Assessing Residual Value of Heavy Construction Equipment Using Predictive Data Mining Model

Hongqin Fan[1]; Simaan AbouRizk[2]; Hyoungkwan Kim[3]; and Osmar Zaïane[4]

Abstract: Construction equipment constitutes a significant portion of investment in fixed assets by large contractors. To make the right decisions on equipment repair, rebuilding, disposal, or equipment fleet optimization to maximize the return of investment, the contractors need to predict the residual value of heavy construction equipment to an acceptable level of accuracy. Current practice of using rule-of-thumb or statistical regression methods cannot satisfactorily capture the dynamic relationship between the residual value of a piece of heavy equipment and its influencing factors, and such rules or models are difficult to integrate into a decision support system. This paper introduces a data mining based approach for estimating the residual value of heavy construction equipment using a predictive data mining model, and its potential benefits on the decision making of construction equipment management. Compared to the current practice of assessing equipment residual values, the proposed approach demonstrates advantages of ease of use, better interpretability, and adequate accuracy.

DOI: 10.1061/(ASCE)0887-3801(2008)22:3(181)

CE Database subject headings: Construction equipment; Data analysis; Information management; Decision making; Knowledge-based systems.

Introduction

The owning and operation of construction equipment constitutes a significant portion of yearly spending for large contractors engaging in equipment-intensive projects such as earth moving, highway, and industrial installations. According to Stewart (2006), the total construction equipment replacement value in North America of the top 250 construction and mining related companies reached nearly US$100 billion in 2006. To minimize the equipment cost per unit of service or maximize the stream of profits generated from the equipment investment, the contractors need to make the right decisions on equipment acquisition, repair/replacement/disposal, and reshuffle their equipment fleet on a regular basis in response to rapid changes in construction markets.

Among all the factors impacting such decisions, equipment residual value is cited as one of the most important, yet uncertain, with no consensus on the method of determination (Perry and Glyer 1990; Lucko and Vorster 2004). The residual value of construction equipment is the expected selling price in the market at a point of its service life. When the need arises for the determination of equipment residual value for equipment management decisions, such as equipment repair, rebuild, disposal, or replacement, the equipment under consideration has not yet been subject to the pricing process in the market (e.g. auction), therefore the market value of a piece of equipment can only be estimated based on experience, historical auction cases, or postulated formulas. As a wide variety of factors exert impact on the market value of construction equipment, including age, manufacturer, model, intensity of use, care, as well as market supply and demand, it is not surprising that no industrial criteria currently exist for an evaluation on the price of used construction equipment.

Research on estimating the depreciation and residual values of heavy equipment is conducted extensively in the agricultural and forestry industries, where similar or same equipment is generally used, and equipment cost constitutes a significant portion of the total production cost. Previous research efforts tended to use statistical regression approaches in order to establish functional relationships between the residual value of machinery and the known impact factors (McNeill 1979; Reid and Bradford 1983; Cross and Perry 1995; Unterschultz and Mumey 1996). In particular, the Cross and Perry (1995) method made its way to the American Society of Agricultural Engineers (ASAE) standard, "Agricultural Machinery Management Data" (ASAE 2003) for estimating the residual values of major types of agricultural equipment. In the construction industry, Lucko et al. (2006) investigated the effectiveness of applying a similar statistical approach to estimate the residual value of heavy construction equipment based on equipment auction data.

Several issues make it difficult or impossible to find a universal solution using the statistical regression method. First, different equipment categories display different behaviors in depreciation, as pointed out by Cross and Perry (1995) for agricultural machin-

[1]Graduate Research Assistant, Dept. of Civil and Environmental Engineering, Univ. of Alberta, Edmonton AB T6G 2W2, Canada. E-mail: hfan@ualberta.ca

[2]Chair, NSERFC/Alberta Construction Industry Research Chair, Professor, Dept. of Civil and Environmental Engineering, Univ. of Alberta, Edmonton AB T6G 2W2, Canada. E-mail: abouRizk@ualberta.ca

[3]Assistant Professor, Dept. of Civil and Environmental Engineering, Yonsei Univ., Seoul 120-749 (corresponding author). E-mail: hyoungkwan@yonsei.ac.kr

[4]Associate Professor, Dept. of Computing Science, Univ. of Alberta, Edmonton AB T6G 2E8, Canada. E-mail: zaiane@cs.ualberta.ca

Note. Discussion open until October 1, 2008. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on April 5, 2007; approved on June 26, 2007. This paper is part of the ***Journal of Computing in Civil Engineering***, Vol. 22, No. 3, May 1, 2008. ©ASCE, ISSN 0887-3801/2008/3-181–191/$25.00.

ery. The same observation holds true for construction equipment; large, heavy-duty, special-purpose equipment depreciates faster than small, multifunction equipment Second, some influencing factors on equipment residual value are dynamic and changing constantly, and their degree of impact may fluctuate over time (e.g. microeconomic indicators), and some other factors are very difficult to quantify as input for the regression model (e.g., technological progress and renovation). Third, the regression model is more appropriate for a specific class of equipment in a narrow range of model series, but this makes it difficult to collect sufficient samples to warrant statistical significance. Last, multicollinearity is a widely acknowledged problem for statistical regression that influences both the stability and accuracy of the derived statistical regression model. Some features of data samples (e.g., equipment age versus usage, and equipment age versus condition ratings) for equipment price evaluation have a relatively high coefficient of correlation.

