as sum-of-years or average annual investment methods (Peurifoy, 2006) can provide an approach for calculating residual value, but it is an uncertain number that depends on not only the unique individual situation of equipment, but economic environments of the specific markets. Lucko and Mitchell Jr (2010) concluded it is not an actual realized dollar amount, but a forecasted value determined

by taking into account many independent variables of the machine itself as well as the economic environment.

Although it is almost impossible to develop a completely accurate system to predict the residual value of heavy construction equipment, many scholars made great efforts to develop predictive models that are competitive and close to reality. The regression model for predicting residual value of equipment, which was widely used in the agricultural industry (T. L. Cross & Perry, 1995; Fairbanks, Larson, & Chung, 1971; McNeill, 1979) was introduced into construction industry by Lucko (2003), based on auction records retrieved from an online construction equipment database. A spatial hedonic price function, proposed by Ponnaluru, Marsh, and Brady (2012), explains the variability in auction price as a function of heavy machines' characteristics, and can be used to amend regression formulas. Fan, AbouRizk, Kim, and Zaïane (2008) discussed the feasibility and effectiveness of the regression model and proposed a data mining algorithm called AutoRegression Tree. With a set of "if-then" split conditions, a single regression tree algorithm could provide a good interpretation of residual value prediction model, yet along with discussions about issues such as the overfitting problem. In machine learning, overfitting is a phenomenon that the model "memorized" details of training set but is less accurate for generalizing new examples for prediction. The model could be too "sensitive" to data especially for small datasets.

As a hybrid of multidisciplinary field, data mining helps people extract information implicitly stored in large datasets. In the field of construction equipment management, it is a useful

approach to help decision makers predict the residual value of heavy machines. This paper steps forward to presents a few other models to calculate residual values of heavy construction equipment with the help of different data mining techniques. Regression decision tree, instance-based learning and random forest will be used to predict the up-to-date auction records from an equipment-owning company, and comparisons will be presented and discussed.

1.3 Research objectives

The objective of this research is to develop simulation models for fleet managers or other decision makers to track and predict maintenance costs and the residual value for different categories of machines, which can help when making decisions about purchases, rebuilding, or replacement. This objective will be achieved by accomplishing the following sub-objectives:

- Find the optimum SMR intervals to build the CCM. Different datasets are created based on different SMR intervals. To simulate and predict the maintenance cost of different types of machines, cumulative cost models are applied to each dataset and statistical analysis is conducted to compare and select the optimum dataset for the model.
- Predict the market price of heavy construction machines using different methods. Besides a traditional regression tree method, k nearest neighbor and random forest are used to train and test the market price predictive model. A discussion will be included to compare the three algorithms and the best situation to use.

1.4 Research Methodology

The flow of this research is illustrated in Figure 1. In order to accomplish the proposed objectives, the following processes are followed:

- Identify and collect maintenance cost data for different types of machines, with 250 units ranging from heavy rigid frame trucks (up to 320T), excavators (up to 360T), and shovels (up to 800 T).
- Import and clean historical maintenance data before analyzing. An adjusted inflation rate is applied in order to sum values in different years.
- Evaluate the performance of CCM on various categories of heavy machines.
- Evaluate the impact of different SMR-intervals on the CCM models.
- Collect and process up-to-date auction records for heavy equipment in the North American market, and apply alternative data-mining algorithms through Waikato Environment for Knowledge Analysis (WEKA). The main purpose of this part of the research is to predict residual value of heavy equipment.
- Identify and rank factors that have a potential influence on the residual market value of heavy machines. Original auction data were re-organized categorized to grab corresponding information.
- Evaluate the performance of alternative data-mining methods and recommend the optimum method for different situations.
- Present the results and contributions that this research makes to the body of knowledge.

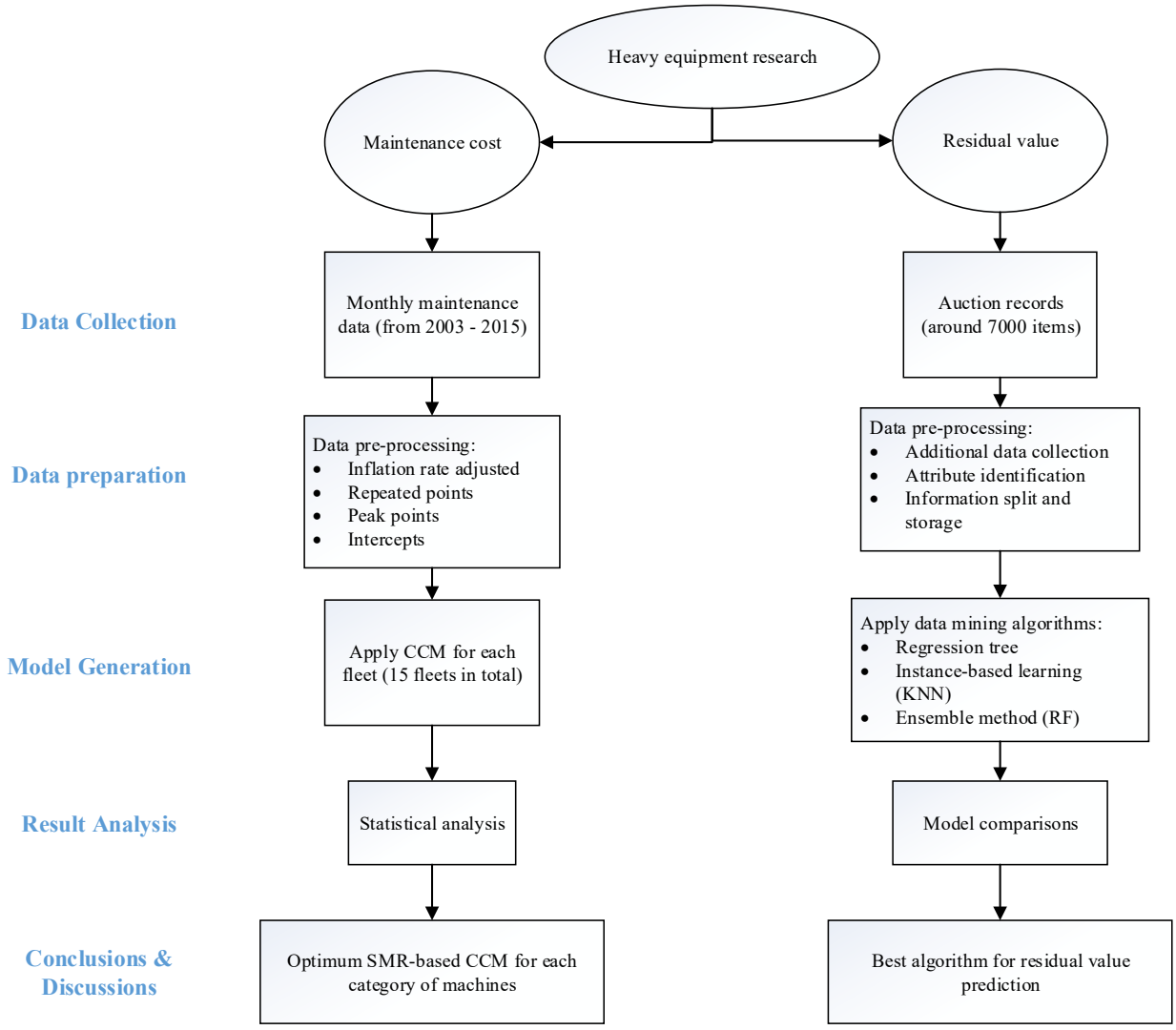


Figure 1: Research Methodology

1.5 Document Structure

- This thesis contains five chapters and several appendices. The following are brief descriptions of the contents for each chapter and appendix, respectively.
- *Chapter 1 – Introduction* introduces concepts and the scope of construction equipment research. Research objectives are presented.

- *Chapter 2 – Literature review* provides background about equipment management, data mining technologies and current practices of equipment management study.
- *Chapter 3 –Cumulative Cost Modeling of Heavy Equipment Maintenance Data* applies the CCM to different classes of heavy construction equipment. A detailed discussion is provided.
- *Chapter 4 – Residual value prediction models* presents different data mining algorithms to predict the residual market value of heavy equipment. Detailed discussion and comparisons regarding prediction error, running time etc. are provided. Recommendations to combine CCM with residual value prediction values are provided.
- *Chapter 5 – Contributions and limitations* summarizes contributions and limitations of this study. It also provides recommendations for future research.

Chapter 2: Literature review

2.1 Maintenance Cost

Decisions about heavy equipment should be made based on sound economic principles. Once the equipment needed for a particular project is identified, the cost and duration of the equipment and other necessary auxiliary instruments are planned (Valli, Jeyasehar, & Saravanan, 2013). Economic issues surrounding construction equipment have been discussed extensively in both academia and by industry manufacturers. Nunnally (2004) discussed productivity and cost of different construction procedures such as loading and hauling. Peurifoy, Schexnayder, and Shapira (2006) provided detailed discussion about repair costs and production rates of dozers, scrappers, excavators and other construction machines. Douglas (1975) explored construction equipment in greater detail from the view of the engineering economy. Vorster (1980) studied the age-cost-reliability relationship as well as the organizational aspects of managing construction equipment and proposed the cumulative cost model (CCM), which will be implemented in this study. Equipment manufacturers have also specified the general cost for hydraulic excavators in life-time analysis.

Several decades ago, these methods were mainly concerned with soil and the machine, independent of each other (Drakatos, 1975; E. Manatakis & Drakatos, 1978; Soltynski, 1968). Such soil-machine systems suggested using different equipment based on soil conditions, which is unrealistic for large scale implementation. Practically speaking, the economics of construction equipment relate to ownership and operating costs including purchase price, residual value,

economic life, repair and maintenance cost, and availability (Kannan, 2011). E. K. Manatakis and Drakatos (1993) presented a new method for the analysis of operating costs of construction equipment, specifically covering rear-dump trucks used for earth moving over a period of six years. The data collected consisted of operating hours and operating costs, which include maintenance, repairs, lubricating oil, fuel, tire repairs and personnel. This model was verified to evaluate construction equipment while also considering, as criteria, operating costs and the equipment's life period.

When considering maintenance costs, age can take the form of calendar years, age in cumulative hours of use, or age in units of production. When making a decision about a repair, it should be possible to estimate the life earned as a consequence of the repair. Vorster (1980)'s cumulative cost model (CCM) provides numerical and graphical solutions to many equipment management problems. Mitchell Jr (1998) developed a regression model by using a quadratic function that employed field data to represent repair costs based on the number of cumulative hours that a machine had been used. This expression can be incorporated into the CCM, where it can be used to identify optimum economic decisions. It can also provide construction engineers with a valuable tool for better understanding the nature of repair costs as they relate to production fleets. Mitchell, Hildreth, and Vorster (2010) attempted to use the CCM by implementing two different methodologies for calculating repair cost: life-to-date (LTD) repair costs and the period-cost-based (PCB) model. They also pointed out that heavy machines tend to require repairs as a result of use rather than simply the passage of time. Thus, tracking operating time is

of great significance to estimating repair costs. Bayzid (2014) compared LTD and PCB cumulative models when the calculating maintenance cost of equipment. With the help of data mining analysis, Bayzid developed different models for each equipment class.

Previous research in this area commonly employed a regression model by ordinary least squares (Duncan, 2015; D. J. Edwards, Holt, & Harris, 1999; Gillespie, 2004; E. K. Manatakis & Drakatos, 1993). A time series approach provides further insights into modeling maintenance costs of construction equipment. Moore (1976) found that time series has an inherent autocorrelation among observed cost series using linear regression analysis. Box and Jenkins (1976) established an autoregressive model, the Box-Jenkins method, which has become a popular way to model equipment failures based on transformed data. A methodology is presented for predicting life cycle maintenance expenditure over the useful life of tracked hydraulic excavators (David J Edwards, Holt, & Harris, 2000). The authors utilized a centered moving average to analyze the time series of the maintenance cost of construction machines and isolated its trend of changes. Besides the time series approach, Yip, Fan, and Chiang (2014) presented a comparative study on the applications of general regression neural network (GRNN) models and conventional Box–Jenkins time series models to predict the maintenance cost of construction equipment. The authors concluded that both the Box-Jenkins models and the GRNN models can be used to estimate maintenance cost time series and the forecasting of maintenance cost intervals instead of point values. Data mining technology has also been applied in equipment economic estimation (Bayzid, Mohamed, & Al-Hussein, 2016; Fan et al., 2008).

2.2 Residual value of equipment

Depreciation values the decrease of assets. There are different methods of depreciation, including straight line depreciation, double declining balance method, sum-of-years-digits methods, etc. Asset-owning companies make acquisition and disposal policies based on depreciation results. A considerable amount of research is focused on the depreciation of agricultural and forestry equipment. Bates, Rayner, and Custance (1979) introduced a simulation model to discuss how inflation adjustment affects the optimal replacement age of farm tractors. Besides using the included age as the only explanatory attribute predicting optimal replacement of farm tractors, Reid and Bradford (1983) added new features including horse power, average met farm income, manufactures, and technological change time-index. And T. L. Cross and Perry (1995) found that macroeconomic variables were significant variables for most types of agricultural machinery. A few remaining value functions were developed by T. Cross and Perry (1996) based on auction sales data including 12 types of farm equipment, and a double square root functional form was found to be the best form for modeling changes in equipment values over time. Furthermore, the American Society of Agricultural Engineers (ASCE) recommended a generalized regression formula for estimating residual percentage (Fan et al., 2008). Similar research was also conducted on logging equipment in the forestry industry by Cubbage, Burgess, and Stokes (1991).

Multi-linear regression analysis was conducted on the residual value of construction equipment. Lucko (2003) introduced the residual value of construction equipment studies from forestry and

agricultural industries. With statistical considerations for performing residual value analysis, a second-order polynomial of equipment calendar age with additive factors (manufacturer, condition rating, auction region, and macroeconomic indicators) appears good based on auction records of track dozers ranges between 100-199 horsepower (Lucko, Anderson-Cook, & Vorster, 2006). By examining different types of construction machines (i.e. articulated trucks, medium dozers, small excavators, and track loaders), intuitive contour diagrams of the total hourly cost were introduced (Lucko et al., 2007).Statistically, the residual value was found to contribute significantly to the total hourly cost. Lucko also conducted quantitative research of incongruous economic data and changing economic conditions (Lucko, 2010, 2011).

As an interdisciplinary subfield of computer science, data mining technology benefits data analysis by extracting patterns and knowledge from large amounts of data in an accurate and effective approach (U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996; U. M. Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996; Han, 2006). In the construction research field, Fan, AbouRizk, and Kim (2007) developed a general framework for building a construction equipment management decision-support system with integrated data mining, showing advantages with rule-based or statistics-based support approach. To predict residual values of heavy construction equipment, Fan et al. (2008) exemplified data mining process with autoregressive tree (ART) algorithm by selecting auction records of 8,589 wheel loaders. In comparison with the results obtained from artificial neural network (ANN) and multivariate linear regression (MLR), the authors concluded that ART model performed better. In change

management, the “generalized, unbiased, interaction, detection and estimation” (GUIDE) regression tree algorithm was introduced by M.-J. Lee, Hanna, and Loh (2004) to quantify the cumulative impacts caused by change orders, while a K Nearest Neighbor (KNN) based knowledge-sharing model was proposed by Chen (2008) to prevent unnecessary expense and loss for severe change order disputes. By constructing a multitude of decision trees, Random Forests (RF) algorithms could keep the bias low with a relatively reduced variance, and correct for the decision tree’s habit of overfitting to the training set (Friedman, Hastie, & Tibshirani, 2001; Prasad, Iverson, & Liaw, 2006). While KNN and RF algorithms have been widely used in other civil engineering area (Harvey & McBean, 2014; Zhang & Haghani, 2015; Zhou, Li, & Mitri, 2016), they have never been used to predict residual value of heavy construction equipment. This thesis discusses the performance if these two algorithms in predicting the residual value of construction equipment, and detailed comparisons are also presented.

Chapter 3: Cumulative Cost Modeling of Heavy Equipment Maintenance Data

3.1 Methodology

In order to find the relationship between the maintenance cost and machine age, a regression method was chosen to track the cumulative cost trend as the machine age increases. Equipment managers often make decisions based on fixed operating hour intervals. For example, some companies perform regular maintenance on equipment every 2000 operating hours. In this study, based on several interviews with field engineers and equipment managers, it was learnt that operating hours of heavy machines collected as field data, was sometimes roughly estimated by operator after the machine had been operated. While the records of operating hours should be accurate in the long run, they might have some "noisy" points and fluctuations which could conceal the true trend of the maintenance costs. The machine age in cumulative hours of use can be likened to the odometer readings in automobiles. In this research, the service meter reading (SMR, unit in hours) is used as the master variable in this research.

Using the field data, Mitchell Jr (1998) applied 19 regression models on 17 fleets of construction equipment and found that the cumulative cost model (CCM) was best suited for economic decision making within the equipment environment. The CCM, first proposed by Vorster (1980), proved itself to be very helpful in making economic decisions. It provides not only a valid numerical solution to equipment management issues, but also an intuitive graphical depiction of the problem being analyzed. With the help of CCM, it is possible to describe and understand changes in total costs, average costs, and marginal costs (Mitchell Jr, 1998). Figure 1 shows a

geometric representation of the CCM.

Figure 2: The Cumulative Cost Model (Vorster, 1980; Mitchell Jr, 1998)

In this model, equipment age is used as the abscissa. The cumulative cost, which is normally expressed as sum of all transactions to date, was the ordinate. All owning and operating costs can be depicted in the CCM (Mitchell Jr, 1998). As noted earlier, SMR is used to indicate machine age. The cumulative cost curve originates at the cumulative cost representing the initial purchase price of the machine. A straight line is drawn directly from the origin to the point on the cumulative cost curve. The slope of this line, T_t, graphically represents the average cost during a given time. The optimum economic life, L*, is defined by a geometric tangent from the origin to the cumulative cost curve. T* can be found the slope of this tangent line, representing the optimum average cost. (See Figure 2).

Mitchell Jr (1998) organized cumulative cost data into four data sets, including all but repeated points, 500-hour intervals, average of 500-hour intervals, and final data points, leading the author to conclude that a 500-hour interval data set is the optimum data set that should be used in CCM. Using field data from 15 fleets, nearly 250 heavy machines, this study developed 11 different

data sets based on different SMR intervals. Raw data was cumulatively calculated by season, semi-annually, annually, and at 500, 1000, 1500, 2000, 2500, 5000, 7500 and 10,000 hour interval. Statistical analysis was conducted on every data set and the optimum one was selected for specific machine models, as described in the following sections. Equation 3-1 below, represents the regression models of maintenance cost of heavy machines developed in the aforementioned research (Mitchell Jr, 1998; Mitchell et al., 2010; Vorster, 2009):

$$\boldsymbol{CC = A \times H_w + B \times H_w^2} \qquad \textbf{[3-1]}$$

CC – cumulative maintenance cost for the heavy machine

Hw – life-to-date hours (SMR) worked by the machine

A – coefficient that measures the rate at which the total cost increases with age

B – coefficient that measures the curvature as opposed to the slope of the line

In this thesis, a few assumptions of CCM need to be clarified. To begin with, the weather of where the equipment was operated was assumed to be normal and consistent. Secondly, historical maintenance data were assumed to be accurate and human errors were excluded. Finally, skills of equipment operators were assumed to be average and same.

3.2 Data Source

This study consists of records stored in company's accounting system. The downside of field data is that it can contain "noisy" points, which might distort the reliability of records associated with a particular machine. It is assumed that the more machines that are included in this study,

the less influence that these distortions will have. This research contains 250 heavy machines with 19 models for different types of machines. The cumulative cost model was developed for each type of machine.

Mitchell Jr (1998) indicated that data would pass through multiple hands before entering into the accounting databases, which might result in inconsistency. The company had its own data collection procedures, which have been studied to validate the accuracy of the collected data. The company uses the Financial Analysis and Control Tool (FACT) as its accounting report software. For this research, all the necessary field data was obtained through FACT inquiries. The data extraction process, FACT Finder, has minimal impact on the production accounting system, and only reads data from the database. In other words, FACT users cannot write any data or make any changes to the database. The data selection is intuitive and powerful, using attribute information within the accounting system to organize the data for filtering, ordering and pagination. Reports are generated using predefined layouts; for instance, trial balances, balance sheets, income statements, job profit reports, equipment cost reports, general and administrative cost reports.

The company tracked SMR increases related to actual hours of usage for an individual machine. Maintenance costs were separated and categorized in different accounts. Figure 3 shows a breakdown of equipment maintenance costs. In the company's accounting database, maintenance costs can be divided into running repairs, undercarriage, ground engaging tools, preventive maintenance (PM) services, and tires. In each of these categories except for tires, there are

several sub-accounts which contain parts, labor, sub-contractors, transportation and others. The costs associated with equipment maintenance and repairs were separated in this research. The component that is most pertinent to maintenance considerations for this study is running repairs. Tires, undercarriage components, ground engaging tools, and preventive maintenance share a common characteristic in that they do not increase with machine age, and have much shorter lives than the equipment with which they are associate (Mitchell Jr, 1998; Vorster, 2009). The goal of this research is to determine whether it is economical to buy, sell, or make other economic decisions (e.g. rebuild, replacement, etc.) at a certain time. However, most of other costs are dependent on not only machine age, but on many other factors. Taking all these operation costs into consideration will affect the trend of running repairs as the machine ages. Mitchell Jr (1998) points out that "the useful lives of these 'expendables' are highly dependent upon local conditions and operator skills". However, this is not true for repair parts and labor wherein an increase in machine age is the single most significant factor in determining how long a machine should be kept (Vorster, 2009). As a result, this research focuses on the cost of running repair accounts for parts and labor as a dependent variable, and SMR as the independent variable. Cost of equipment overhauls were excluded in this research. All the needed information was obtained from FACT.

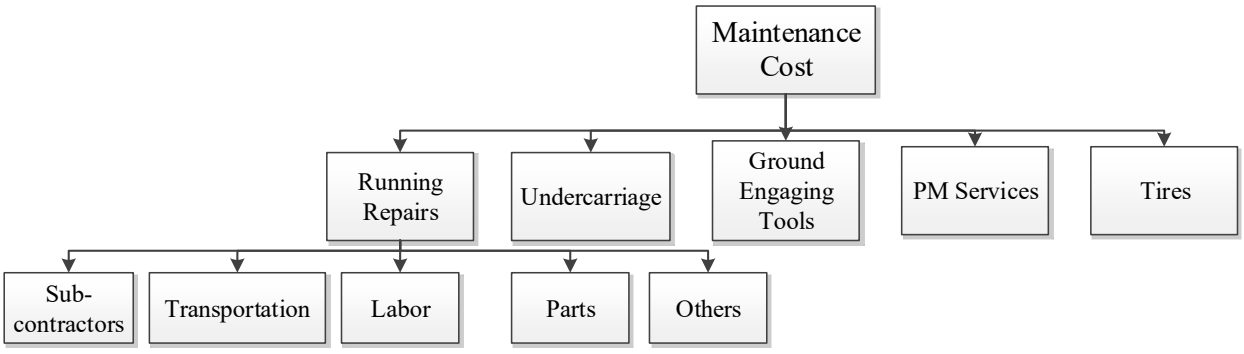


Figure 3: Breakdown of maintenance cost from company's database

Caterpillar 785 was selected as an example to demonstrate how the CCM for each type of heavy equipment is developed. This fleet contains 26 individual machines; monthly data is available from January 2003 to May 2015. All of the data are managed in several spreadsheets in Microsoft Office Excel 2010. However, before conducting any data analysis, adjustments to data points were required as detailed below.

3.3 Inflation rate adjustment

As the available maintenance cost records are from 2003 to 2015, inflation rates must be considered before using maintenance data. A computational form of the inflation equation in Equation 3-2 below was proposed by Jones (1982):

$$\boldsymbol{P(t_2) = P(t_1) \times \frac{I(t_2)}{I(t_1)}} \qquad \textbf{[3-2]}$$

Where I(t) represents an inflation index at time t. In the above equation, t_1 represents the time a transaction occurred, and t_2 represents the base time against which the transaction will be indexed. In this research, inflation rates were determined from company internal agreements and Statistics Canada. These include the labor rates from Overburden Agreements between the

company and Union of Operating Engineers and the *Equipment & Machine Price Index (construction area specified).* Labor rates can be used as a measurement of change in the wages earned by maintenance providers, while Equipment &Machine Price index (MEPI) provides estimates of price changes for machinery and equipment purchased by industry. In "Construction Equipment Policy", the author of this book recommends a composite index that contains combinations of indices for machinery price, petroleum, etc.(Douglas, 1975). A similar composite index is developed in this research. Labor rates will be applied to labor costs, while the MEPI will be applied, as the inflation rate, to the repair part. The sum of these two results is obtained on a monthly basis and is used in the CCM. The following equation provides a standard process of obtaining indexed points:

$$\boldsymbol{T(t_2) = P(t_2) \times \frac{MEPI(t_2)}{MEPI(t_1)} + L(t_1) \times \frac{Labor\ rates(t_2)}{Labor\ rates(t_1)}} \quad \textbf{[3-3]}$$

where T(t) represents the sum of the indexed cost at time t, P(t) stands for the repair parts cost at time t, L(t) stands for the repair labor costs at time t, and the base time t_1 is set as January 2003.

3.4 Data cleaning

As noted above, raw data needs to be cleaned and transformed prior to being used in the CCM. The following section describes three issues that were encountered regularly and how they were corrected.

3.4.1 Repeated points

This is a common issue for all heavy construction machinery in this study. When maintenance cost data and SMR records were summed up cumulatively, data was often repeated within certain ranges of time. There are two main reasons for this. First, it is possible that the machine was kept idle for a period of time, resulting in both maintenance cost and SMR remaining the same over different time periods. The second reason is that when a machine is at the shop for major repairs, the SMR remains constant while the cumulative cost increases. In both cases, the machine is counting more than one point for the same cumulative hours; therefore, these repeated points could influences the regression model with other machines. An example of repeated points is shown in Table 1 (Please note that all the data published in this paper have been processed and standardized for the confidentiality purposes).

Table 1: Repeated points example

Date	Machine #1		Machine #2		Machine #3	
	Cumulative Cost	SMR (1000 hrs)	Cumulative Cost	SMR (1000 hrs)	Cumulative Cost	SMR (1000 hrs)
29-Feb-08	813.98	0.53	630.76	0.27	0.00	2.03
31-Mar-08	991.39	0.73	2914.43	0.67	2384.72	2.47
30-Apr-08	11225.55	1.22	3433.50	1.24	7915.86	3.01
31-May-08	~~12769.79~~	~~1.22~~	~~6017.17~~	~~1.24~~	~~10283.55~~	~~3.01~~

The above data illustrate how field data appear in the company's accounting system. To address this issue, all but one of these repeated points are eliminated. In this study, the first set of points was kept as the cumulative cost for a machine which should include all maintenance cost up to the point when SMR were reached (Mitchell 1998). All costs incurred after are to be added in next interval.

3.4.2 Peak points

Another common issue in what is that the repair cost for each piece of machinery was occasionally recorded as negative. This could be due to the accounting method used at the company. Occasionally repair costs were overcharged earlier in the year, and the accounting department used negative points to balance the cost and keep it accurate. As the repair cost and SMR data were processed in a cumulative way, these negative points resulted in a few peaks and fluctuations when maintenance cost increased with the SMR records. Before developing any regression models, these points need to be eliminated (see Table 2).

Table 2: Peak points example

	Machine #1	
Date	Cumulative Cost	SMR (1000 hrs)
31-Aug-09	70899.48	15.17
30-Sep-09	~~80134.37~~	~~15.20~~
31-Oct-09	77796.36	15.70
30-Nov-09	80357.31	16.16

3.3.3 Intercepts

For a certain type of machines, not all started working at the same time. Theoretically, when the SMR is zero (i.e. the machine is brand new), the cumulative cost should be zero. However, within the company's system, there are a few scenarios where maintenance work has already been performed but the SMR remains near the zero level, or maintenance costs is zero while SMR increases. As shown in Equation 3-1, the starting point of the regression line begins at the origin (0, 0). All points that intercepts elsewhere on the axis have been eliminated.

3.5 Model generation

After the data has been cleaned, the regression generation model is developed based on Mitchell Jr (1998). In this study, a few data sets were developed based on the original monthly data. Raw data were cumulatively calculated by season, semi-annually, annually, and at 500, 1000, 1500, 2000, 2500, 5000, 7500, and 10,000 hour intervals. Using the linear interpolation method, data points in different SMR intervals are calculated based on the pairs of nearest points. Using Equation 3-1, this study developed several equations based on different data sets. Figure 4 plots a CCM model based on the original data set. This process (CCM Model) was repeated for each of the 18 machine types.

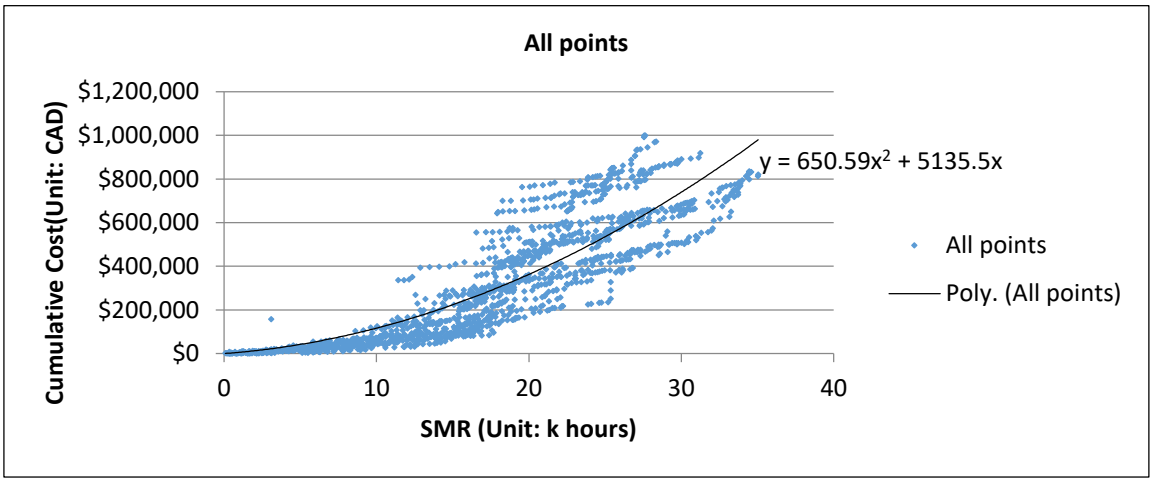


Figure 4: All points data set (Caterpillar 785)

3.6 Analysis of results

Regression analysis was conducted for each type of machine using the Microsoft Excel add-on "Analysis ToolPak". Taking a fleet of Caterpillar 785 trucks as an example, Table 3 summarizes

the regression analysis results. Results of other categories of machines can be found in Appendix A.

Table 3: Regression Analysis Summary for CAT 785

CAT 785 – 26 machines		Coefficient	P-Value	R Square	Adjusted R Square	F	MSE(E+10)	Number of points
All Points	SMR	5135.54	1.7232E-13	0.9065	0.9057	7287.41	1.64	1506
	SMR^2	650.59	4.296E-100					
Seasonal	SMR	5802.63	1.3576E-06	0.9059	0.9037	2421.17	1.60	505
	SMR^2	627.33	7.0474E-33					
Semi-annual	SMR	5688.56	0.00113075	0.9058	0.9013	1168.44	1.67	245
	SMR^2	624.05	2.592E-16					
Annual	SMR	2414.17	0.26858736	0.9054	0.8980	708.38	1.56	150
	SMR^2	775.08	2.6889E-14					
500 SMR Interval	SMR	3079.45	2.8621E-06	0.8947	0.8939	5432.17	1.26	1281
	SMR^2	728.93	1.016E-116					
1000 SMR Interval	SMR	3247.32	0.00049171	0.8941	0.8923	2675.15	1.27	636
	SMR^2	718.82	5.1487E-58					
1500 SMR Interval	SMR	3297.59	0.00420612	0.8931	0.8905	1754.85	1.29	422
	SMR^2	719.10	2.0212E-38					
2000 SMR Interval	SMR	3100.98	0.01924877	0.8945	0.8910	1322.81	1.28	314
	SMR^2	726.62	7.2682E-30					
2500 SMR Interval	SMR	2749.38	0.0646914	0.8936	0.8892	1057.95	1.29	254
	SMR^2	741.19	7.9035E-25					
5000 SMR Interval	SMR	2495.95	0.26368117	0.8849	0.8753	445.91	1.42	118
	SMR^2	749.66	6.2361E-12					
7500 SMR Interval	SMR	447.27	0.86798787	0.8924	0.8774	306.86	1.31	76
	SMR^2	822.48	1.6846E-09					

These data sets can be divided into two categories. The first, including seasonal points (one data pair every three months), semi-annual points (one data pair every six months), annual points (one data pair every 12 months) is based on calendar age. It is noted that there are no significant improvements between these three data sets and the original monthly data set. In fact, because each individual machine has its own schedule, some machines could have been idle for months,

and some maintenance information could have been lost if data were processed this way. The second category is based on machine age (i.e. SMR intervals). In practice, equipment managers usually make maintenance decisions based on fixed SMR intervals. This category provides a reference to help the company to design maintenance policies.

Generally, the R square value provides an indication of the regression model's performance. The measured R^2 is interpreted as how closely the data fit the regression line, falling between zero and one. If the absolute value of R^2 is larger than 0.75, it indicates a good fit of data (Peck & Devore, 2011). Instead of using the R^2 value, in multiple regression models the adjusted R^2 value is often used to avoid gaining a goodness-of fit statistic by adding more variables. Statistically, the R square value is an important measure, but not the only measure of how closely the data fit the regression line. As shown in Table 3, however, there is little difference between different data sets when focusing on the R square or adjusted R square value. In this study, regardless of how data are organized, good fitness is obtained when using the second-order polynomial equations suggested by Mitchell Jr (1998).

P-value is used to test statistical hypothesis. Assuming the null hypothesis is true, p-value is the probability of achieving a result that is "more extreme" than what was actually observed. In this study, the null hypothesis refers to a position where there is no relationship between SMR and the cumulative maintenance cost of machines. The threshold value, denoted as α, is set to be 1%. Significance of the model coefficients are ascertained using the following criteria:

P - value $\leqslant$ 0.01—acceptable; P- value > 0.01—unacceptable

This test is performed on all coefficients of models. For CAT 785, since the p value for SMR coefficient needs to be under 0.01, data sets with SMR intervals exceeding 2000 hours are eliminated.

Residual mean square error, often abbreviated as MSE, is the ratio of residual sum of squares to its degrees of freedom value, which is roughly a mean of the squared errors in using the regression trend line to predict explainable variable y. In practice, a lower MSE value indicates higher accuracy of regression models. For the CAT 785 rigid frame truck in this model, data sets with 500 and 1000 SMR intervals had a better performance than the other data sets. The "F" value, which is the mean square for regression divided by MSE, is also chosen to identify the model that best fits the population. Other than original data sets, 500 and 1000 SMR intervals have a higher F value indicating a better performance than others.

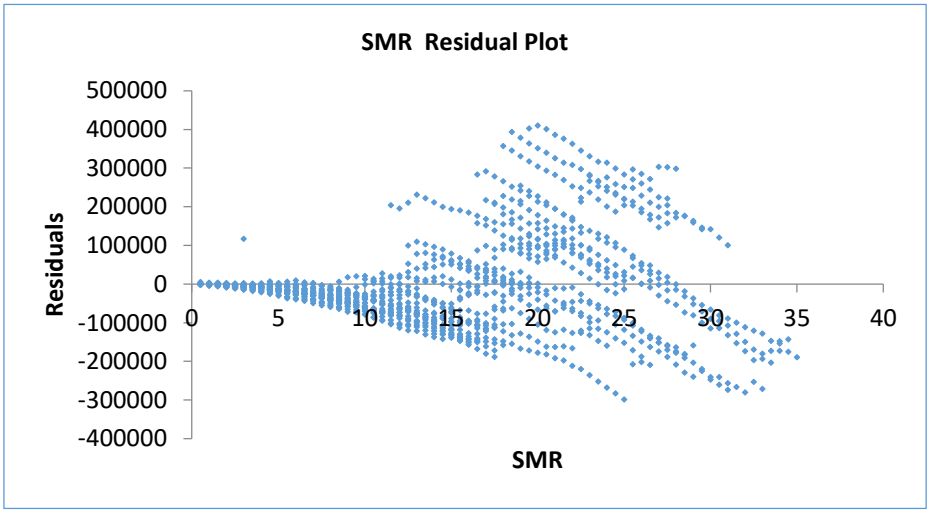


Figure 5: Residual plot for CAT 785 (500 SMR Interval)

Additionally, the residual plots are also analyzed. Figure 5 shows residual plots of CAT 785 at 500 SMR intervals. Ideally, the points should fluctuate randomly around 0, and there should be

no obvious regular changes in trends as SMR increases. However, as seen in Figure 5, as SMR increases from 0 to 20,000 hours, variation of residuals increases. The increasing spread from left to right in Figure 5 shows a heteroscedasticity problem. A pattern like this indicates that the residual standard deviation of y is not constant – the data are more spread out when SMR is around 20k area, suggesting that the variance of y is not the same at each x value but rather increases with x. According to Peck and Devore (2011), when the data points are of varying quality, it is better to select the best-fit line by using weighted least squares (WLS) than ordinary least squares (OLS). Besides, this divergence trend of residual is not uncommon for all 19 different types of machines. It often indicates that the model can be improved, or that a few variables might be missing. In fact, even for the fleet with same type of machines, each individual has its own schedule and work conditions, and operator skills could also make a difference on maintenance costs.