This paper introduces an approach for predicting the residual value of construction equipment using the data mining technique. Data mining is an interdisciplinary science of statistics, machine learning, database, information theory, visualization, etc., with the objective of discovering valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al. 1996). As a hybrid of multiple disciplines, data mining integrates their perspective plausible features for knowledge extraction, validation, presentation, and deployment. The technique for which we advocate is a predictive data mining algorithm, the AutoRegression Tree (ART) proposed by Meek et al. (2002). This technique is utilized in this research to build a tree-structured nonlinear regression model based on large amounts of construction equipment auction data. The ART algorithm is chosen in this research for the building of predictive data mining models in consideration of its high interpretability and accuracy. A prediction model represented as a tree structure is more meaningful to decision makers in construction equipment management, because a tree-type model is more accurate in mimicking the nondeterministic nonlinear relationship between the input-output variables and is much simpler to interpret. The logical meaning of the tree-type model is inherently apparent, providing a convincing and transparent reasoning path to the prediction. Simply put, the analysis by means of a tree (or decision tree) is to settle on a set of "if-then" split conditions that allow for accurate predictions via careful data partitioning. After validation, the data mining model is embedded in a construction equipment information system for the prediction of equipment residual value. The integration of a data mining model with the equipment information system provides the user with the capability of browsing through the visualized data mining model, making prediction on the fleet equipment while being informed of the reasoning process. In addition, the automated processing of data mining makes it possible to update the model in real time to reflect recent changes in the equipment auction market. After comparing the data mining approach with the statistical regression approach, the paper concludes that the data mining model can better capture the complex and dynamic relationship between equipment residual value and its various influencing factors, and is easier to integrate with the current equipment management information system.

Literature Review

Depreciation is defined as the decrease in the residual value of equipment over time. Considering the fact that equipment cost constitutes a significant portion of the total production cost and the different options available for equipment acquisition, accurate evaluation of depreciation is crucial for a successful business in agriculture, mining, or construction. Therefore much research was devoted to quantifying the depreciation patterns of heavy equipment.

One of the most comprehensive studies on depreciation of agricultural equipment was the one conducted by Cross and Perry (1995). Their research objective was to identify the most fitted mathematical functions for modeling the relationship between equipment residual value (RV) and the known explanatory variables as stated in the following general form:

$$\mathrm{RV} = f(\mathrm{age}, \mathrm{usage}, \mathrm{care}, \mathrm{manufacturer}, \mathrm{auctionType}, \mathrm{region}, \mathrm{microeconomicsVariables}) \tag{1}$$

Given the fact that different types of equipment display different behaviors of depreciation, and that depreciation normally occurs in a nonlinear way, Cross and Perry (1995) proposed to transform the dependent variable RV and the two most statistically significant explanatory variables *age* and *annual hours of use* using the following Box–Cox transformation:

$$y^{\lambda} = \begin{cases} \dfrac{(y^{\lambda} - 1)}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases} \tag{2}$$

Depending on the transformation parameters λ obtained by maximum likelihood estimation or the Bayesian methods from the data, the Box–Cox transformation enables the regression model on RV to encompass a wide variety of functions, such as linear, exponential, and logarithmic. Using the Box–Cox transformation, Cross and Perry (1995) estimated price models for nine types of machinery and equipment used in agricultural production. Based on their research results, ASAE (2003) recommended a generalized regression formula for the estimation of residual value percentage (residual value divided by original list price) using *equipment age* and *annual hours of use* with different coefficients for different types of equipment.

Other similar research includes that conducted by McNeill (1979), Reid and Bradford (1983), Perry et al. (1990), and Unterschultz and Mumey (1996). To evaluate and compare different research results and functional forms, Dumler et al. (2000) and Wu and Perry (2004) each evaluated and compared different functional forms and debated over their applications in the agricultural industry.

Statistical regression was applied for building prediction models for the residual value of heavy construction equipment (Lucko and Vorster 2004; Lucko et al. 2006, 2007). Upon identification of

influential factors to the residual value of construction equipment, the researchers proposed several forms of multiple linear regression models (plain model, best model, and trade model) based on equipment auction data after statistical tests. The research confirmed that although equipment age is indeed the most significant influential factor of equipment residual value, other factors including the manufacturer, condition rating, auction region, and microeconomic indicators also contributed to the "goodness-of-fit" of the prediction model with statistical significance (Lucko et al. 2006). In their work, they used a narrow range of equipment, namely, the track dozers of 74.57–148.39 kW (100–199 horsepower) were selected as an implementation example to illustrate the methodology of applying regression analysis to predict the residual value of heavy construction equipment.