3.7 Conclusion

This study contains 19 different types of heavy machines, including rigid frame trucks, excavators and shovels. By using CCM, different data sets were created and analyzed to select the optimum data set to represent each type of machine. The main criteria for choosing optimum data sets are R^2 value, F-value, P-value, and MSE. Generally, 500 SMR intervals and 1000 SMR intervals are selected as the optimum data sets for most heavy machines. In most cases, construction companies already have established maintenance policies, which are usually based

on fixed SMR intervals. Equations and statistical analysis from this study will provide a reference for equipment managers to use when creating maintenance cost model and making economic decisions. The residual plots for each regression model indicate that factors apart from SMR records, such as working conditions, operator skills and other related attributes could have a potential influence on maintenance cost considerations. As this type of data are unavailable in this research, further information is needed to quantify such influences. Methods of quantifying such mentioned attributes and evaluating the method of WLS are recommended to be discussed and investigated in future research.

Chapter 4: Residual value prediction models

The owning and operating costs of heavy construction equipment constitute a significant expenditure for companies, especially those engaging in earth moving and industrial installation projects. Construction equipment management involves managing equipment resources not only to satisfy project requirements but to achieve maximum return on disposed assets. Therefore, it is useful to develop a reliable model to help decision-makers estimate the residual value of a heavy machine before purchasing or disposing of it.

4.1 Methodologies

Predicting a numeric value is a common task for data mining, based on the training dataset with predictor variables. Fan et al. (2008) indicated that although different data mining algorithms used unique methods for model inference, they had a few common features such as intuitive visualization, making them superior to traditional regression models. With progress in computer hardware, algorithms can be run very fast on personal computers. Data mining software and language packages such as Weka (Hall et al., 2009) are widely used and becoming increasingly reliable. This section introduces a basic instance-based learning (IBL) method and an ensemble-based learning method to predict equipment residual value.

4.1.1 Instance-based learning

IBL is a family of learning algorithms that compares target instances with instances seen in

training datasets. Derived from the nearest neighbor pattern classifier (Aha, Kibler, & Albert, 1991), training instances are searched for an instance that is most closely related to a new target instance. Since the stored training instances themselves represent the main knowledge, no model is learned.

K nearest neighbor (KNN) is a member of IBL family and is widely used throughout both academia and industry (Chen, 2008; Dang, Zhang, Zhang, & Zhao, 2005; B.-H. Lee & Scholz, 2006; Rosa, Ebecken, & Costa, 2003). KNN is a non-parametric method that can be used for both classification and regression. While the output for classification is a class membership voted by a majority vote of its k (k is a positive integer) nearest neighbors, the output for regression is a numeric value averaged, or weighted averaged by its k nearest neighbors. For simplicity, consider a regression problem with response variable Y and two independent variables X_1 and X_2. The numerical value can be predicted using the following generalized form: $y = f(x_1, x_2)$. The expression of "nearest neighbor" indicates that a distance should be inspected from a geometric perspective. Euclidean distances are calculated from the query example to training examples as:

$$\boldsymbol{d(t,p) = \sqrt{(t_{x_1} - p_{x_1})^2 + (t_{x_2} - p_{x_2})^2}} \qquad \textbf{[4-1]}$$

where d(t, p) is the distance between the target instance (t) and training points (p); t_{x1} and p_{x1} represent the value of the X_1 attribute of t and p; and t_{x2} and p_{x2} represent the value of the X_1 attribute of t and p. Next, training points are ordered by distance in an ascendant way. K points will be selected from the top of the distance list as the nearest training points for the target. Finally, the weighted average Y values of these selected points will be assigned as the predicted

value for the target instance. Figure 6 shows a diagram illustrating a dataset of 20 training instances (stars in the diagram) and a target instance falling onto the diagram (represented by a red triangle). If k is set to be 3, for instance, three nearest "neighbors" (i.e. p_1, p_2, and p_3) are found and the weighted average value of y_1, y_2, and y_3 is calculated.

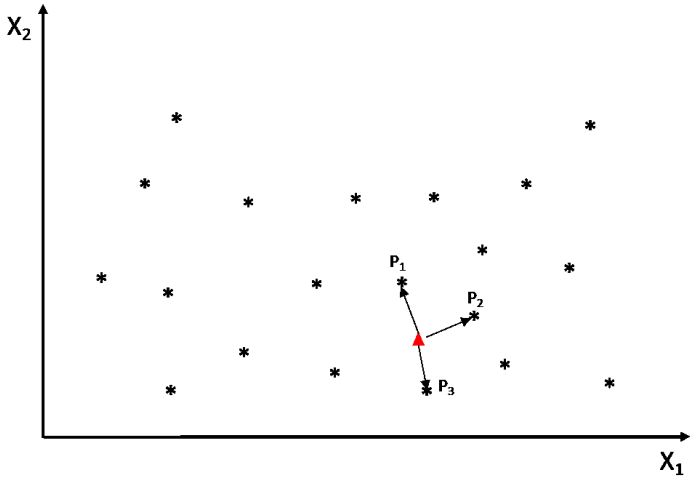


Figure 6: Examples of KNN

Now we consider a generalized version of the above example: a regression problem with n features and one response variable. Provided m observations, each observation consists of (x_{i1}, x_{i2}, … , x_{in}, y_i) for i = 1, 2, …, m. Distances are calculated as

$$\boldsymbol{d_i(t,p)} = \sqrt{\sum_{j=1}^{n}(t_{x_{ij}} - p_{x_{ij}})^2} \qquad \textbf{[4-2]}$$

Finally, a weighted average Y value of these selected k "nearest neighbors" is allocated to the target instance. A 1/distance is used as the weight for calculating average values of chosen points. The reason for using weighted average value instead of average value of selected response variables is based on the assumption that a "similar" instance will have a similar value. In other words, the closer the neighbor is to the target instance, the more biased predicted value of target

will be.

Problems related to features property need to be considered when calculating distances. For numeric features, such as the machine age and operating hours of a machine, data should be scaled or normalized before processing. Normalization adjusts independent variables measured on different scales, avoiding biased influences on distance value. Meanwhile, for categorical or nominal variables such as the machine's manufacture, the value of $(t_{x_{i1}} - p_{x_{i1}})$ is set to be 0 if the value of the training point is the same as that of the target point while 1 if they are different (suppose x_{i1} is a categorical variable of the i th observation).

4.1.2 Ensemble method

As an ensemble-based algorithm, a random forest (RF) is operated by constructing a mass of single decision trees in training, combined through majority voting (classification) or averaging (regression) the individual trees. Ho (1995) first created an algorithm of "random decision forests" introducing random features selection. In 2001, Breiman (2001) developed the "random forest" algorithm using a combination of the random features selection methods (Amit & Geman, 1997; Ho, 1995, 1998) and Breiman's "bagging" (also known as bootstrap) idea (Breiman, 1996). Detailed instructions of bagging and random features selection are given in the following paragraphs.

A random forest is an ensemble of decision trees. If no modification is made to the trees, each tree will be exactly the same and no improvement will be shown in the results. In order to make our trees effective, a variation has to be introduced into each individual decision tree model.

Each tree will be constructed slightly differently, therefore will make different predictions. The word “random” in “random forest” refers to this kind of variation. As mentioned earlier, two main methods are used to create variation – bagging and random feature selection.

In a random forest, each individual tree is trained by the whole dataset. Instead, a random sample of the data, or a “bag”, is trained by sampling with replacement from the original data. A recent development in ensemble techniques has been the widespread use of bootstrap (or bagging) to generate a diverse data subset for training base models (Zhang & Haghani, 2015). When sampling with replacements, after a row from the original data is selected, that row will be restored. In other words, some observations can be picked up again in other samples while other observations might be “left out” of the sample. A few rows from the original data may appear in the “bag” multiple times. For example, given a training dataset with a total of m records, n new training sets are created by sampling with a replacement from the original dataset. Each new set has a sample size of s. The n base trees are trained using the generated n training set and combined as the output. Figure 7 provides a pseudo-code for random forest algorithm.

The hypothesis behind the random forest is the property of instability. The more “diverse” each tree is to construct a forest, the stronger the combined prediction will be. Otherwise, if each base tree of the forest is similar in how it make predictions, the boost from the ensemble will be negligible. In our former example, each bagged tree is grown on data samples randomly drawn with replacements from the original dataset. If s is a relatively large number, the learned trees are usually similar to each other. An averaging (regression problem) or majority voting

(classification) of these base trees does not improve prediction accuracy.

The random forest is distinguished from other bagged regression trees in that instead of using all of the features for each individual tree, it only allows a random subset of features at each splitting node of the tree. We then compute the information gain (Kent, 1983) for each feature in the random sample and pick the one with the highest value to split on. We repeat the same process to select the optimal split for a node, but only evaluate a constrained set of features that are selected randomly. This introduces variation into the trees and makes for stronger ensembles.

```
# Given a data set with a total number of m observations with F input variables (i.e. features)
# N bagging samples are randomly drawn with replacement, with a sample size of s
# Now we are going to construct N individual trees based on bagging samples
For n = 1 to N, do:
    # Start with the root node
    Grow a tree through following loop repeatedly, do until:
        Randomly select f variables as a subset of F (i.e. f < F);
        Calculate the information gain for each features in f;
        Find the best feature (with maximum information gain) in f as a split node;
        Split the node into two daughter nodes;
    End;
      Output an individual base tree Tn;
End;
# A random forest is constructed containing N individual trees.
Output for random forest (regression) is: 1/N ∑_{n=1}^{N} T_n
```

Figure 7: Pseudo-code of random forest algorithm

One of the major advantages of random forests over a single tree is they "overfit" less. Although each individual tree in a random forest varies, the average of each tree's prediction is less sensitive to the input data than a single tree is. Generally, a random forest is a powerful algorithm for dealing with large datasets, but it also takes a longer creation time and is harder to interpret. A

detailed discussions about this will be provided in the “Discussion and limitations” section.

4.2 Data description and pre-processing

The residual value prediction of heavy construction equipment is based on available information about individual machines and the economic environment. The primary data source for this research is from a construction company doing business in North America, whose fleet management department gathered resale information for different categories of heavy machines. As pointed out by Fan et al. (2008), although it is possible to build a single model for all categories of heavy machines, the model of scale would be of poor quality and difficult to interpret. In this research, the modeling process is exemplified by selecting the category of articulated trucks for model generation and comparison.

Auctions of articulated trucks throughout North America (including Canada, the United States and Mexico) from 2011-2015 are studied. Economic indicators (i.e., Real National GDP and GDP Growth) are obtained from Statistics Canada and the U.S. Bureau of Economic Analysis.

Original auction datasets zipped all the information together into four columns, “Description Data”, “Location Data”, “SMR Data”, and “Price Data”. SMR is the abbreviation for “service meter hour” which is an expression of how many hours a vehicle really served excluding the idle time. As the SMR is available in the dataset of this research, this study differs from previous research as it expresses “the usage of equipment” by using only the machines’ calendar age. In addition to addressing “how old the machine is”, SMR answers the question of “how many hours

has the machine served". Another concern with the use of the calendar age (i.e., age in years) is that as a machine nears the end of its useful life, it might be used less and less (Terborgh, 1949). Table 4 gives examples of original auction records.

Table 4: Examples of original auction records

Number	Description Data	Location Data	SMR Data	Price Data
1	2007 CATERPILLAR 740 6x6 Articulated Dump Truck	Orlando, FL, USA Thursday Oct 23, 2014	Meter Reads :10551 Hr	Sold for: 140000 USD
2	2004 JOHN DEERE 400D 6x6 Articulated Dump Truck	Chehalis, WA, USA Wednesday Oct 22, 2014	Meter Reads :10522 Hr	Sold for: 80000 USD
3	2007 CATERPILLAR 740 6x6 Articulated Dump Truck	Verona, KY, USA Thursday Oct 16, 2014	Meter Reads :7881 Hr	Sold for: 165000 USD
4	2007 JOHN DEERE 250D 6x6 Articulated Dump Truck	Chilliwack, BC, CAN Wednesday Oct 15, 2014	Meter Reads :6570 Hr	Sold for: 105000 CAD
…	…	…	…	…

To effectively collect information from original auction records, useful information such as the manufacturer and auction location are "sliced" out of "Description Data" to act as separate predictor features. Specifications such as rated payload (t), body capacity (cy), and horse power (HP) are also collected. For a particular model of articulated trucks, machines can vary based on different payload or capacity. To identify an individual machine more accurately, information about a few representative specifications is collected from manufacturers' manuals.

As a significant macroeconomic indicator of a country, gross domestic product (GDP) is used in various models as an independent variable predicting residual values of heavy construction machines (Fan et al., 2007; Lucko, 2003, 2011; Lucko et al., 2006; Lucko & Mitchell Jr, 2010). Instead of using nominal GDP (also known simply as "GDP"), real GDP is used in this research. A major difference between these two macroeconomic indicators is that real GDP is adjusted as

per changes in the general price level while nominal GDP is not. In other words, real GDP is adjusted for inflation, which means that it shows the actual picture of a country's economic environment. Another issue that should be noted is that either nominal GDP or real GDP is represented by an absolute number, usually in US$ billion. As this dataset includes auction records through the US, Canada, and Mexico, numbers are much different from each other. For example, the real GDP for Mexico in 2014 was around US$1,056 billion while for the US, it was US$15,962 billion, indicating a 15 times larger of economy than Mexico. The author of this thesis believes that the growth of GDP (in percentage) is a better representation of the economic environment, which might have more impact on the heavy construction equipment market.

Table 5 shows the selected features for predicting the residual value of heavy machines.

Table 5: Features of datasets

Machine age	Calculated based on the built year and the year of auction
Brand (Make)	Manufacturer of articulated trucks
Model	Model of articulated trucks
Rated payload(t)	Weight that the articulated truck can carry
Body capacity(cy)	Struck-heaped volume of the truck body
Horse power(HP)	Engine horsepower
Location	States where the auction occurs
Auction year	The year when auction occurred
Real GDP (US$ Billion)	Refers to the measure of GDP adjusted according to the general price level
Real GDP growth	Increment of real GDP divided by real GDP of previous year
SMRs	Service meter hours, refers to the usage of equipment

The author of this thesis interviewed a few professionals from the construction company and noted that the residual percentage was often used instead of the equipment resale price in constant dollars when the company made decisions about acquisitions or disposal. This is verified in research by Lucko (2003) and Lucko et al. (2007). Inflation rates (based on consumer

price index), and exchange rates between Canadian Dollars and US Dollars are applied to adjust the auction price. Residual value percentage can be obtained by dividing the auction price by the initial value of the machine.

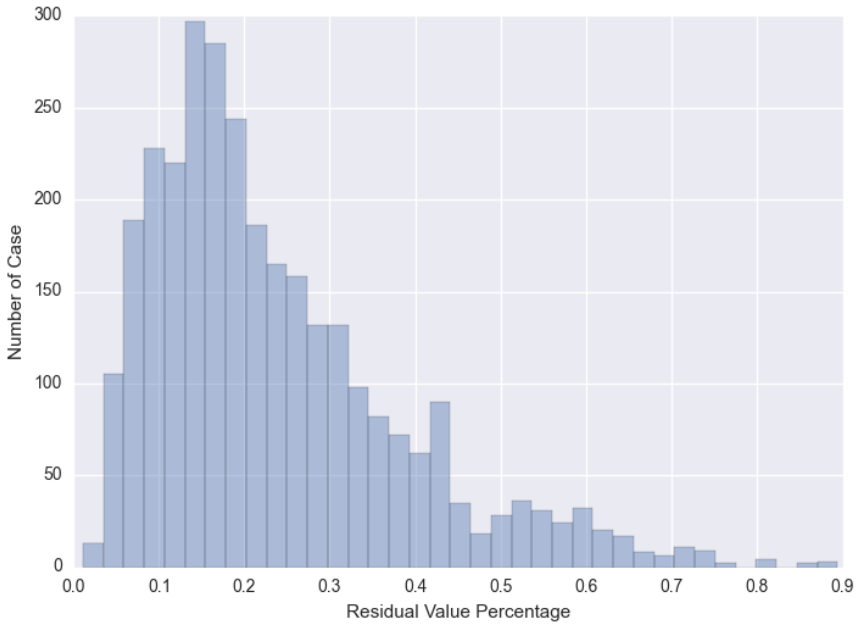


Figure 8: Residual value percentage of articulated trucks

Figure 8 illustrates a general view of the dataset. A total number of 3,044 cases of articulated trucks is obtained from data pre-process. Most of machines are auctioned within 0-50% residual value percentage. It appears that the residual value of equipment drops quickly in the early years, but these declines flatten out after a few years' usage. Acknowledging this trend, most cases in our research fall into the interval where machines are regularly used and sold. In the next sections, predictive models are built to simulate the trend of heavy equipment residual value.

4.3 Model generation and optimization

Before a predictive model is built, an attribute selection algorithm called ReliefF (Kononenko, 1994; Robnik-Šikonja & Kononenko, 1997) is run upon the dataset to identify which feature has a dominant impact on equipment residual value. By exploiting local information given by different contexts, this algorithm could offer a unified observation on estimating attribute quality in regression problem. In addition, a tenfold cross-validation method is utilized to rank features. The results are shown in descending order: machine age, model, real GDP growth, SMRs, brand, rated payload (t), body capacity (cy), horsepower (HP), real GDP (US$ Billion), and location. Based on the author's interview with professionals from a construction company, it is found that the SMR is always listed as the top place to predict residual value in current practice. However, this research provides a different view of the feature selection that machine age might have much more power of impacting on residual value, or at least becomes as a dominant factor especially during a certain period of usage. Figures 9 and 10 provide comparative scatterplots of machine age vs. residual value percentage and SMRs vs. residual value percentage. It could be seen that both machine age and SMRs have an obvious negative relationship with residual value percentage. Another interesting fact is that the real GDP growth ranks much higher than the absolute value of GDP, validating our conjecture that the auction market for articulated trucks is more sensitive to the growth of real GDP instead of real GDP itself. After features ranking, data mining methods are generated and validated after a tenfold cross-validation method.

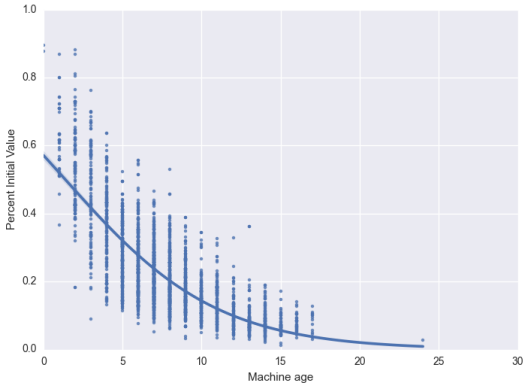


Figure 9: Machine age vs. residual value percentage

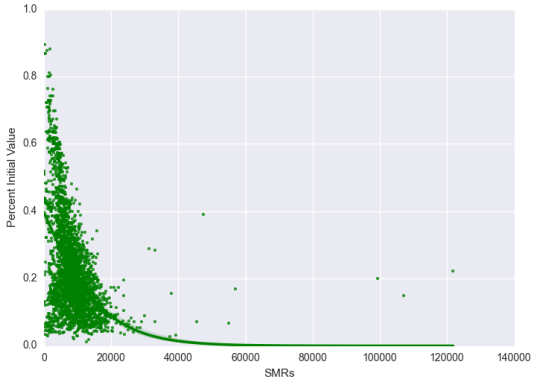


Figure 10: SMRs vs. residual value percentage

To validate the prediction of different data mining models, a few measures are selected as criteria in this research:

- Correlation coefficient (CC): this is a value that evaluate how much y (real value) and ŷ (predicted value) correlated with each other, falling between 0 and 1. The higher the value is, the stronger correlation indicated
- Mean absolute error (MAE): this is the average distance between the predicted value and actual value

$$\boldsymbol{MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|} \qquad \textbf{[4-3]}$$

where N is the total number of cases, $\hat{y}_i$ is the predicted value of i th case, and y_i is the actual value of i th case.

- Root mean absolute error (RMSE): RMSE provides another way to estimate distances between y and ŷ

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}})^{\mathbf{2}}} \quad \textbf{[4-4]}$$

- Relative absolute error (RAE) : RAE scales error to the mean, which estimates how much y differs from its average value

$$\boldsymbol{RAE} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}|\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}}|}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}|\bar{\boldsymbol{y}} - \boldsymbol{y}_{\boldsymbol{i}}|} \quad \textbf{[4-5]}$$

  where $\bar{\boldsymbol{y}}$ is the mean actual value.

- Root relative squared error (RRSE): RRSE is similar to RAE, scaling error to the mean

$$\mathbf{RRSE} = \sqrt{\frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}})^{\mathbf{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\bar{\boldsymbol{y}} - \boldsymbol{y}_{\boldsymbol{i}})^{\mathbf{2}}}} \quad \textbf{[4-6]}$$

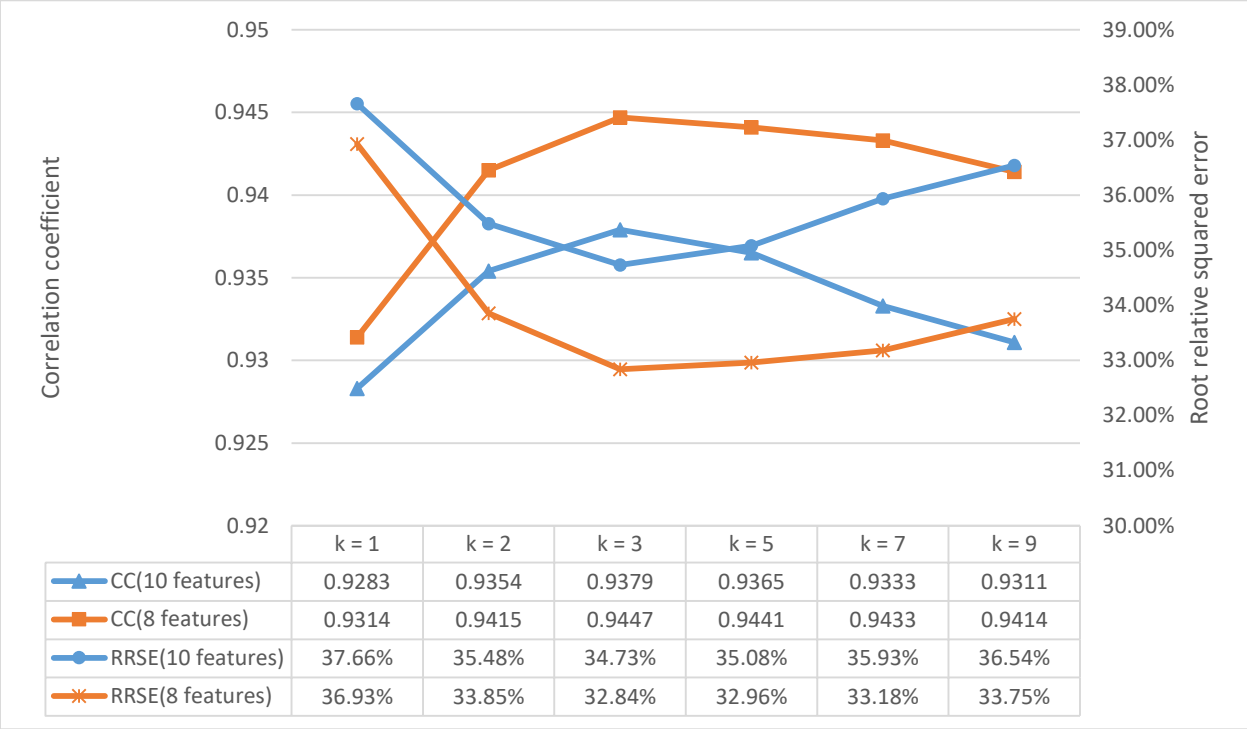
4.3.1 KNN Model

A k nearest neighbor (KNN) algorithm is implemented into the articulated trucks dataset which contains 3,044 auction cases. KNN includes a main parameter to control model structure: the number of k, which represents the number of instances taken into account to determine the target subset. As there is no strict method to calculate the optimum value of k, different numbers of k are tested in this research. Table 6 shows statistical results between different selected k values. When k equals 3, the model performs best, with the highest correlation coefficient (CC) and lowest error.

Table 6: Results of KNN models

Dataset with 10 features	K = 1	K = 2	K = 3	K = 5	K = 7	K = 9
Correlation coefficient	0.9283	0.9354	0.9379	0.9365	0.9333	0.9311
Mean absolute error	0.0319	0.0317	0.0319	0.033	0.0341	0.0349
Root mean squared error	0.0556	0.0523	0.0512	0.0517	0.053	0.0539
Relative absolute error	37.79%	27.62%	27.78%	28.76%	29.68%	30.35%
Root relative squared error	37.66%	35.48%	34.73%	35.08%	35.93%	36.54%
Total number of instances	3044	3044	3044	3044	3044	3044

Another finding of interest is a demonstration of features ranking obtained during data pre-process procedure. The features selection procedure showed that the auction location and real GDP (US$ Billion) are the two attributes that contribute least to residual value prediction. Next, a new dataset excluding location and real GDP information is created to test, and it turns out that a CC increases to 0.9447 while the root relative squared error decreases to 32.84%. In this case, given k was equal to 3. Details are shown in Figure 11. With a “shrunk” dataset, the model with the best performance was also found when k equaled 3. In this paper, k is set to be 3 and a weighted average function is used to obtain the final output



	k = 1	k = 2	k = 3	k = 5	k = 7	k = 9
CC(10 features)	0.9283	0.9354	0.9379	0.9365	0.9333	0.9311
CC(8 features)	0.9314	0.9415	0.9447	0.9441	0.9433	0.9414
RRSE(10 features)	37.66%	35.48%	34.73%	35.08%	35.93%	36.54%
RRSE(8 features)	36.93%	33.85%	32.84%	32.96%	33.18%	33.75%

Figure 11: Statistical results for different number of feature

4.3.2 Random forest model

The RF algorithm includes a few parameters for model structure control. The number of iterations indicates how many trees a forest has. Generally, increasing the number of iterations is the easiest tweak. The accuracy increase function is logarithmic-like, which means that the number going from 10 trees to 100 trees will make a bigger difference than that going from 100 to 200 trees. Although there is never any harm in constructing more trees, and an appropriate increase in trees could provide a better prediction because the model can obtain sufficient information from various individuals, a trade-off between tree building and time efficiency exists. That is, the more trees there are to be built, the more time will be spent on model construction. Usually, the number of iterations will not exceed 200.

Another main parameter is the number of tried attributes. This indicates the total number of features selected in an individual run. Usually, a square root of the number of features is assigned. For example, if the training dataset has 100 predictive features, 10 ($\sqrt{100}$) attributes are randomly selected at each split node of individual trees. Another typical calculation for the number of attributes is using$(\log_2 N + 1)$.

In this research, the square root of predictive attributes is assigned to the number of tried attributes and a random forest with 200 individual trees is built. A ten-fold cross validation is implemented to train the model. Table 7 shows a summary of the RF result. Model comparisons are discussed in the next section

Table 7: Result of RF Model

Correlation coefficient	0.9602
Mean absolute error	0.0289
Root mean squared error	0.0417
Relative absolute error	25.1809%
Root relative squared error	28.282%
Total number of instance	3044

4.4 Model comparison

A benchmark algorithm in this research is M5P, a single regression decision tree algorithm that combines a traditional decision tree with the linear regression functions at each node. See Quinlan (1992) and Wang and Witten (1996) for more details.

To avoid random uncertainty of the algorithms' performance, besides implementing a ten-fold cross validation, each algorithm (KNN, RF, and M5P) is repeated 10 times with different seeds.

That means, 100 calls of each classifier with training data and test again test data to get a better statistical perspective. The original dataset with 10 features (i.e., machine age, model, real GDP growth, SMRs, brand, rated payload (t), body capacity (cy), horsepower (HP), real GDP (US$ Billion), and location) of 3,044 cases of articulated trucks' auction records is used to make a comparison. Table 8 shows a general comparison of different criteria. Results are calculated as the mean value of overall 10 runs. The table shows that both the KNN and RF models perform better than the M5P, which is built on a single regression tree algorithm. The higher predictive error of the single regression tree algorithm can be explained by the single regression tree's high sensitivity to small perturbations in data. In other words, a small amount of the changes could result in a totally different tree. This feature becomes obvious when a value with a few attributes value "out-of-sample." A single decision tree like M5P might provide a "bad guess," while another ensemble tree algorithm like RF or an instance-based algorithm like KNN can weaken or even negate the predictive error to a certain extent. Discussions about the three algorithms are provided in the next section.

Table 8: Average results for three different models

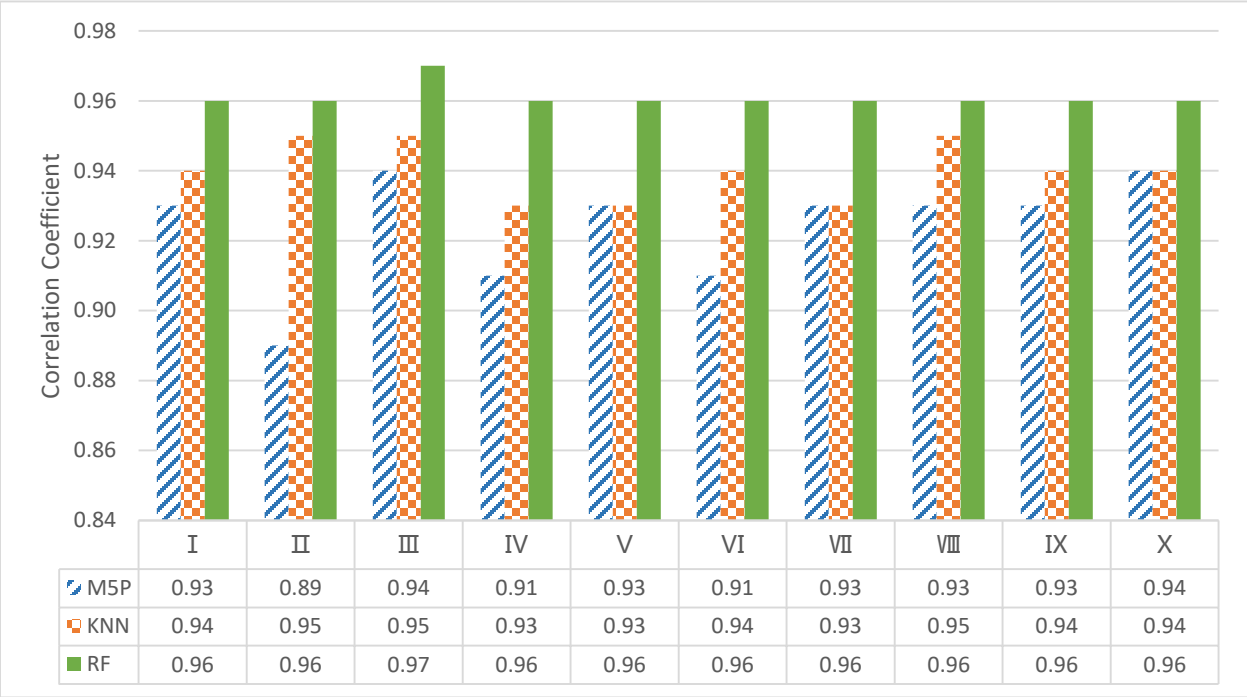
Model	CC	MAE	RMSE	RAE (%)	RRSE (%)
M5P	0.9240	0.0365	0.0552	31.87	37.56
KNN	0.9400	0.0313	0.0498	27.29	33.92
RF	0.9610	0.0286	0.0411	24.96	27.93

Figure 12 illustrates the result of CC for each run. While all three models have a good performance with a CC value above 0.9, RF stands out with an average CC value of 0.96. Table 9 shows the standard deviation of CC for each run of each model, which indicates that both KNN

and RF are much more stable and reliable; unlike a single regression tree algorithm, KNN and RF both demonstrate low standard deviations for each run. An illustrative comparison of RRSE between different models can be found in figure 13. In terms of RRSE, RF models have a relative low value around 28% while M5P models have much higher values than the other two. Other criteria including MAE, RMSE, and RAE) also indicate that RF models perform best (See Appendix B)

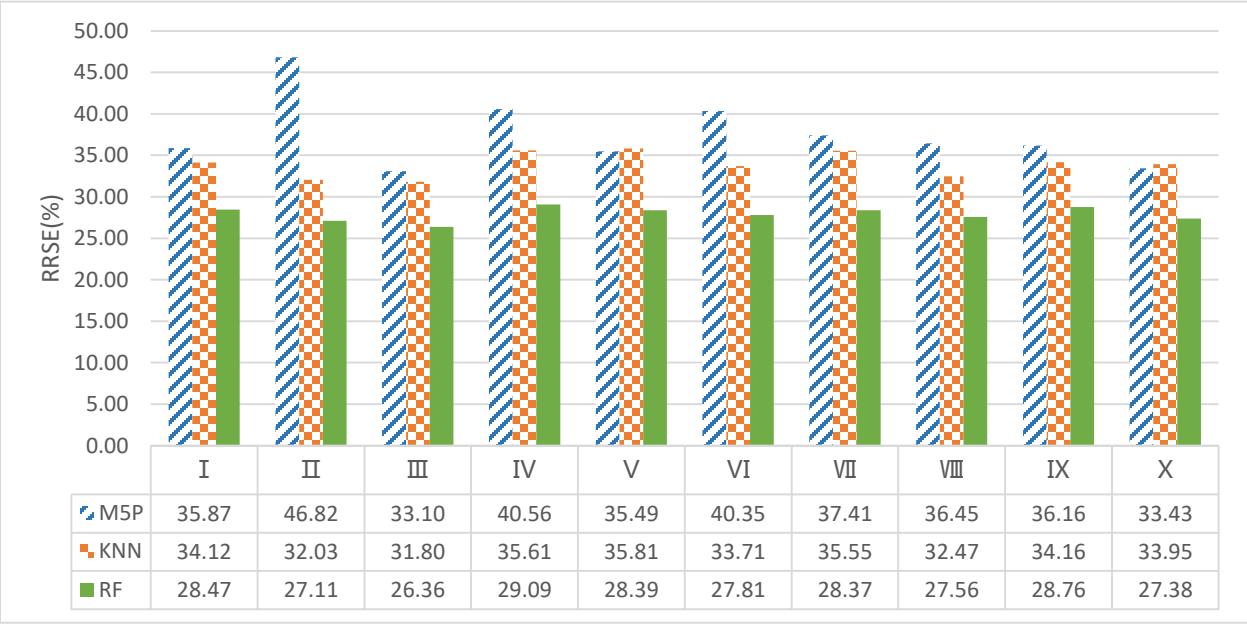
Table 9: Standard deviation of CC for each run of three models

Run	M5P	KNN	RF
I	0.04	0.01	0.01
II	0.12	0.01	0.00
III	0.02	0.01	0.01
IV	0.08	0.02	0.01
V	0.03	0.01	0.00
VI	0.04	0.01	0.00
VII	0.04	0.01	0.01
VIII	0.02	0.01	0.01
IX	0.04	0.01	0.00
X	0.01	0.01	0.01



	I	II	III	IV	V	VI	VII	VIII	IX	X
M5P	0.93	0.89	0.94	0.91	0.93	0.91	0.93	0.93	0.93	0.94
KNN	0.94	0.95	0.95	0.93	0.93	0.94	0.93	0.95	0.94	0.94
RF	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.96	0.96

Figure 12: Comparisons of CC for each run for M5P, KNN, and RF



	I	II	III	IV	V	VI	VII	VIII	IX	X
M5P	35.87	46.82	33.10	40.56	35.49	40.35	37.41	36.45	36.16	33.43
KNN	34.12	32.03	31.80	35.61	35.81	33.71	35.55	32.47	34.16	33.95
RF	28.47	27.11	26.36	29.09	28.39	27.81	28.37	27.56	28.76	27.38

Figure 13: Comparisons of RRSE for each run for M5P, KNN, and RF

In addition, training time is also considered a comparative indicator of these three models. Speed differences are not that noticeable in model generation because the dataset being studied is relatively small. However, as long as there is a live connection with auction companies or any

related database to obtain up-to-date information, time efficiency is critical, especially when dealing with large datasets. According to the original dataset, the KNN model performs much faster than the other two (Figure 14). This is because the instance-based learner method stores training instances to represent the main knowledge, so no model is learned. Therefore, KNN is effective when the training set is large. Another finding of interest is that although it takes more training time for M5P model than RF model, test time is reverse. M5P, like other single regression tree algorithms, will split attribute values at each node until a minimum number of instances is achieved (in our case, the number is set by default which is 4). This takes much time if the dataset has many attributes and values ranges in a large interval. Once a tree is built, however, new tested instances will follow the trunks of the tree and take little time to leaf, as is shown in our experiment (Figure 14). As an ensemble tree algorithm, RF is a combination of many single trees. It thus seems plausible to expect that building hundreds of trees should take longer time than building only one tree. However, as RF randomly chooses a subset of features at each split node and only a bagging of data is trained, it depends. Figure 14 shows that regarding model training time, RF runs faster than M5P in our case. In terms of tested instance, it will take more time for RF model due to its much more complexity than a single tree. See Figure 14 for more details.