In the area of construction research, there have been many applications involving prediction using inferred models from data. Lee et al. (2004) used the GUIDE regression tree algorithm to quantify the cumulative impact of change orders on productivity; Arditi and Pulket (2005) applied a boosted decision tree for predicting the outcome of construction litigation; Kim et al. (2004) proposed a neural network-based classification system for automatic assessment of aggregate quality using laser imaging results.

Data Mining for Prediction of Equipment Residual Value

Prediction of a numerical value is a common data mining task to infer the most likely value of a response variable based on the known predictor variables, and can be represented in the following generalized form: $y=f(x_1,x_2,\ldots,x_n;r_1,r_2,\ldots,r_k)$, where y=target variable of the continuous data type, x_i $(i=1,2,\ldots,n)$=predictor variables of either categorical or continuous data types, and r_i $(i=1,2,\ldots,k)$=model parameters. Instead of a mathematical or statistical function as defined traditionally, $f(\ )$ stands for a data mining model representing the discovered patterns or rules from observation data by a data mining algorithm. The model parameters r_i are introduced in some algorithms to fine-tune the model structure or incorporate prior knowledge into the model generation process.

Predictive data mining uses complex computer algorithms to search through the data and generalizes the rules and patterns reflecting the relationship between the target variable and predictor variables. The "divide-and-conquer" and "heuristic" methods are two representative ones for inferring predictive models from data:

- The divide-and-conquer method: The algorithm searches over the data space and recursively partitions it into subspace, where more pure information or promising relations can be found. For example, the algorithms of a decision tree family use a measurement such as information gain or chi-square test to search for most informative splitting of data space by an input variable as well as a split-on value so that the partitioned data space contains more pure information on the prediction results. Using the prediction problem for equipment residual value as an example, after the algorithm identifies the input variable *equipment age* as the most relevant and informative feature to the residual value, it would have a tendency of splitting the data using *equipment age*. Exemplar algorithms using "divide-and-conquer" method are C4.5 by Quinlan (1993), Categorization and Regression Tree (CART) by Breiman et al. (1984) and ART (Meek et al. 2002).
- The heuristic method: Both artificial neural network (ANN) (Anderson 1995) and the support vector machine (Burges 1998) use a trial and error method to iteratively obtain an optimized predictive model based on predefined error measurement.

Though different algorithms use different methods for model inference, they have some common features which make them excel over traditional statistical regression approaches for predictive modeling on equipment residual value:

1. The models are inferred from data (recorded facts) with minimum user input. In contrast to the hypothesis-and-testing approach used in statistical regression, search and generalization is used in predictive data mining for the inference of patterns or rules in training data. The assumptions on statistical distributions or postulated functional forms make a statistical regression model subjective, and vary from one model to another; whereas, a data mining model is derived by an algorithm based on the available data, and involves minimum user interference in model generation.
2. Data mining models are represented by a computer model capable of storing complex rules and patterns by utilizing data structures, algorithms, and indexes. Therefore, complex rules and patterns that exist in data can be uncovered and represented.
3. Many data mining models, such as the family of decision trees or Bayesian inference, can be visualized in an intuitive manner for human interpretation.
4. Data mining models adapt to changes easily. As data mining can be designed as an automated process in a computer system, a data mining model can be updated in real time after the updating of the data sources.

AutoRegressive Tree Algorithm

The AutoRegressive Tree is a data mining algorithm proposed by Meek et al. (2002) to establish a nonlinear relationship between a set of explanatory variables and a target numeric variable through the exploration of the training data set. Much research (e.g., Chipman et al. (2002)) proved that, for a large number of domain problems, the data space can be partitioned into subregions where a simple linear regression model exists for each subregion. As the partitioning of the data space can be conveniently expressed in a decision tree structure with subsets of data residing in tree leaves where regression models are grown, this type of model is called treed regression (Alexander and Grimshaw 1996). Different approaches have been proposed to induce the decision tree structure with linear regression models at its leaf nodes, such as m5 by Quilan (1992), RETIS by Karalic (1992), and Bayesian treed models by Chipman et al. (2002). Though different algorithms generate similar treed models with a common goal of partitioning data space into subsets in such a way that the overall goodness-of-fit of the model to training data is maximized, different mechanisms and measurements are used to partition the data space and build local linear models. The ART algorithm uses the Bayesian technique to generate the tree structure and model parameters.

The Bayesian updating technique is a statistical inference method for model induction based on both prior assumption and observed facts. In contrast to the traditional statistical method assuming the model parameters are fixed, the Bayesian updating method considers the model parameters as changing variants, which can be described by the current statistical distribution. A

prior distribution of model parameters is assumed based on past experiences or subjective judgment. However, if the factual information is available, the prior distribution is updated by the likelihood that the observed factual data fall into the prior distribution. This updating process draws prior probability distribution closer to its true distribution and hence the posterior probability of model parameters is more accurately obtained.

The ART algorithm uses this posterior probability of model structure s to compare different alternatives of tree topology in terms of their goodness-of-fit to the training data set. Based on the Bayesian theory

$$p(s|d) = \frac{p(s)p(d|s)}{p(d)} \tag{3}$$

where $p(s|d)$=given training data d, the probability of fitting model structure s; $p(s)$=prior probability of model structure s; $p(d|s)$=marginal probability of observing data d, given model structure s; and $p(d)$=prior probability of data d.

Because the prior probability $p(d)$ of the training data set is a constant, Meek et al. (2002) defines $p(s)p(d|s)$ as the Bayesian score for the ART model. The first product $p(s)$ is the assumed structure prior, which is a subjective judgment on the probability distribution of model parameters whereas the second product $p(d|s)$ is the marginal likelihood of training data falling into the assumed prior distribution for given structure s.

For each candidate model structure in the data partitioning process, ART builds a normal multilinear regression model for the subset of data at each leaf node. Assuming the linear model parameters at the leaf nodes are independent from each other, the Bayesian model score is calculated as

$$\text{score}(s) = \prod_{i=1}^{L} \text{LeafScore}(l_i) \tag{4}$$

The LeafScore in Eq. (4) at each leaf node is calculated according to Eq. (3), using an assumed prior distribution $p(s)=0.1^{|\theta|}$ (θ=number of model parameters) and normal likelihood function. See Meek et al. (2002) for details on the Bayesian score calculation for each leaf node.

The ART algorithm uses the divide-and-conquer method to partition the data space and builds regression models at each leaf node. The pseudocode of the algorithm is shown in the following:

List 1. Pseudo-code of AutoRegressive Tree Algorithm

```
1.  #Start with the root node
2.  build a linear regression model at the root node
3.  calculate Bayesian model score
4.  #Compare alternative splitting options
5.  For each input attribute A
6.      #Determine candidate split values split[]
7.      # in case of categorical attribute
8.      If input attribute A is a categorical attribute
9.      set split[]= distinct nominal values of A
10.     else
11.     #in case of continuous attribute
12.     set split[]= 7 splitting points of 8 equal-probability areas
        assuming the input attribute conforms to a Normal distribution
13.     end if
14. # Loop through every split-on value to evaluate current splitting
    option
15. For each value in split[]
16.     partition the data using current attribute A and split-on value
17.     build linear regression model for each leaf node
18.     calculate Bayesian model score
19.     calculate increase in model score compared with model prior
        to splitting
20.     store current splitting parameters and model score
21. end For
22. End For
23. choose attribute and split-on value which leads to highest
    increase in model score
24. #continue with the recursive partitioning
25. split data using the selected attribute and split-on value
26. recursively repeat the above process for each subset of data
27. #Terminate splitting process
28. If   the splitting will not increase the model score or
         the number of cases in the leaf node is less than specified
         threshold value
29. terminate splitting
30. end If
```

To improve the model accuracy, the ART algorithm uses a dynamic splitting method proposed by Chickering et al. (2001) to determine candidate values for data splitting. Instead of determining these values at the beginning of the algorithm and using them for all the subsequent partitioning, the algorithm recalculates the candidate split-on values for features of the data subset at each step of partitioning:

- For a categorical attribute, it uses distinct nominal values in the subset of data (lines 8 and 9 in List 1), and
- For a continuous attribute, it uses seven intermediate points that split the attribute values into eight equal-probability areas assuming normal distribution (lines 10–12 in List 1).

ART Model for Prediction of Equipment Residual Value

Although it is theoretically possible to build a single predictive data mining model for all types of heavy construction equipment so long as they are fully represented by training data, the model of this scale would be of poor quality and difficult to interpret. Therefore, separate data mining models are built for each major category of heavy construction equipment. In this section, the data mining process is exemplified by selecting the equipment category of wheel loaders for model building and validation.

Data Sources

The primary data source for model building is Last Bid™, an online construction equipment database covering up-to-date auction results across the U.S. and international markets (Prism Business Media Inc. 2005, Equipment Watch Business, San Jose, CA). The wheel loader auctions across the United States and Canada from 1996 to 2005 are selected with the available information on make, model, year of build, auction year, conditions, auction locations, and transaction price. The "usage of equipment" information is missing from this data source because "it is difficult to confirm the data with confidence" (Vorster 2004). Other potential factors of influence on auction results, [i.e., gross domestic product (GDP) and yearly construction investment] are obtained from the U.S. Bureau of Economic Analysis and Statistics Canada.

Feature Selection

The residual value of a wheel loader is influenced by various features which have a potential impact on its market transaction price. To enable the data mining model to capture the inherent relationship, all the factors of potential influence to the residual value should be fully identified, and some features need to be transformed either to fit the model input or improve the model accuracy. Two examples of feature transformation for this model are *equipment age* and *auction location*. *Equipment age* measures the number of years the equipment has been in service at the time of auction, and has a direct impact on equipment residual value, therefore, it is derived and used together with *auction year* to describe the timeline of the auctioned equipment. For predictor variable *auction location*, the state/province is given in the auction data as a characterization attribute. To better represent the location variable, a simple transformation is conducted to derive the country of auction and the region of auction as two additional candidate attributes for this variable. A calculation of information gain based on the information theory (Shannon 1948) determines that the region of auction is the best attribute among the three (country, region, state/province) to represent the *auction location* because it has the maximum discriminating power on the response variable of *auction price* in the data set.