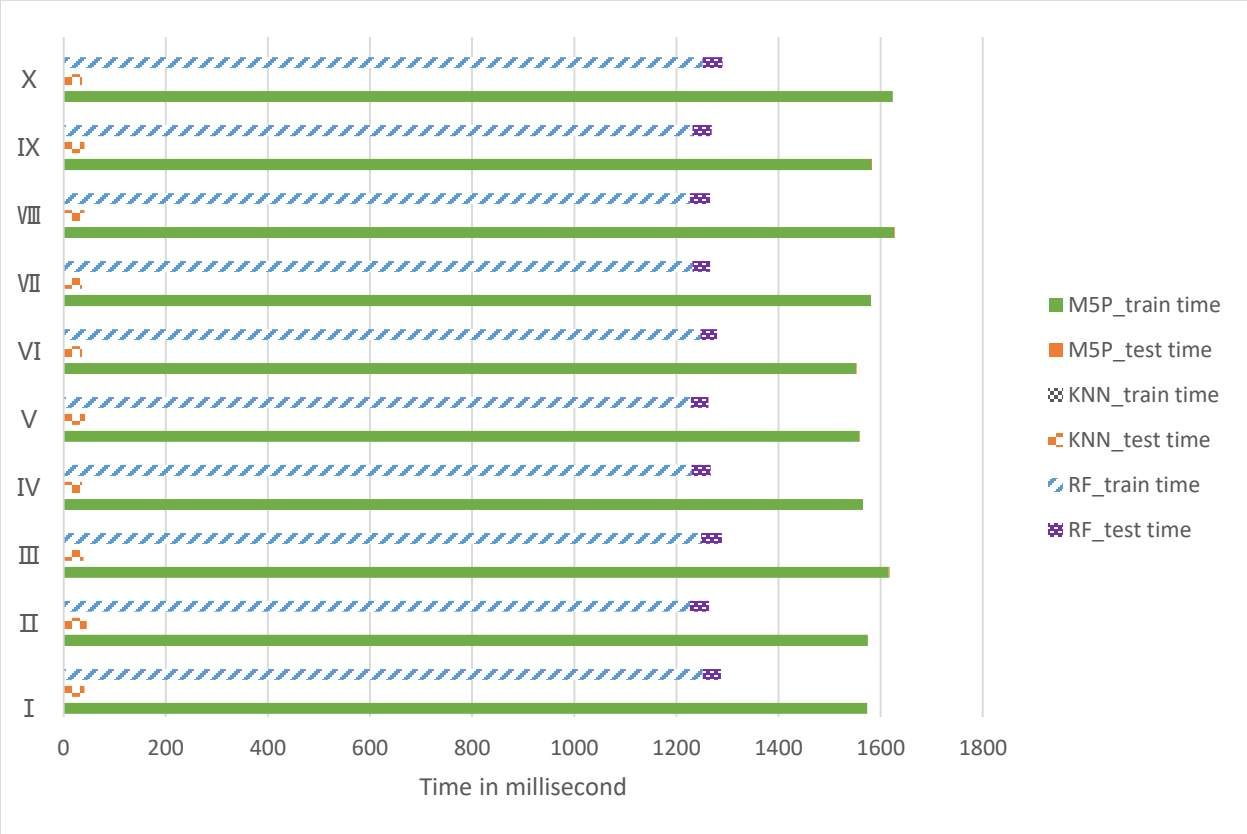


Figure 14: Comparisons of running time for each run for M5P, KNN, and RF

4.5 Discussions and limitations

The main objective for residual value of heavy construction equipment research is to capture the market trend and provide reliable results. As an ensemble method, the RF model has unique features that distinguish it from other data mining algorithms, by enhancing diversity through randomly "bagging" data and selecting different predictive features at each splitting node. Based on the comparison between selected models, it was found that RF has the highest CCs as well as the lowest prediction error. Besides, RF also overcomes the instability of using a single regression tree algorithm like M5P. When there is any perturbation in a dataset such as missing

values or a portion of outliers, a single tree will change to a different one, which indicates a relatively high sensitivity to perturbations in data. Furthermore, even if a cross-validation method is used when building a model, overfitting may still happen. Overfitting is a scenario in which the model performs well in a training set but not in a testing set. Compared to a single regression tree, RF assembles slightly different individual trees and calculates the mean value of them as the output. Resistance to overfitting and accurate predictions make RF a reliable method to conduct residual value research of heavy machines.

Complexity and interpretation is always considered as a criteria to evaluate models. A single regression tree provides a clear path of directions when a tested instance is given. The instance can be "visualized" and "tracked" though decision tree. Besides of single regression tree algorithm like M5P in this research, KNN also performs well in model complexity and interpretation. Although it is hard to plot data onto a diagram because of difficulties in expressing multi-dimensional points, KNN can be done simply in any auto-calculated spreadsheet. In fact, a few spreadsheets are developed by the author using Microsoft Excel, to store data and to implement KNN algorithms for the equipment owning company. To compare, on the other hand, the result of RF is hard to interpret because it is the average results of many unique individual trees. Both single decision tree algorithm and KNN can perform as a "white box" for users to understand, track, and manipulate. Practically, KNN is a user-friendly as well as low-cost algorithm that can be easily developed and maintained, especially for construction companies that have only a relative small dataset.

There is often a trade-off between model accuracy and speed (Wickelgren, 1977). As mentioned earlier, KNN and RF are good examples to explain this. Although KNN takes less time to build, it loses prediction accuracy compared to RF. On the other hand, both the RF and KNN models outperform the M5P model in terms of both accuracy and speed.

As only 3,044 records of articulated trucks are studied, one issue of this research is the size of the dataset. To test how sensitive the model is to the size of the dataset, two subsets are randomly selected from the original dataset, with 1,000 records and 2,000 records respectively. Different models are generated and analyzed based on these two new datasets. Results show that as the dataset becomes smaller, all the data mining models (M5P, KNN, and RF) perform worse according to the result of CC, MAE, RMSE, RAE, and RRSE. However, intra-comparisons between the three data mining models are similar to this research in terms of accuracy, predictive error, and running time. It is believed that a larger dataset will provide more stable and robust models to predict the residual value of articulated trucks.

4.6 Combination with CCM

Vorster (1980) provided an approach to combine the cumulative cost model of maintenance cost and residual value prediction. Figure 15 illustrates a trend of cumulative cost as machine age grows.

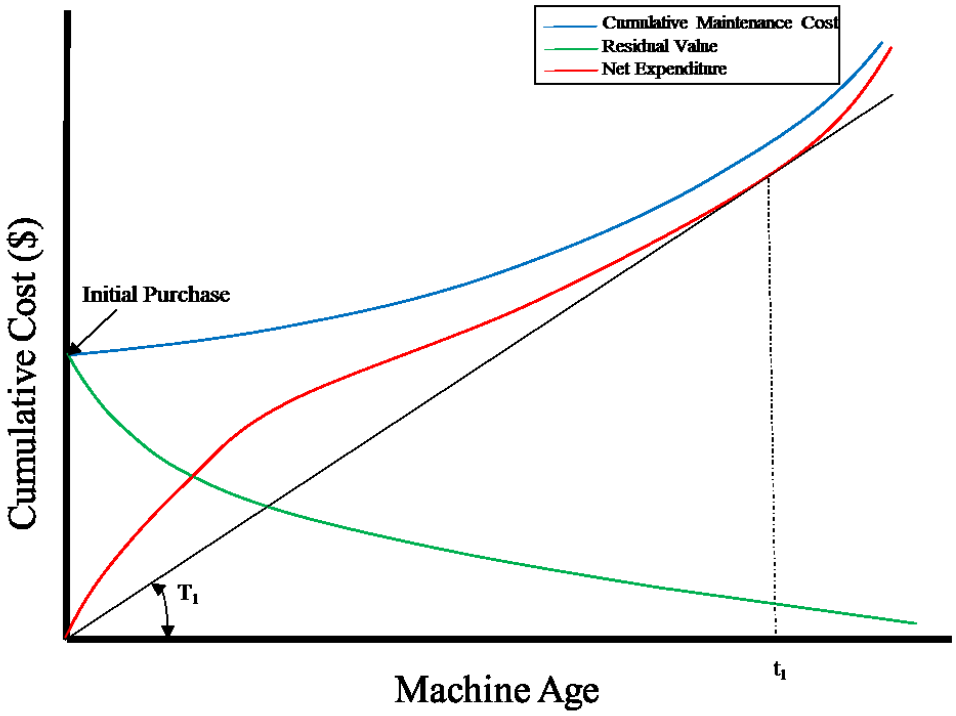


Figure 15: Combination of maintenance cost and residual value

Starting from the initial purchase price, the gross expenditure line increases as maintenance occurs, while the residual value drops dramatically in the early life of an asset. The net expenditure line equals the gross expenditure minus the residual value of the equipment at a given time. As the residual value decreases, the net expenditure converges with the gross expenditure line. This could be an intuitive explanation of how maintenance cost combines with the residual value of heavy equipment. Although the net expenditure keeps increasing, the increasing rate changes. Figure 15 shows that the increasing rate of the net expenditure "slows" down in the early half of the machine age, while it "speeds" up in the next phase. In Figure 15, T_1 is the slope of the line originating from the original point to the point on the net expenditure curve, which also represents the average expenditure during time t_1. The optimum economic life of equipment can be achieved when the line becomes a geometric tangent to the net expenditure

curve. In other words, that is the best time to sell the machine instead of keeping it because the average expenditure reaches its lowest point and it is going to increase as the machine grows old. However, in this research, there are limited machine units which have both maintenance cost data and residual value information from auction markets. The author of this research recommends the following: collect sufficient unit information that overlaps in these two fields, and draw the net expenditure line to build decision-making models.

Another issue to be noticed is that only machine age is used as a predictive variable in Figure 15. In this chapter, it is demonstrated that residual value of heavy equipment is impacted by multiple factors such as machine age, manufacturer, and even economic environments. Integration of multiple factors needs to be addressed.

4.7 Conclusion

This research used three kinds of data mining algorithms to build predictive models for the articulated trucks residual value. Detailed procedures were explained for data description and pre-processing, model generation, and model comparisons. All models in this research were validated by repetitive 10-time runs as well as ten-fold cross validation for each individual run. CC, MAE, RMSE, RAE, and RRSE were selected as criteria to compare performances for different models. Model interpretation and running time were also discussed to provide a thorough review of pros and cons for different models.

There are very few studies that discuss KNN and RF model applications in construction

equipment management. To the best knowledge of the author, I did not find any other studies on the applications of data mining technology on construction equipment residual value predictions other than the regression tree methods developed by Fan et al. (2008). There is also no discussion or comparison on the performance of a single regression tree, KNN, and RF in construction equipment residual value prediction. Compared to a single regression tree, the proposed KNN and RF models could provide more accurate results with less running time. While building a KNN model takes little time and is easy to interpret, it is less accurate than RF models. When decision makers are considering heavy equipment acquisition or disposal, this research could provide detailed instructions regarding model accuracy, interpretability, and speed. In short, the RF model offers superior performance in terms of prediction accuracy, and a KNN model is helpful for making a quick and simple simulation of residual percentage.

This study considers only a relative small dataset of articulate trucks in auction market with selected 10 features. Further research can incorporate more related features with expanded dataset, and it is recommended to try other data mining algorithms to get competitive prediction results of residual percentage of heavy equipment. Besides, combination of maintenance cost and residual value research is discussed in this chapter. A few recommendations are provided for future study based on the result of this research.

Chapter 5: Contributions and limitations

5.1 Research summary

This research concerns economic issues related to heavy construction equipment in terms of both maintenance cost and residual value prediction. Based on historical running repair data from January 2003 to July 2015 obtained from a construction company, cumulative cost models were applied to each rate group (B25 - B55, D35 - D95). Because running repair expenditures increase as machine age grows, models were developed in this research to help decision makers estimate maintenance costs. In current practice, equipment-owning companies always develop maintenance policies at a certain time interval for the same categories of machines. Therefore, different data sets were created and analyzed to select the optimum interval to represent the trend of running repairs as machine age grows. Generally, it is found that 500 SMR intervals and 1000 SMR intervals offer the best statistical performance. However, results of other data sets also provide reference points to decision makers when the maintenance policy is established based on other SMR intervals. Statistical results for each category of machine can be found in Appendix A. The residual plots for each regression model indicates that factors other than SMR records (e.g., working conditions, operator skills and other related attributes) could influence maintenance cost considerations.

This research also looked at the residual value prediction of heavy construction equipment. As construction equipment expenditures always comprise a significant part of a construction company's budget, this study is necessary and useful in that it can help decision makers to

reliably predict market price of heavy machines in auction markets. For this research, original auction records from the North American market were collected and processed. Additional attributes such as the machine's specifications and national macroeconomic indicators were also collected to devise better predictions. In current practice, a company uses single regression models when predicting residual values which turns out to be of low accuracy. Different data-mining algorithms, including a single regression tree, k nearest neighbor, and random forest, were utilized to build distinct prediction models for this research. The analysis of the output of these algorithms shows that the random forest (RF) algorithm has a better residual value prediction performance in terms of correlation coefficient and error rate. Regarding running time, k nearest neighbor (KNN) has a faster mode than the other two (M5P, RF). Compared to the single decision tree algorithm, which is the up-to-date method in residual value prediction, both KNN and RF demonstrate advantages of ease of use and better reliability. The models generated from this research could be helpful in residual value prediction of heavy construction equipment.

5.2 Research contribution

This research has a number of academic contributions in the heavy equipment management area. The main contributions are discussed below:

- This research examined a wide range of heavy construction and mining machines, with 15 different fleets of 250 units. For each unit, historical maintenance cost data were collected and processed. Cumulative cost models were built and validated for each category of

equipment.

- Compared to previous research, more SMR-interval-based datasets were created when building cumulative cost model (CCM). Instead of just using 500 SMR intervals, this research created three calendar-based datasets (seasonal, semi-annual, and annual), and SMR interval-based datasets (500, 1000, 1500, 2000, 2500, 5000, and 7500 SMR intervals, respectively). Statistical tests and analysis were provided to identify optimum datasets for different categories of machines. These different models can help a company to develop maintenance policies and provide a point of reference for fleet managers to estimate maintenance costs at certain SMR levels.
- Residual plots for CCM were included and discussed in maintenance cost research. The residual plots for each regression model indicate that factors apart from SMR records such as working conditions, operator skills and other related attributes could influence maintenance cost considerations. Methods of quantifying such mentioned attributes should be discussed and investigated in future research.
- For residual value prediction of heavy equipment, new attributes were included in this research. Equipment specifications, macroeconomic indicators including real GDP value and real GDP growth were included as predictive variables. A ranking algorithm of attributes regarding correlation with residual value percentage was also provided.
- In addition to the single decision tree method, an instance-based algorithm KNN and an ensemble algorithm RF were implemented to predict the percentage of residual value of

equipment. Compared to a single decision tree, both KNN and RF offer better performance in terms of correlated coefficient and prediction error.

- When comparing the performance of different algorithms in residual value prediction, a series of criteria were set, and detailed discussions provided regarding the prediction accuracy, interpretability, running time, etc.

5.3 Research limitation

This research has a few limitations:

- Due to updates of the construction company's database, historical maintenance data are only available since January 2003. Previous data are not available, which limits the number of machines qualified for this study.
- CCM models were built for only 15 equipment fleets, mainly in rigid frame trucks and excavators. Because of an insufficient amount of data, this research did not test for maintenance costs for other categories of heavy equipment.
- The maintenance cost research work was based on historical maintenance cost data from a certain construction company; therefore the research output does not represent all construction equipment.
- Regarding residual value prediction, this research exemplified a prediction model only for articulated truck categories. A more comprehensive prediction model covering other categories of construction equipment could be built based on sufficient data from other

categories of equipment.

- As it is validated that for all three algorithms, residual value prediction model are sensitive to data size (see *Chapter 4*). Because the auction records employed in this thesis were limited, it is possible that some results were biased. The data-mining models would be more useful and reliable if sufficient instances of auctions records were available.

5.4 Recommendations for future study.

There are a few recommendations for future research in the same area:

- For a maintenance cost study of heavy machines, it is better to obtain historical data with a wide-range of records for life-time analysis. Larger data pools with more records of units will be helpful to generalize and identify maintenance cost patterns for each category of heavy machines.
- For residual value predictions in this research, the basic auction records were provided from the company. These records are stored in MS Excel spreadsheets. In future studies, it will be better if an automatic management system can be generated to query as many as auction instances as possible with up-to-date auction records. Models built on limited data sources might influence prediction accuracy.
- In the research about residual value prediction, a single decision tree, an instance-based method (KNN) and an ensemble method (RF) were compared to obtain an optimum method with the best performance. For future studies, many other algorithms such as boosting

methods could be tested to find more accurate models.

- Besides the historical maintenance cost data and predictable features of residual value, other related information could be identified and collected for future research. New models of heavy equipment in recent years are equipped with telematics systems which help record and process data, such as kinematic global positioning system (GPS) and application of radio frequency identification (RFID). Such information can be collected and used in heavy construction equipment research.

Reference

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning, 6*(1), 37-66.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation, 9*(7), 1545-1588.

Bates, J., Rayner, A., & Custance, P. (1979). Inflation and farm tractor replacement in the US: a simulation model. *American Journal of Agricultural Economics, 61*(2), 331-334.

Bayzid, S. M. (2014). *Modeling Maintenance Cost for Road Construction Equipment.* University of Alberta.

Bayzid, S. M., Mohamed, Y., & Al-Hussein, M. (2016). Prediction of maintenance cost for road construction equipment: a case study. *Canadian Journal of Civil Engineering, 43*(5), 480-492.

Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: forecasting and control, revised ed*: Holden-Day.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Chen, J.-H. (2008). KNN based knowledge-sharing model for severe change order disputes in construction. *Automation in Construction, 17*(6), 773-779.

Cross, T., & Perry, G. (1996). Remaining value functions for farm equipment. *Applied engineering in agriculture, 12*(5), 547-553.

Cross, T. L., & Perry, G. M. (1995). Depreciation patterns for agricultural machinery. *American Journal of Agricultural Economics, 77*(1), 194-204.

Cubbage, F. W., Burgess, J. A., & Stokes, B. J. (1991). Cross-sectional estimates of logging equipment resale values. *For. Prod. J, 41*(10), 16-22.

Dang, Y., Zhang, Y., Zhang, D., & Zhao, L. (2005). *A KNN-based learning method for biology species categorization.* Paper presented at the International Conference on Natural Computation.

Douglas, J. (1975). *Construction equipment policy*: McGraw-Hill.

Drakatos, P. (1975). *The Application of Similarity Methods During the Experimental Research on Soil Compaction Machines.* PH. D., Thesis, Thessaloniki.

Duncan, K. C. (2015). The Effect of Federal Davis-Bacon and Disadvantaged Business Enterprise Regulations on Highway Maintenance Costs. *ILR Review, 68*(1), 212-237.

Edwards, D. J., Holt, G. D., & Harris, F. C. (1999). *Predicting maintenance expenditure on construction plant: model development and performance analysis*.

Edwards, D. J., Holt, G. D., & Harris, F. C. (2000). Estimating life cycle plant maintenance costs. *Construction Management & Economics, 18*(4), 427-435.