The usage of equipment (accumulated operation hours of a wheel loader) is considered an important factor on equipment residual value, but it is not available from the data source. Assuming normal use of equipment in its lifetime, the age and hours of use have a high coefficient of correlation [e.g., 0.75 in a research conducted by Perry et al. (1990) on farm tractors]. Therefore, it is safe to ignore this variable while including the age in years in training data to represent the usage of equipment.

To determine the condition rating of a piece of equipment with minimal bias, evaluation of equipment needs to follow the detailed guidelines set out by the equipment auctioneer, and is usually carried out by accredited equipment appraisers. The determination of condition rating for construction equipment is also explained by Lucko et al. (2006).

Finally, the following features are selected for building the predictive data mining model for equipment residual value:

- Make: Manufacturer of wheel loader.
- Model: Model of wheel loader.
- Horsepower: The rated engine horse power (HP).
- Age in years: Obtained based on the year of build and the auction year.
- Auction year: The year at which auction occurred.
- Auction location: The auction region (United States Southeast, Southwest, West, Mideast, Northeast, and Canada).
- Condition rating: The rating of equipment in terms of physical conditions (new, excellent, very good, good, and fair).
- Annual construction investment: Annual construction investment in the United States and Canada in US$ million at the year of auction.
- GDP: The Gross Domestic Product in the United States and Canada in US$ billion at the year of auction.

To measure the equipment price in constant dollars, the response variable *auction price* is indexed to the year 2000 based on the consumer price index obtained from the U.S. Bureau of Labor and Statistics as well as Statistics Canada.

Data Quality Control

Data quality in data mining measures the overall fit of data to knowledge generation. In addition to the general requirements, such as consistent format, no missing values and outliers, for data quality in decision analysis, the data should be representative of all the features in full range and unbiased quantity. If there is a systematic lack of attribute values (e.g., lack of equipment auction cases in a price range, or lack of an auction region in the training data) the predictive model for equipment residual value would have a poor accuracy of prediction for the defined domain problem.

The attribute values of "unknown" for *equipment condition* are considered as missing values, and replaced by the mode value of "good." Some auction cases of transaction prices over US$200,000 are removed from the collected data as isolated cases resulting from customized build or special attachments. Finally, a total of 8,589 effective cases are obtained for model generation.

A preliminary check is conducted as to the representation of predictive features and auction results by the training data. Fig. 1 shows the histogram of the discretized auction price, the auction prices from 6,000 to 200,000 are binned into 14 cohorts based on Sturge's rule [number of bins $=1+3.3\log(N)$, where $N=$number of data points], with no obvious missing data in any price cohorts. The representation data on each predictor variable is checked in the same principle based on its frequency diagram (for categorical variables) or histogram (for continuous variables).

Model Generation and Validation

The ART algorithm includes two parameters for model structure control: One is the coefficient of complexity γ controlling the growth of the tree; a higher value increases the likelihood of node splitting to generate a bushy tree, and the other is the minimum number of cases M in each leaf node. The sensitivity analysis of γ and M on prediction accuracy, using 90% of the data for training and the remaining 10% for validation, found out that prediction accuracy is not sensitive to γ but is sensitive to M. The model parameters γ and M are determined as 0.5 and 20, respectively, for the final model generation.

To verify the stability and accuracy of the predictive model for equipment residual value, a ten-fold cross-validation method is used for model generation and validation. The wheel loader training data comprised of 8,589 effective cases is randomly divided into 10 partitions of approximately equal size, each containing around 859 cases. A data mining model is generated and validated for ten iterations according to the following procedure: hold each partition and use the remaining of the entire data set as the training data to generate the ART model, and then use the reserved partition as out-of-sample data for the validation test.

Test results indicate that the ten models are similar in their tree topography and linear regression functions at leaf nodes. Fig. 2 partially shows the structure of the derived tree model at the first few levels. A comparison of the ten models found that the top levels of the tree structure starting from the root node are the same, whereas some nodes at the bottom levels vary slightly.

In addition to the regression tree, the algorithm also generates an output which ranks the predictor variables as per their discriminating power on equipment auction price. The same ranking is generated from all the ten iterations as in the following in a decreasing order: *equipment age, horsepower, make, auction year, GDP, Construction Investment, model*, and *condition*. The fact that *equipment age, horsepower*, and *make* are the top three relevant features for prediction can be observed from the tree structure: the three features are most frequently selected at the top levels to partition the data space (Fig. 2). Another finding of interest is that the algorithm ranks *equipment model* as one of the

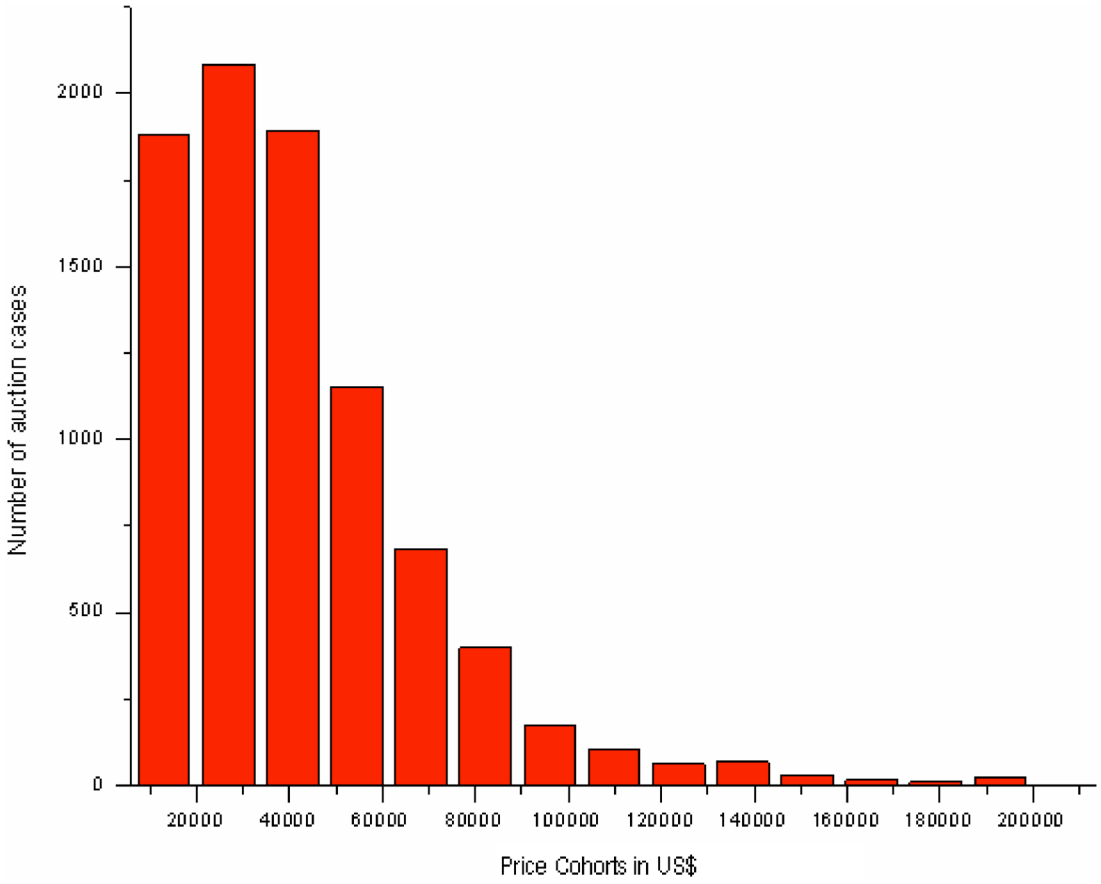

Fig. 1. Histogram of wheel loader auction price

least predictive features on equipment residual value even though equipment model is typically one of the decisive factors on its residual value. To explain this result, only the top 50 equipment models with a large number of auction cases accounting for the over 70% of the total cases are selected for the model building, and it turns out that *equipment model* is ranked as one of the powerful features for prediction. In consideration that there are over 300 models of wheel loaders in the training data, the information conveyed by this feature on equipment residual value is noisy, therefore the algorithm cannot identify *equipment model* as a persuasive explanatory variable on equipment residual value.

To cross validate the prediction accuracy of the data mining model, three measures are used to evaluate prediction errors in each iteration:

(1) Relative squared error (RSE): Evaluates the percentage of the total squared error between the predicted value and actual value out of the total squared error if using the average of actual values as a prediction. That is, the total squared error is normalized by dividing it by the total squared error of a simple default predictor using the average of the actual values for prediction

$$\text{RSE} = \frac{\sum_{i=1}^{n} (p_i - a_i)^2}{\sum_{i=1}^{n} (a_i - \bar{a})^2} \tag{5}$$

where p_i is the predicted value; a_i is the actual value; and $\bar{a}$ is the average of actual values in the validation sample.

(2) Root relative squared error (RRSE): Takes the root of RSE so that the error rate is reduced to the same magnitude as the response variable (i.e., the same dimensions as the quantity being predicted).

(3) Mean absolute error (MAE): The average of absolute prediction errors (i.e., without taking into account the sign of the error). Contrary to the mean squared error, which tends to exaggerate the effect of outliers, the MAE is not affected by extreme values and all sizes of errors are treated evenly.

The ten-fold cross validation generated similar test results on RSE, RRSE, and MAE as summarized in Table 1. On average, after using the predictive model, the total squared error is reduced by 94.5% compared to using the average value as the prediction. This error reduction in the same magnitude is 76.7% and the mean absolute error of prediction results is US$4,248. About 1% of the cases are indicated as "unpredictable" by the model; this is because when a test case falls into a leaf node where no regression is available due to insufficient information for the creation of a regression model, or if the number of cases is less than 20.