Fairbanks, G. E., Larson, G. H., & Chung, D.-S. (1971). Cost of using farm machinery. *Transactions of the ASAE, 14*(1), 98-0101.

Fan, H., AbouRizk, S., & Kim, H. (2007). Building intelligent applications for construction equipment management. *Computing in Civil Engineering (2007)*, 192-199.

Fan, H., AbouRizk, S., Kim, H., & Zaïane, O. (2008). Assessing residual value of heavy construction equipment using predictive data mining model. *Journal of Computing in Civil Engineering, 22*(3), 181-191.

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

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining* (Vol. 21): AAAI press Menlo Park.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1): Springer series in statistics Springer, Berlin.

Geraerds, W. (1983). *The Cost of Downtime for Maintenance: Preliminary Considerations*: University of Technology. Department of industrial engineering & management Science.

Gillespie, J. S. (2004). The replace repair decision for heavy equipment VTRC ; 05-R8: Charlottesville, Va., Virginia Transportation Research Council, [2004].

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Han, J. (2006). Data mining : concepts and techniques. In M. Kamber (Ed.), (2nd ed. ed.). San Francisco, Calif. :: Morgan Kaufmann ;.

Harvey, R. R., & McBean, E. A. (2014). Predicting the structural condition of individual sanitary sewer pipes with random forests. *Canadian Journal of Civil Engineering, 41*(4), 294-303.

Ho, T. K. (1995). *Random decision forests.* Paper presented at the Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence, 20*(8), 832-844.

Jones, B. W. (1982). *Inflation in engineering economic analysis*: John Wiley & Sons.

Kannan, G. (2011). Field studies in construction equipment economics and productivity. *Journal of Construction Engineering and Management*(10), 823. doi:10.1061/(ASCE)CO.1943-7862.0000335

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika, 70*(1), 163-173.

Kononenko, I. (1994). *Estimating attributes: analysis and extensions of RELIEF.* Paper presented at the European conference on machine learning.

Lee, B.-H., & Scholz, M. (2006). A comparative study: Prediction of constructed treatment wetland performance with k-nearest neighbors and neural networks. *Water, Air, and Soil Pollution, 174*(1-4), 279-301.

Lee, M.-J., Hanna, A. S., & Loh, W.-Y. (2004). Decision tree approach to classify and quantify cumulative impact of change orders on productivity. *Journal of Computing in Civil Engineering, 18*(2), 132-144.

Lucko, G. (2003). A statistical analysis and model of the residual value of different types of heavy construction equipment.

Lucko, G. (2010). Modeling the residual market value of construction equipment under changed economic conditions. *Journal of Construction Engineering and Management, 137*(10), 806-816.

Lucko, G. (2011). Modeling the residual market value of construction equipment under changed economic conditions. *Journal of Construction Engineering and Management, 137*(10), 806-816.

Lucko, G., Anderson-Cook, C. M., & Vorster, M. C. (2006). Statistical considerations for predicting residual value of heavy equipment. *Journal of Construction Engineering and Management, 132*(7), 723-732.

Lucko, G., & Mitchell Jr, Z. W. (2010). Quantitative research: Preparation of incongruous economic data sets for archival data analysis. *Journal of Construction Engineering and Management, 136*(1), 49-57.

Lucko, G., Vorster, M. C., & Anderson-Cook, C. M. (2007). Unknown element of owning costs—Impact of residual value. *Journal of Construction Engineering and Management, 133*(1), 3-9.

Manatakis, E., & Drakatos, P. (1978). Computerized method of constructions equipment cost estimating. *AACE Transactions*, 352-358.

Manatakis, E. K., & Drakatos, P. A. (1993). A new method for the analysis of operating costs of construction equipment. *International journal of production economics, 32*(1), 13-21.

McNeill, R. C. (1979). Depreciation of farm tractors in British Columbia. *Canadian Journal of*

Agricultural Economics/Revue canadienne d'agroeconomie, 27(1), 53-58.

Mitchell Jr, Z. W. (1998). A statistical analysis of construction equipment repair costs using field data & the cumulative cost model.

Mitchell, Z., Hildreth, J., & Vorster, M. (2010). Using the cumulative cost model to forecast equipment repair costs: Two different methodologies. *Journal of Construction Engineering and Management, 137*(10), 817-822.

Moore, T. K. (1976). *Environmental Effects on Maintenance Costs for Aircraft Equipment*. Retrieved from

Nunnally, S. W. (2004). Construction methods and management (6th ed. ed.). Upper Saddle River, N.J. :: Pearson Prentice Hall.

Outlaw, K., & Young, K. (2014). Construction Equipment Making the Grade, but. *MANUFACTURING ENGINEERING, 153*(4), 87-+.

Peck, R., & Devore, J. L. (2011). *Statistics: The exploration & analysis of data.* Cengage Learning.

Perspectives, G. C., & Economics, O. (2011). Global construction 2020: A global forecast for the construction industry over the next decade to 2020. *Final report, 3*.

Peurifoy, R. L. (2006). Construction planning, equipment, and methods. In C. J. Schexnayder & A. Shapira (Eds.), (7th ed. ed.). Boston :: McGraw-Hill Higher Education.

Peurifoy, R. L., Schexnayder, C. J., & Shapira, A. (2006). *Construction planning, equipment, and methods* (7th ed.). Boston: McGraw-Hill Higher Education.

Ponnaluru, S. S., Marsh, T. L., & Brady, M. (2012). Spatial price analysis of used construction equipment: the case of excavators. *Construction Management and Economics, 30*(11), 981-994.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems, 9*(2), 181-199.

Quinlan, J. R. (1992). *Learning with continuous classes.* Paper presented at the 5th Australian joint conference on artificial intelligence.

Reid, D. W., & Bradford, G. L. (1983). On optimal replacement of farm tractors. *American Journal of Agricultural Economics, 65*(2), 326-331.

Robnik-Šikonja, M., & Kononenko, I. (1997). *An adaptation of Relief for attribute estimation in regression.* Paper presented at the Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97).

Rosa, J., Ebecken, N. F., & Costa, M. (2003). Towards on an optimized parallel KNN-fuzzy classification approach. *WIT Transactions on Information and Communication Technologies, 29*.

Soltynski, A. (1968). Physical similarity and scale effects in soil-machine systems. *Similatudes physiques et effets de la modification d'echelle (French), 5*, 31-43. doi:10.1016/0022-4898(68)90109-2

Terborgh, G. W. (1949). Dynamic equipment policy (1st ed. ed.). New York: McGraw-Hill Book

Co.

Valli, P., Jeyasehar, C. A., & Saravanan, J. (2013). A General Analysis of Equipment Cost and Management in Flyover Project. *IUP Journal of Structural Engineering, 6*(4), 60.

Vorster, M. (1980). *A systems approach to the management of civil engineering construction equipment.* Stellenbosch: Stellenbosch University.

Vorster, M. (2004). How to estimate market value. *Construction Equipment, 107*(6), 64.

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

Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica, 41*(1), 67-85.

Yip, H.-l., Fan, H., & Chiang, Y.-h. (2014). Predicting the maintenance cost of construction equipment: Comparison between general regression neural network and Box–Jenkins time series models. *Automation in Construction, 38*, 30-38. doi:10.1016/j.autcon.2013.10.024

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies, 58*, 308-324.

Zhou, J., Li, X., & Mitri, H. S. (2016). Classification of Rockburst in Underground Projects: Comparison of Ten Supervised Learning Methods. *Journal of Computing in Civil Engineering*, 04016003.

Appendices

Appendix A: Data analysis result for 15 fleets – CCM Model

B25 – CAT 777 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	13345.29957	8.5E-100	0.882637184	0.882139029	8434.3	0	2.5898	2245
	x2	222.33484	8.76E-26						
Seasonal	x	12823.78683	6.07E-32	0.881732708	0.880219491	2754.8	0	2.5303	741
	x2	238.5608039	1.18E-10						
Semi-annual	x	13589.54984	2E-19	0.884765423	0.881898496	1493.4	5.69E-183	2.5586	391
	x2	213.7080597	2.69E-05						
Annual	x	15168.26595	1.18E-10	0.879741766	0.873967239	709.6	1.10E-89	3.1346	196
	x2	172.04906	0.024728						
500 Interval	x	9310.977959	1.95E-51	0.861040947	0.860423625	5716.1	0	2.072	1847
	x2	334.8064865	4.25E-47						
1000 Interval	x	9247.01957	3.21E-26	0.861584917	0.860351531	2872.7	0	2.0729	925
	x2	339.659074	5.96E-25						
1500 Interval	x	9584.404406	3.47E-19	0.859701552	0.857838319	1875.1	1.73E-261	2.0911	614
	x2	321.6253603	9.88E-16						
2000 Interval	x	9027.252818	3.14E-13	0.861077238	0.858590507	1419.4	8.65E-197	2.0875	465
	x2	348.2919331	1.15E-13						
2500 Interval	x	8859.550128	1.32E-10	0.86286235	0.859746904	1148.3	5.96E-158	2.0893	367
	x2	351.6910519	1.41E-11						
5000 Interval	x	9227.208775	7.33E-06	0.859951149	0.853399144	534.21	9.27E-75	2.2306	176
	x2	347.027452	8.91E-06						
7500 Interval	x	7229.541452	0.000689	0.869911843	0.861151469	431.32	1.33E-57	1.9315	131
	x2	375.8743334	2.7E-06						

B30 – CAT 785 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	5135.5417	1.7232E-13	0.90646082	0.905733735	7287.4121	0	1.6381	1506
	x2	650.58792	4.296E-100						
Seasonal	x	5802.6343	1.3576E-06	0.90589919	0.903724044	2421.1657	1.50E-258	1.6046	505
	x2	627.32533	7.0474E-33						
Semi-annual	x	5688.5583	0.00113075	0.90580929	0.901306444	1168.4361	4.56E-125	1.6741	245
	x2	624.05009	2.592E-16						
Annual	x	2414.1674	0.26858736	0.90541726	0.898021426	708.38372	3.36E-76	1.5625	150
	x2	775.08126	2.6889E-14						
500 Interval	x	3079.447	2.8621E-06	0.89474843	0.893883603	5432.1685	0	1.2621	1281
	x2	728.92794	1.016E-116						
1000 Interval	x	3247.3173	0.00049171	0.89405611	0.892311721	2675.15	0	1.2733	636
	x2	718.82087	5.1487E-58						
1500 Interval	x	3297.5895	0.00420612	0.89312185	0.89048643	1754.8544	2.28E-204	1.2857	422
	x2	719.10302	2.0212E-38						
2000 Interval	x	3100.9811	0.01924877	0.8945101	0.89096686	1322.8145	8.23E-153	1.275	314
	x2	726.62008	7.2682E-30						
2500 Interval	x	2749.381	0.0646914	0.89357679	0.889186218	1057.9522	5.00E-123	1.2944	254
	x2	741.1859	7.9035E-25						
5000 Interval	x	2495.9534	0.26368117	0.8849	0.875287071	445.90965	6.61E-55	1.4171	118
	x2	749.6635	6.2361E-12						
7500 Interval	x	447.26749	0.86798787	0.8923974	0.877429794	306.85785	2.94E-36	1.3078	76
	x2	822.48234	1.6846E-09						

B45 – CAT 930D Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	17494.71121	4.23903E-34	0.916618275	0.91593259	8684.49815	0	8.3656	1582
	x2	1020.068054	1.32493E-79						
Seasonal	x	16900.74091	3.98208E-11	0.913517559	0.9114201	2735.827597	9.95E-276	8.6912	520
	x2	1042.994389	6.6454E-27						
Semi-annual	x	15645.90427	4.16755E-06	0.914437007	0.91030939	1405.379388	8.54E-141	7.9038	265
	x2	1078.395559	1.29617E-16						
Annual	x	19699.05889	6.78657E-05	0.922408884	0.9143067	790.556881	3.37E-74	9.0366	135
	x2	956.8185785	1.16229E-07						
500 Interval	x	12150.8062	1.37896E-22	0.902839686	0.90217231	7638.244406	0	6.817	1646
	x2	1166.103232	3.8468E-116						
1000 Interval	x	12503.66105	1.25676E-12	0.902446371	0.90110462	3783.565709	0	6.8739	820
	x2	1148.80938	1.49452E-57						
1500 Interval	x	12125.99296	2.01082E-08	0.903140277	0.90111655	2526.86057	3.60E-275	6.8784	544
	x2	1167.080635	1.22918E-39						
2000 Interval	x	12943.33482	2.07242E-07	0.903025333	0.90031005	1881.018247	4.13E-205	6.989	406
	x2	1129.544493	5.15357E-29						
2500 Interval	x	12815.34452	7.59132E-06	0.899564973	0.89610449	1424.113036	4.02E-159	7.138	320
	x2	1138.209226	2.42302E-22						
5000 Interval	x	10957.939	0.007621079	0.898144046	0.89079834	661.3339831	7.95E-75	7.2591	152
	x2	1181.19546	4.06484E-12						
7500 Interval	x	11490.30201	0.023653836	0.900102423	0.88864516	432.4921365	1.92E-48	7.386	98
	x2	1162.988014	3.88944E-08						

B45 – KOM 830E Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	3732.256167	0.027722736	0.966988437	0.9639855	5038.295515	5.52E-255	1.0426	346
	x2	1749.634496	9.79679E-58						
Seasonal	x	1565.801133	0.596066816	0.9693373	0.96013495	1770.323185	6.27E-85	1.0046	114
	x2	1878.512495	1.1444E-21						
Semi-annual	x	2632.827083	0.558348935	0.962880701	0.93876165	557.7135237	5.69E-31	0.9827	45
	x2	1812.513945	1.84736E-09						
Annual	x	7285.951639	0.235625024	0.963290962	0.92984874	406.7393411	1.84E-22	1.3844	33
	x2	1529.721223	2.14553E-05						
500 Interval	x	2393.425572	0.138379711	0.964177402	0.96029792	3593.197834	3.12E-193	0.8377	867
	x2	1809.11454	1.15784E-55						
1000 Interval	x	1942.049086	0.393735966	0.96457332	0.95672918	1797.002666	5.92E-96	0.8244	134
	x2	1837.27949	4.99277E-29						
1500 Interval	x	1938.056598	0.50250893	0.962908403	0.95098781	1129.272352	1.87E-62	0.8803	89
	x2	1843.201192	4.51891E-19						
2000 Interval	x	1858.210116	0.579546975	0.962764965	0.9468075	840.3338747	1.16E-46	0.8923	67
	x2	1857.163106	1.04706E-14						
2500 Interval	x	388.5057128	0.91843597	0.962102167	0.94175123	647.3616935	1.81E-36	0.9069	53
	x2	1930.043198	2.88469E-12						
5000 Interval	x	1476.385537	0.82053564	0.959306078	0.91200181	259.3106351	1.57E-15	1.0864	24
	x2	2143.484846	9.29102E-06						
7500 Interval	x	1635.14648	0.777727848	0.972925597	0.90445397	269.5144201	6.66E-12	0.8698	17
	x2	1779.493206	2.8007E-05						

B50- HIT EH4500 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	-24185.9961	6.23834E-11	0.96642	0.961521529	3036.252154	1.07E-155	8.0706	213
	x2	1739.469424	6.17337E-47						
Seasonal	x	-24049.7341	3.39439E-05	0.9550864	0.943970291	999.4534721	1.35E-63	9.3225	96
	x2	1627.876512	4.18157E-19						
Semi-annual	x	-19493.0806	0.019524812	0.9498407	0.930393329	511.2852745	2.26E-35	11.362	56
	x2	1492.299614	2.15162E-09						
Annual	x	-23114.3865	0.028032609	0.9535815	0.91870086	308.1468652	2.91E-20	11.084	32
	x2	1581.057185	7.90085E-07						
500 Interval	x	-28704.5661	1.9095E-12	0.9592964	0.953904117	2274.295543	2.07E-134	7.9442	195
	x2	1760.39728	2.88648E-41						
1000 Interval	x	-28750.5591	7.2955E-07	0.9589516	0.948107332	1121.350901	8.40E-67	8.0579	98
	x2	1759.815172	3.2038E-21						
1500 Interval	x	-29475.8423	3.07808E-05	0.9591912	0.942928517	752.1437358	1.08E-44	8.1301	66
	x2	1788.717828	4.67547E-15						
2000 Interval	x	-25718.1615	0.001513384	0.9564147	0.934673346	526.6444509	6.59E-33	8.4648	50
	x2	1672.714219	1.08566E-10						
2500 Interval	x	-29349.4882	0.002199723	0.9562847	0.928076161	404.6926276	2.12E-25	8.4383	39
	x2	1775.399315	8.107E-09						
5000 Interval	x	-33543.8616	0.010190802	0.9682667	0.907576523	259.3575624	6.43E-13	7.1348	19
	x2	1895.983074	1.02968E-05						
7500 Interval	x	-27910.9549	0.248576683	0.9514844	0.834982702	88.25373222	3.53E-06	10.002	11
	x2	1756.511649	0.020214526						

B55 – HIT EH5000 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	48067.194	2.6172E-34	0.952471451	0.95072557	6011.995779	0	31.401	602
	x2	1067.5912	6.55759E-21						
Seasonal	x	48854.373	2.11022E-12	0.950410985	0.94516304	1916.57566	9.66E-131	32.601	202
	x2	1043.6452	2.16731E-07						
Semi-annual	x	44919.521	2.65388E-06	0.95289961	0.94263392	1031.793575	6.05E-68	33.0698	104
	x2	1198.2402	1.89546E-05						
Annual	x	38713.867	0.002524566	0.962020267	0.940837007	620.5808834	5.04E-35	26.4542	51
	x2	1312.7444	0.000564535						
500 Interval	x	21831.712	6.33319E-10	0.972270658	0.970097869	8292.371636	0	18.0899	475
	x2	1944.781	9.3879E-57						
1000 Interval	x	20936.165	3.26109E-05	0.972329907	0.967938155	4111.392054	1.92E-182	18.2784	236
	x2	1978.5373	1.84538E-29						
1500 Interval	x	22671.035	0.000160246	0.973127603	0.96662841	2860.819577	3.09E-124	17.807	160
	x2	1912.596	2.89789E-20						
2000 Interval	x	18903.124	0.007801928	0.972780171	0.963924828	2072.799591	6.25E-91	18.2469	118
	x2	2047.636	2.99005E-16						
2500 Interval	x	20003.911	0.011910303	0.972330354	0.961160032	1616.47157	7.96E-72	17.6993	94
	x2	2009.1485	1.03205E-12						
5000 Interval	x	21314.322	0.066179971	0.97199729	0.948633592	763.6382352	2.55E-34	18.1227	46
	x2	1969.9028	1.49126E-06						
7500 Interval	x	16050.483	0.240574666	0.972156997	0.940036904	558.6506684	4.84E-25	19.1538	34
	x2	2182.8467	1.06899E-05						