To evaluate the dispersions of prediction errors, the prediction errors in US$ are transformed into error percentage rates after being divided by their perspective actual values, and presented in box plots for each iteration. As shown in Fig. 3, the ten boxplots show similar characteristic values (i.e., quartile values at 25, 50, and 75% denoted by the top, middle, and bottom lines of the box). The prediction error rates for the middle 50% cases (the cases with prediction errors between upper and lower quartiles) are less than 8.5%, whereas a high degree of dispersion is observed from all the ten sets of test results. Each set of test results has a small

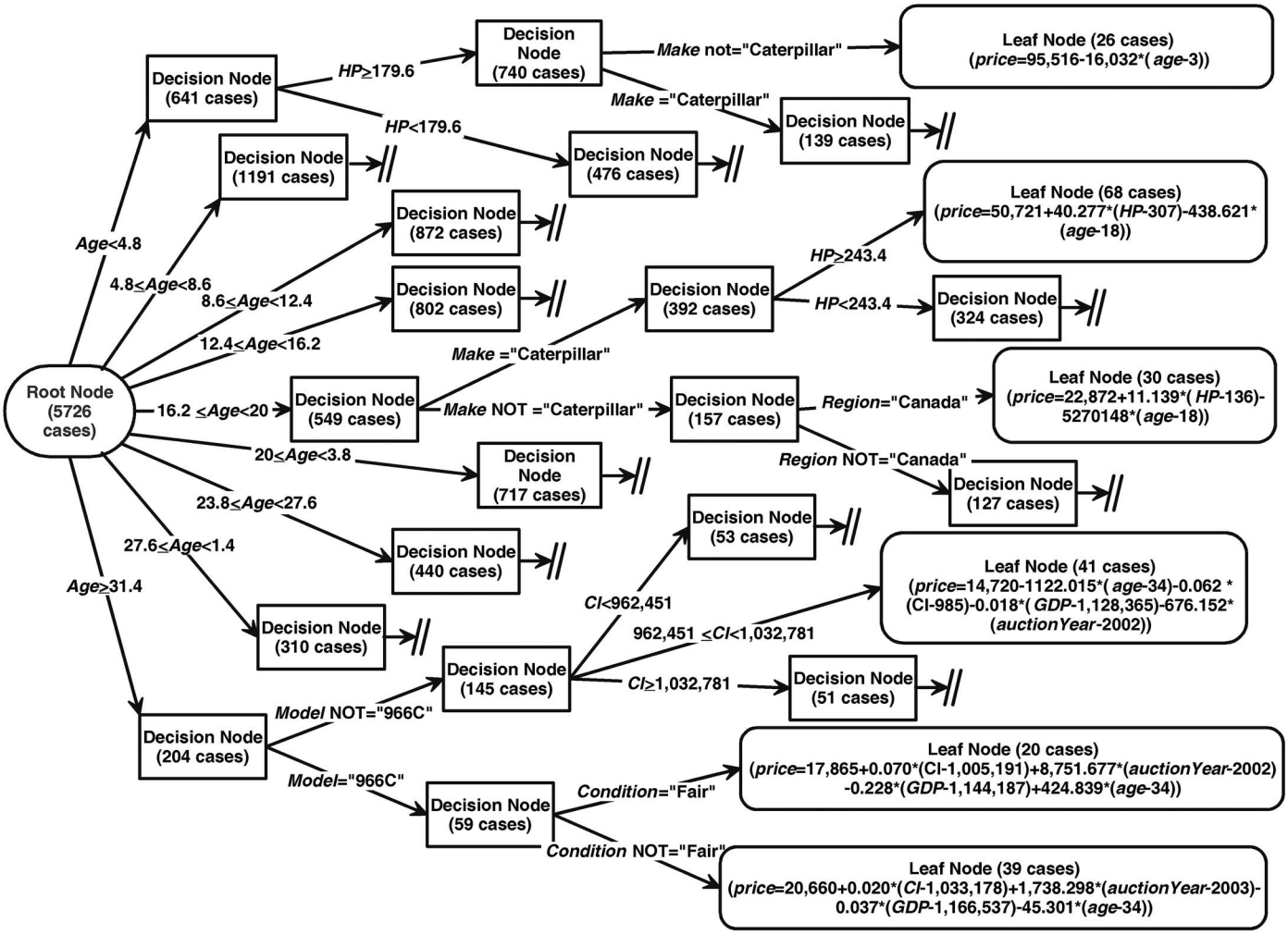

Fig. 2. Partial ART data mining model for prediction of equipment residual value. (Obtained using Microsoft SQL Server 2005 Analysis Services; for lack of space, most of the decision paths are not shown here.)

number of outliers with prediction errors outside 1.5 times interquartile range from the median. The prediction errors of each set follow a bell-shaped distribution, yet with open ends on both sides.

Both the cross-validation tests and boxplots verify the stability of the inferred data mining model. The mean absolute error of US$4,248, representing the deviation of predicted market prices of the equipment from their transaction prices, is less than 10% of the average equipment transaction price and deemed acceptable for heavy construction equipment.