B55 – KOM 930E Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	47428.066	9.74382E-40	0.894028534	0.892722783	3572.858803	0	24.9581	849
	x2	1547.9493	9.33195E-25						
Seasonal	x	17425.874	0.000194589	0.943492419	0.939874242	2437.724118	1.69E-182	13.3993	294
	x2	3032.2085	2.71236E-38						
Semi-annual	x	54850.818	1.77291E-08	0.871750616	0.863138025	445.2236989	6.84E-59	27.769	133
	x2	1153.8829	0.005009273						
Annual	x	62601.769	1.42779E-06	0.864208723	0.848860192	235.477001	1.46E-32	27.7902	76
	x2	747.24064	0.150553262						
500 Interval	x	20195.192	3.33202E-13	0.945023264	0.943773291	7253.974015	0	13.0145	846
	x2	3031.8272	1.0654E-104						
1000 Interval	x	20329.207	2.46467E-07	0.944824889	0.942324474	3613.188065	8.55E-266	13.1437	424
	x2	3022.2458	8.36114E-53						
1500 Interval	x	20179.805	2.65817E-05	0.945131438	0.941377457	2420.164866	2.00E-177	13.0743	283
	x2	3021.4535	4.69292E-36						
2000 Interval	x	21259.162	0.000122939	0.9450042	0.940097243	1847.194726	1.02E-135	13.1574	217
	x2	2971.9144	3.07534E-27						
2500 Interval	x	22692.394	0.000272195	0.943916613	0.937740687	1439.01564	2.80E-107	13.5807	173
	x2	2876.9171	4.34602E-21						
5000 Interval	x	23986.327	0.007230825	0.945648099	0.932631409	704.6441318	1.60E-51	14.2844	83
	x2	2800.663	8.31635E-11						
7500 Interval	x	28842.507	0.009342459	0.945766765	0.925493049	453.4108251	3.31E-33	15.3166	54
	x2	2474.8597	1.1648E-06						

D35 – EX500 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	MSE(E+10)	Observations
All Points	x	-5197.984372	7.59088E-12	0.8819505	0.880797871	3623.446011	0	0.437	972
	x2	1311.89146	5.8354E-116						
Seasonal	x	-5337.282794	4.55727E-08	0.8753866	0.873177165	1787.816835	1.21E-230	0.3818	511
	x2	1311.438511	1.01928E-66						
Semi-annual	x	-4843.218722	0.00011114	0.8764999	0.872652302	1036.185325	4.44E-133	0.3667	294
	x2	1262.943925	1.78145E-38						
Annual	x	-4429.522445	0.008013689	0.87771	0.871222737	620.8348079	2.12E-79	0.4079	175
	x2	1236.74715	7.5768E-23						
500 Interval	x	-1417.663431	0.029048365	0.87989	0.878289828	2563.995136	0	0.3128	702
	x2	968.280776	5.43411E-86						
1000 Interval	x	-2520.987457	0.013392752	0.8804697	0.876811071	1127.00982	1.33E-141	0.3203	308
	x2	1065.995137	2.18876E-41						
1500 Interval	x	-814.245538	0.46318262	0.8771069	0.872287584	831.4781161	1.57E-106	0.3185	235
	x2	911.8430432	3.307E-28						
2000 Interval	x	-1361.004189	0.293268172	0.8799663	0.873492154	634.1311761	4.26E-80	0.3163	175
	x2	954.8809407	1.25273E-22						
2500 Interval	x	-925.0097542	0.518111829	0.879952	0.87206434	520.4301019	8.02E-66	0.3383	144
	x2	922.1807753	1.45725E-18						
5000 Interval	x	-2450.132613	0.270806372	0.8783833	0.860579855	227.5104701	2.79E-29	0.3451	65
	x2	1028.31058	1.49727E-09						
7500 Interval	x	2324.233434	0.339322892	0.8894427	0.862355898	164.9241474	4.79E-20	0.3741	43
	x2	659.9174094	4.56941E-05						

D40 – EX600 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	-3360.54557	0.0004998	0.9525247	0.938921144	772.44833	3.16E-51	0.077	79
	x2	743.0931	1.073E-22						
Seasonal	x	-3271.30601	0.0332431	0.9290744	0.903576165	275.08488	1.7457E-24	0.1094	44
	x2	727.9034074	1.444E-10						
Semi-annual	x	-3614.4935	0.2069544	0.9023144	0.856577469	110.84304	1.5519E-12	0.2173	26
	x2	782.1473865	4.835E-05						
Annual	x	-3654.44137	0.446566	0.8629608	0.775496222	40.931676	4.36628E-06	0.328	15
	x2	786.3725035	0.010702						
500 Interval	x	-991.515337	0.3591961	0.9108549	0.894838013	347.40046	4.28E-36	0.0798	70
	x2	528.2549397	8.467E-11						
1000 Interval	x	-1242.92369	0.4358942	0.9097057	0.876666488	166.23581	1.25E-17	0.0884	35
	x2	542.3439201	5.1E-06						
1500 Interval	x	-951.270145	0.6326942	0.9067301	0.857035964	106.93724	9.76E-12	0.095	24
	x2	518.908108	0.0003516						
2000 Interval	x	-126.725377	0.9610681	0.8914023	0.817495796	61.562242	1.16E-07	0.1078	17
	x2	459.5335951	0.0132284						
2500 Interval	x	-742.633706	0.77444	0.9065626	0.822452023	63.065286	4.30E-07	0.089	15
	x2	509.9322481	0.0078117						
5000 Interval	x	-2866.21025	0.3181557	0.9673398	0.760807774	74.045792	0.000691687	0.051	7
	x2	689.8458537	0.0079011						
7500 Interval	x	-373.670477	0.9576627	0.9125036	0.368755397	10.429041	0.213891717	0.1029	4
	x2	473.0408379	0.4041913						

D50 – EX800 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	2577.72141	0.068024004	0.916789	0.913938453	2093.35219	1.50E-205	1.2796	382
	x2	963.555461	9.99695E-37						
Seasonal	x	-75.329809	0.967022392	0.9191501	0.913578676	1102.753625	2.4659E-106	1.1018	196
	x2	1060.82818	1.18598E-24						
Semi-annual	x	548.174831	0.845968344	0.9039265	0.893870806	512.7741689	7.33028E-56	1.4012	111
	x2	1035.01545	2.03857E-11						
Annual	x	-977.77282	0.75509964	0.9176729	0.902640613	401.2802458	2.01266E-39	1.1684	74
	x2	1078.93713	4.03383E-10						
500 Interval	x	-520.61369	0.748505759	0.8981419	0.894761942	1437.265075	4.02E-162	1.395	328
	x2	1170.52649	4.28222E-35						
1000 Interval	x	-953.0595	0.679743892	0.8980064	0.891245743	717.57006	3.16E-81	1.3983	165
	x2	1203.47997	7.90959E-19						
1500 Interval	x	-1128.7016	0.694363169	0.8975703	0.887267262	468.8096667	2.28E-53	1.4335	109
	x2	1202.73738	8.75239E-13						
2000 Interval	x	-2451.0818	0.461148158	0.9010183	0.887281002	364.1149969	1.35E-40	1.3971	82
	x2	1304.16496	7.08348E-11						
2500 Interval	x	1387.28174	0.703784757	0.9022585	0.884834001	290.7785643	3.15E-32	1.4877	65
	x2	1064.31314	2.18509E-07						
5000 Interval	x	-1649.5229	0.759492114	0.8949862	0.860454473	136.3608628	4.34E-16	1.5058	34
	x2	1244.96497	9.08473E-05						
7500 Interval	x	-2597.6301	0.751145877	0.9076943	0.843440982	83.58528124	3.39E-09	1.6484	19
	x2	1288.79163	0.006472907						

D50 – EX850 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	5234.84357	8.7856E-10	0.9287554	0.927645308	6289.943692	0	0.9306	967
	x2	1188.31642	2.2052E-106						
Seasonal	x	11110.778	2.27224E-14	0.9089204	0.906388877	2150.5618	1.1978E-224	1.3494	433
	x2	960.173138	1.94977E-30						
Semi-annual	x	13953.2388	2.59822E-12	0.8949738	0.890271539	1001.268225	1.9855E-115	1.5024	237
	x2	774.786887	7.603E-13						
Annual	x	2625.42682	0.260862441	0.9278079	0.920659968	963.8950569	5.70989E-86	1.1301	152
	x2	1496.2656	8.88816E-23						
500 Interval	x	11960.2589	7.51319E-30	0.8704668	0.86892582	2462.890298	0	1.1507	735
	x2	604.543616	3.20531E-25						
1000 Interval	x	11860.6335	2.1406E-15	0.8728094	0.869695584	1242.060801	1.46E-162	1.1332	364
	x2	612.90701	1.62529E-13						
1500 Interval	x	11598.1251	2.91976E-10	0.8723985	0.867680461	817.0090632	2.56E-107	1.1332	241
	x2	625.506396	1.43017E-09						
2000 Interval	x	11076.514	1.595E-07	0.8768237	0.870583824	640.6599905	2.58E-82	1.1326	182
	x2	670.505623	2.2368E-08						
2500 Interval	x	10863.7164	8.80579E-06	0.8720754	0.863842347	466.9715632	1.21E-61	1.1783	139
	x2	681.257885	9.21541E-07						
5000 Interval	x	9489.37329	0.006238604	0.8805149	0.862162651	224.7618815	1.35E-28	1.0905	63
	x2	755.155062	0.000261895						
7500 Interval	x	5312.38728	0.233618392	0.902791	0.874657486	181.0988114	3.74E-20	1.3161	41
	x2	1093.57164	5.63803E-05						

D60 – EX1200 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	57734.52502	6.04908E-35	0.945423563	0.94030426	1784.261339	2.18E-130	9.7226	208
	x2	-100.4245713	0.45377302						
Seasonal	x	49918.43623	2.42481E-13	0.937728209	0.927105491	752.9317841	1.3E-60	10.9787	102
	x2	203.6011617	0.335138041						
Semi-annual	x	46069.06194	2.87303E-06	0.923714795	0.904145973	332.9892964	4.24E-31	13.5566	57
	x2	320.472908	0.315964278						
Annual	x	41252.727	0.001985943	0.905007734	0.87182615	157.1983516	2.81E-17	16.919	35
	x2	474.9972502	0.27597196						
500 Interval	x	51130.2044	1.14285E-29	0.946450504	0.940065356	1458.13075	3.57E-105	7.514	167
	x2	61.54794058	0.646608456						
1000 Interval	x	49731.01416	8.6595E-15	0.945419321	0.932558581	710.181569	4.43E-52	7.6569	84
	x2	113.7112862	0.556720117						
1500 Interval	x	49286.97538	6.86205E-10	0.946858478	0.927355857	481.0772765	1.05E-34	7.7975	56
	x2	153.9527847	0.52705091						
2000 Interval	x	52649.42817	1.79462E-08	0.948243993	0.92127589	357.2678602	2.30E-25	7.7106	41
	x2	5.611724633	0.983607328						
2500 Interval	x	51605.53788	5.50878E-07	0.950330414	0.916470105	296.562195	1.73E-20	7.1797	33
	x2	47.53594922	0.876443369						
5000 Interval	x	49801.34834	0.001154504	0.944414994	0.87404266	127.4284731	1.04E-09	7.8199	17
	x2	121.1567315	0.80084445						
7500 Interval	x	61325.23498	0.002087415	0.952562035	0.836180039	90.36073084	3.23E-06	8.5609	11
	x2	-304.9606176	0.558207931						

D65 – EX1900 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	19410.5633	1.511E-22	0.9815918	0.9772947	6318.85545	1.14E-205	3.7337	239
	x2	1016.93987	8.916E-52						
Seasonal	x	13748.0585	2.774E-07	0.9834474	0.9735779	3059.78491	8.9066E-92	3.2285	105
	x2	1192.97011	5.04E-30						
Semi-annual	x	14584.0399	4.478E-05	0.9825264	0.9649837	1630.64199	4.9705E-51	3.2512	60
	x2	1175.09995	5.364E-17						
Annual	x	12405.5254	0.0147277	0.9808768	0.9480018	795.033401	1.0323E-26	3.9252	33
	x2	1217.90408	1.069E-09						
500 Interval	x	17028.3059	1.162E-18	0.9758669	0.9712949	4528.92773	2.82E-181	3.3948	226
	x2	1086.4579	2.132E-50						
1000 Interval	x	16783.7054	9.458E-10	0.9760037	0.9666947	2237.0218	3.21E-89	3.4572	112
	x2	1091.04905	5.953E-26						
1500 Interval	x	15540.3816	3.951E-06	0.9754965	0.9614622	1453.08532	6.34E-59	3.4383	75
	x2	1133.85701	6.059E-18						
2000 Interval	x	16761.065	5.364E-06	0.9792228	0.9603195	1272.49894	1.60E-45	3.1186	56
	x2	1082.36084	1.02E-14						
2500 Interval	x	16971.5988	6.228E-05	0.9786436	0.9543256	962.310017	3.51E-35	3.1489	44
	x2	1099.58852	1.457E-11						
5000 Interval	x	19481.4593	0.0062787	0.9734599	0.9221329	366.788761	6.62E-16	4.7161	22
	x2	1030.08714	3.878E-05						
7500 Interval	x	17372.3322	0.0146498	0.9839261	0.8992533	367.276702	8.49E-11	2.9776	14
	x2	1084.69605	6.96E-05						

D70 EX2500 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	29771.9109	1.4218E-29	0.98727672	0.9811016	6362.8814	2.05E-155	3.1617	166
	x2	1743.53944	3.61143E-55						
Seasonal	x	37317.8767	2.72383E-13	0.98288248	0.9677016	1923.5593	3.102E-59	4.7033	69
	x2	1441.77132	4.38379E-16						
Semi-annual	x	37480.3282	5.38774E-08	0.9860136	0.9570426	1233.7152	1.845E-32	4.4925	37
	x2	1447.78578	7.41403E-10						
Annual	x	38982.9278	0.000147545	0.98081582	0.9271745	485.6996	2.183E-16	6.2499	21
	x2	1418.9391	3.86384E-05						
500 Interval	x	26584.7875	7.67925E-17	0.98918703	0.9802418	5168.7085	4.86E-111	4.0637	115
	x2	1861.84098	1.16751E-39						
1000 Interval	x	26633.4647	4.86186E-09	0.98883252	0.970776	2479.2799	1.28E-54	2.9718	58
	x2	1857.47371	4.06743E-20						
1500 Interval	x	27179.9107	3.04747E-06	0.98852457	0.960428	1550.5689	6.83E-35	3.2403	38
	x2	1835.53223	4.12917E-13						
2000 Interval	x	24585.8103	0.000227799	0.98840505	0.9494976	1108.1781	3.92E-25	2.1794	28
	x2	1938.11426	3.49636E-10						
2500 Interval	x	24781.9863	0.000966427	0.99024066	0.9370954	963.92612	4.96E-19	2.8121	21
	x2	1946.61888	3.22479E-08						
5000 Interval	x	28576.8529	0.020528342	0.98918539	0.8628336	365.8699	8.28E-08	3.2389	10
	x2	1820.63964	0.000921761						
7500 Interval	x	33098.4733	0.024069768	0.9934588	0.7921506	379.69293	2.746E-05	3.0287	7
	x2	1646.43232	0.0046337						

D72 – EX3600 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	19088.423	2.703E-08	0.992111	0.978112	4527.0689	1.35E-75	0.848	74
	x2	3018.6823	5.499E-28						
Seasonal	x	33911.661	1.217E-12	0.995774	0.968632	4358.8066	1.13E-43	0.516	39
	x2	2314.507	2.117E-15						
Semi-annual	x	32823.552	1.442E-07	0.996722	0.946558	3040.3423	1.54E-24	0.4724	22
	x2	2299.2851	1.189E-09						
Annual	x	30941.144	0.0071916	0.991379	0.899686	632.4541	2.97E-11	1.6103	13
	x2	2600.6214	0.0002443						
500 Interval	x	21567.644	2.674E-11	0.990605	0.976965	3901.3695	6.26E-75	0.6892	76
	x2	2871.9044	4.826E-28						
1000 Interval	x	21939.749	2.749E-06	0.990625	0.962587	1902.023	1.98E-36	0.7313	38
	x2	2818.5808	2.991E-14						
1500 Interval	x	22853.272	8.86E-05	0.991077	0.947211	1277.2779	1.76E-23	0.7004	25
	x2	2844.0149	9.083E-10						
2000 Interval	x	17815.008	0.0090758	0.989706	0.930277	817.22057	7.80E-17	0.8007	19
	x2	3052.5767	1.996E-07						
2500 Interval	x	21185.316	0.0083138	0.990005	0.912313	643.80427	6.20E-13	0.8895	15
	x2	2878.5932	6.694E-06						
5000 Interval	x	14946.686	0.1017651	0.996004	0.795205	623.12727	1.02E-05	0.4294	7
	x2	3446.05	0.0007116						
7500 Interval	x	17423.579	0.2572458	0.992879	0.657171	209.13293	0.004759	1.2202	5
	x2	2827.3186	0.0249763						

D75 – EX5500 Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	20106.7617	0.000735925	0.98093446	0.975304	4656.2852	1.02E-155	44.6242	183
	x2	2118.68321	1.46944E-40						
Seasonal	x	21894.4605	0.00776996	0.98284635	0.969637	2205.9208	4.93731E-68	37.4088	79
	x2	1952.68809	6.14871E-19						
Semi-annual	x	28712.0305	0.019153323	0.98076071	0.955901	1045.0278	2.97427E-35	45.3456	43
	x2	1835.93403	2.63549E-09						
Annual	x	20929.91	0.238239741	0.97878474	0.932366	507.49468	1.66752E-18	53.3175	24
	x2	1963.54861	1.14225E-05						
500 Interval	x	37891.6219	5.26756E-09	0.97229206	0.966796	3280.9839	8.95E-146	51.2402	189
	x2	1775.90436	8.00715E-28						
1000 Interval	x	38181.4835	4.03422E-05	0.97208864	0.960916	1602.0743	1.18E-71	52.288	94
	x2	1763.59548	2.88104E-14						
1500 Interval	x	37988.1861	0.000914911	0.97230559	0.955177	1053.2512	6.94E-47	51.2894	62
	x2	1778.53352	7.24202E-10						
2000 Interval	x	38427.2143	0.003289229	0.9739977	0.951198	842.80818	8.34E-36	52.1327	47
	x2	1772.88415	5.11943E-08						
2500 Interval	x	41046.6698	0.006334784	0.9731284	0.943789	633.74513	1.23E-27	52.678	37
	x2	1701.13174	4.11861E-06						
5000 Interval	x	39995.92	0.072520326	0.97466571	0.910582	307.77754	6.65E-13	55.0516	18
	x2	1757.92988	0.001514098						
7500 Interval	x	41830.563	0.142099389	0.97946566	0.866073	214.64505	1.12E-07	48.7642	11
	x2	1772.94888	0.014291949						

D78 – EX8000 Model

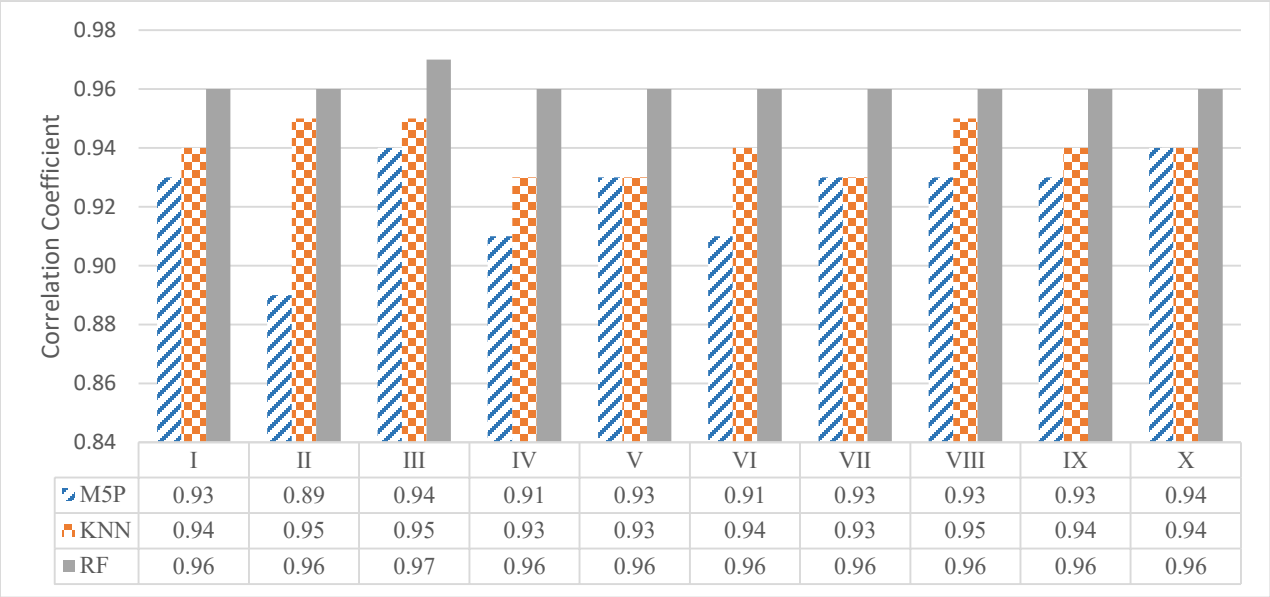
		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	255492.3406	1.28321E-29	0.9596862	0.951980211	1606.86622	2.28E-94	230.818	137
	x2	1049.945352	0.033102233						
Seasonal	x	242145.2998	1.04509E-09	0.9535635	0.933819449	544.17219	1.35878E-35	312.3332	55
	x2	1884.501248	0.044164635						
Semi-annual	x	229795.7121	4.77022E-06	0.9501187	0.917309859	304.761174	4.11931E-21	313.7491	34
	x2	2184.189809	0.070057077						
Annual	x	254401.2672	0.000270232	0.951565	0.893318606	176.816023	4.19566E-12	356.6971	20
	x2	1786.364719	0.257191115						
500 Interval	x	256968.0405	2.90052E-35	0.9556251	0.94944536	1819.72933	1.42E-114	232.046	171
	x2	979.6223407	0.035026513						
1000 Interval	x	257147.3833	4.4873E-18	0.9558556	0.943275501	898.595618	1.71E-56	233.191	85
	x2	978.0928412	0.143749289						
1500 Interval	x	257159.7106	1.65957E-12	0.9542209	0.935546239	583.632365	9.18E-38	240.611	58
	x2	961.1022201	0.241102816						
2000 Interval	x	254357.286	7.8753E-10	0.9568902	0.932054293	466.128665	6.36E-29	233.847	44
	x2	1071.268772	0.247338501						
2500 Interval	x	251732.9246	1.53346E-07	0.953782	0.921087684	330.185442	1.26E-21	250.736	34
	x2	1134.167968	0.297466061						
5000 Interval	x	245943.9421	0.000528516	0.9539156	0.884176683	155.245019	2.78E-10	262.575	17
	x2	1424.424933	0.400961772						
7500 Interval	x	249579.2302	0.006195236	0.9565691	0.840632292	99.1127767	2.26E-06	286.096	11
	x2	1130.403342	0.588988592						

D95 – 495HF Model

		Coefficient	P-Value	R Square	Adjusted R Square	F	F Significance	Residual MS(E+10)	Observations
All Points	x	144124.7226	1.36093E-06	0.9544543	0.94503501	1163.0575	1.019E-74	145.9632	113
	x2	9194.270778	4.8065E-10						
Seasonal	x	161472.4492	0.000905686	0.9533704	0.92845069	429.358138	3.183E-28	157.3241	44
	x2	8592.720654	0.000215477						
Semi-annual	x	185482.5809	0.00406476	0.9528113	0.90917848	242.298388	3.519E-16	160.1027	26
	x2	7319.371084	0.012464836						
Annual	x	186010.9753	0.04039358	0.9583685	0.87824301	149.631826	3.283E-09	168.2573	15
	x2	7233.473816	0.067377214						
500 Interval	x	173733.8695	1.98299E-07	0.947509	0.93708833	911.57037	6.302E-65	152.0969	103
	x2	7914.132386	8.95936E-07						
1000 Interval	x	168149.3694	0.000347334	0.948604	0.92757604	461.418817	1.636E-32	152.1235	52
	x2	8260.406057	0.00029692						
1500 Interval	x	180749.4351	0.0020667	0.9488691	0.91701663	306.201059	1.367E-21	161.1922	35
	x2	7522.880136	0.006327387						
2000 Interval	x	174768.5177	0.015203389	0.94428	0.89837915	194.889151	1.012E-14	165.3855	25
	x2	7875.154119	0.025508829						
2500 Interval	x	192275.5599	0.02118168	0.9440414	0.88537706	151.833261	1.437E-11	172.4257	20
	x2	6909.817212	0.082305797						
5000 Interval	x	181217.628	0.147710532	0.9488277	0.81743111	74.1672186	1.943E-05	210.3016	10
	x2	7497.467076	0.199482382						
7500 Interval	x	194222.2338	0.354367521	0.9323333	0.66541667	27.55665	0.0117292	300.4669	6
	x2	6664.780888	0.515435984						

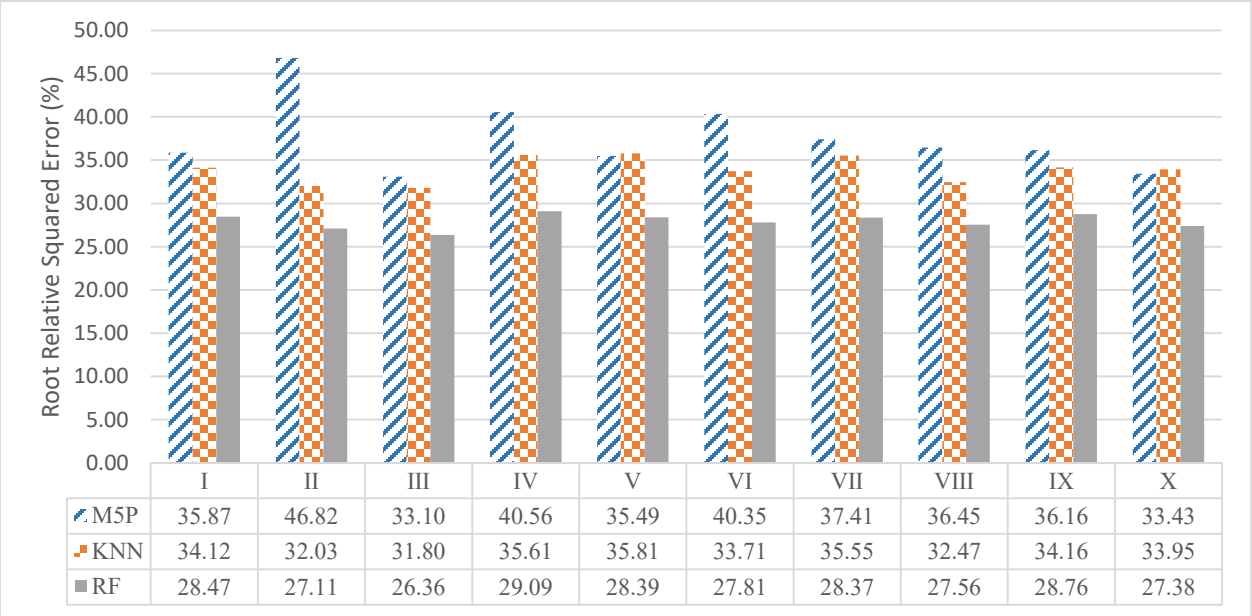
Appendix B: Performance of different models regarding residual value prediction

Correlation coefficient (CC)



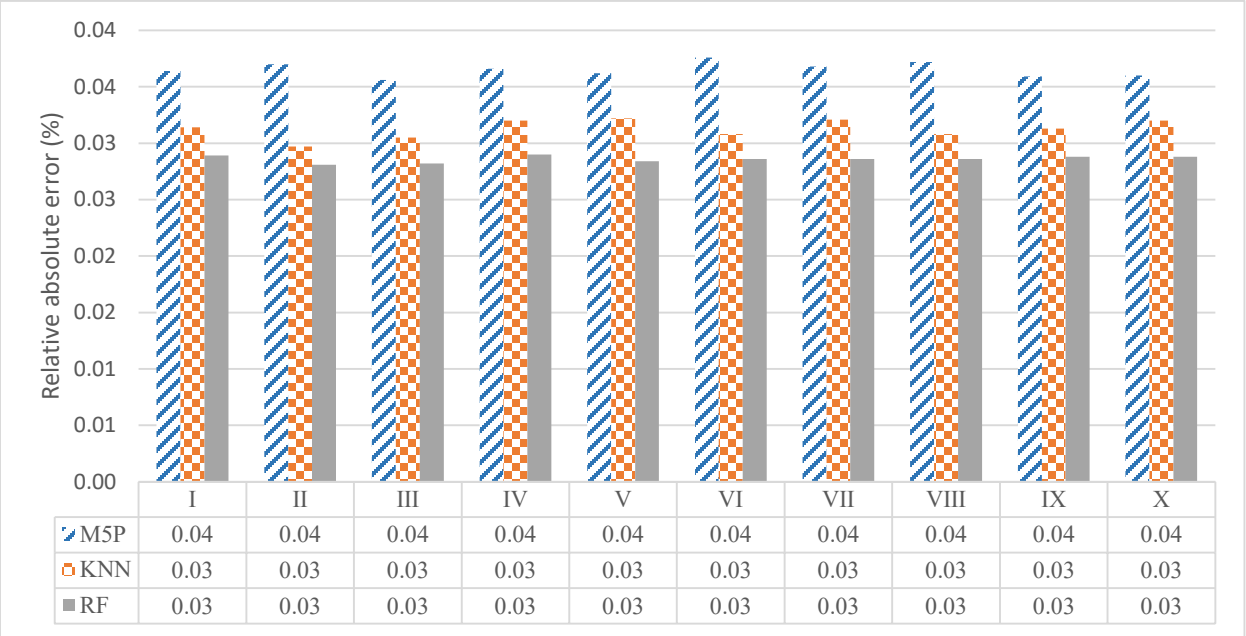
	I	II	III	IV	V	VI	VII	VIII	IX	X
M5P	0.93	0.89	0.94	0.91	0.93	0.91	0.93	0.93	0.93	0.94
KNN	0.94	0.95	0.95	0.93	0.93	0.94	0.93	0.95	0.94	0.94
RF	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.96	0.96

Root relative squared error (RRSE, %)



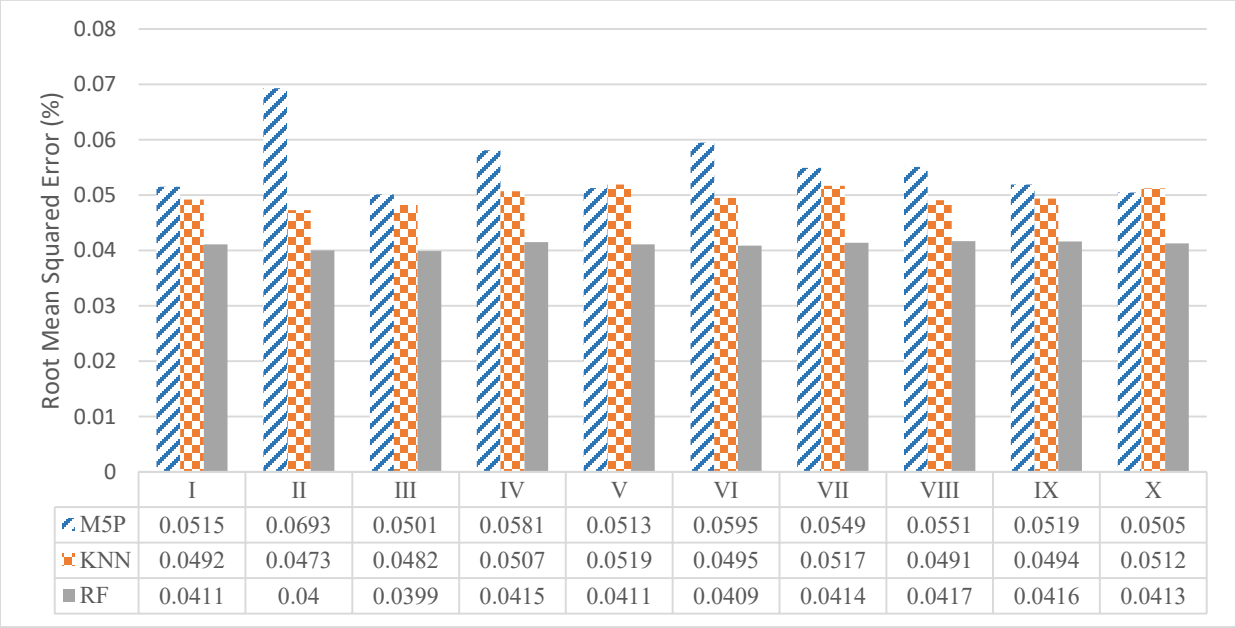
	I	II	III	IV	V	VI	VII	VIII	IX	X
M5P	35.87	46.82	33.10	40.56	35.49	40.35	37.41	36.45	36.16	33.43
KNN	34.12	32.03	31.80	35.61	35.81	33.71	35.55	32.47	34.16	33.95
RF	28.47	27.11	26.36	29.09	28.39	27.81	28.37	27.56	28.76	27.38

Relative absolute error (RAE, %)



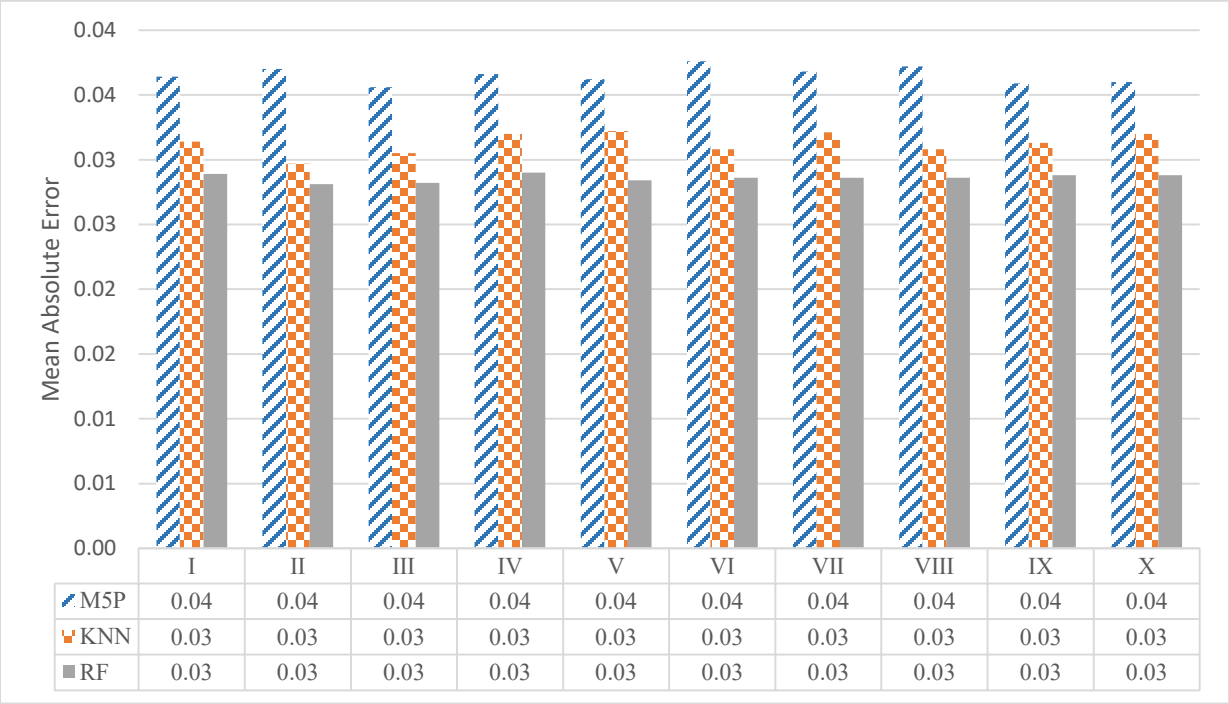
	I	II	III	IV	V	VI	VII	VIII	IX	X
M5P	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
KNN	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
RF	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Root Mean Squared Error (RMSE, %)



	I	II	III	IV	V	VI	VII	VIII	IX	X
M5P	0.0515	0.0693	0.0501	0.0581	0.0513	0.0595	0.0549	0.0551	0.0519	0.0505
KNN	0.0492	0.0473	0.0482	0.0507	0.0519	0.0495	0.0517	0.0491	0.0494	0.0512
RF	0.0411	0.04	0.0399	0.0415	0.0411	0.0409	0.0414	0.0417	0.0416	0.0413

Mean absolute error (MAE)



	I	II	III	IV	V	VI	VII	VIII	IX	X
M5P	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
KNN	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
RF	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Training time (millisecond)

	M5P	KNN	RF
I	1573.44	0	1251.6
II	1575	1.56	1226.6
III	1615.63	1.56	1248.4
IV	1565.63	0	1229.7
V	1557.81	0	1228.1
VI	1551.56	0	1246.9
VII	1581.25	1.56	1231.3
VIII	1626.56	0	1226.6
IX	1581.25	0	1231.3
X	1623.44	0	1251.6

Testing time (millisecond)

	M5P	KNN	RF
I	0	40.63	35.94
II	0	43.75	37.5
III	1.56	37.5	40.63
IV	0	35.94	37.5
V	1.56	42.19	34.38
VI	1.56	35.94	32.81
VII	0	34.38	34.38
VIII	1.56	40.63	39.06
IX	1.56	40.63	37.5
X	0	35.94	39.06

Running time