Table 1. Test Results of Ten-Fold Cross Validation

Partition	Missing	Relative squared error (RSE %)	Root relative squared error (RRSE %)	Mean absolute error (MAE)
1	11/859	5.3	23.0	4,289
2	8/859	5.1	22.6	4,105
3	18/859	7.7	27.8	4,594
4	17/859	3.9	19.8	3,972
5	2/859	6.6	25.7	4,064
6	5/859	7.7	27.8	4,743
7	0/859	5.7	23.9	4,391
8	10/859	3.9	19.8	4,211
9	11/859	4.1	20.3	4,032
10	2/859	5.4	23.2	4,075
Average	8.4/859	5.5	23.4	4,248

The dispersion of error distribution shows that a small percentage of cases have a high degree of deviation, which is attributed to the fact that these isolated cases are not well represented in the model; on the other hand, the *equipment age* being the most important predictor introduces a certain degree of error when being measured in the whole number of years. The predictor accuracy of the current model can be improved if equipment usage data, rather than *equipment age*, is available to gauge the intensity of equipment usage. To improve the prediction accuracy, the final model is built using the entire data set.

Comparison of ART with ANN and Multivariate Linear Regression Models

The same prediction problem is also modeled using the ANN model and the multivariate linear regression (MLR), and validated with 10% holdout test. The prediction accuracy is evaluated on the ANN and MLR models using the same three measures for the ART model, as summarized in Table 2. It shows that both the

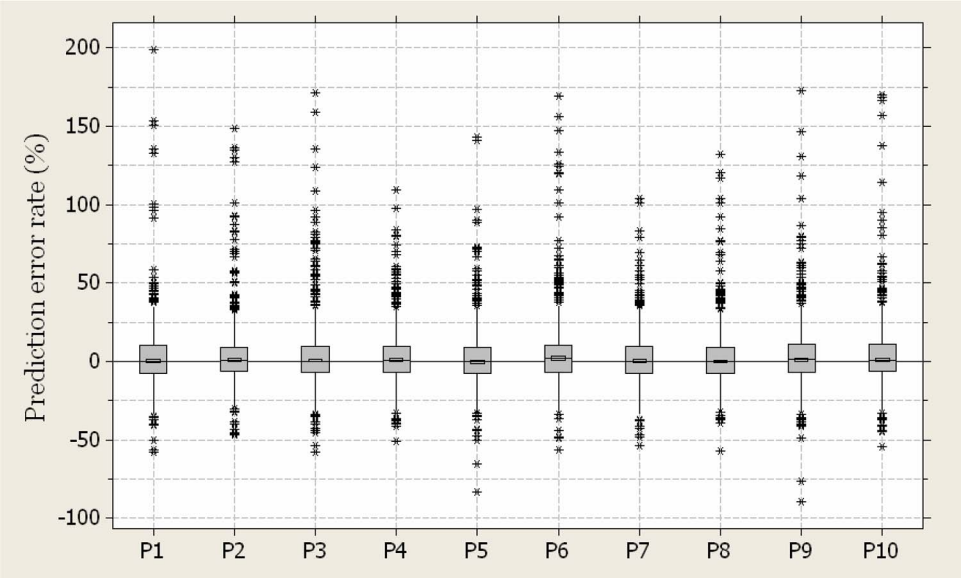

Fig. 3. Boxplots for prediction error rates of cross-validation tests

ANN and MLR models perform worse than the ART model in explaining the variability of auction price and also generate a higher degree of prediction error.

The higher prediction error of the ANN model can be explained by the way it handles the categorical input attributes: Each categorical input attribute is encoded into $N-1$ (N is the number of nominal values for the attribute) binary attribute as model input. A large number of input attributes are created for a predictive model if there are many categorical features as input, such as is the case in this problem. As a result, the large number of input variables quickly decreases the prediction accuracy of the ANN model. The high prediction error of the MLR model is attributed to its over-simplified assumption of the linear relationship between the auction price and its predictor variables using a single statistical regression model.

Deployment of Predictive Data Mining Models for Equipment Residual Value

The data mining models for the prediction of equipment residual value were built for major categories of equipment and deployed in a construction equipment management information system for testing. Fig. 4 shows a screenshot from the system with the integrated data mining module. After choosing a piece of equipment from the fleet database, the user gets the projected market price under given conditions. The sensitivity of the auction price on the sale date and location can be conducted on a piece of equipment by changing the input parameters. The user can also browse through the regression tree model by visually expanding the tree structure, and moving the mouse cursor over a leaf node to view the multivariate linear regression model. The related historical cases in the leaf node used for prediction on the current case are automatically retrieved for comparisons. The visualization of the regression tree model allows the user to reaffirm the predicted transaction price or analyze a doubted prediction.

Due to copyright issues, the online equipment auction data source does not allow for a live connection to external applications. Otherwise, the data mining for prediction of equipment residual value can be designed as a fully automated process. Under a fully automated data mining process, all the procedures including feature extraction and transformation, data cleaning, modeling, validation, and deployment can be implemented programmatically in a streamlined process. One recommendation on such a data mining enabled system is to design it as a two-mode application: user mode and developer mode. The user mode interacts directly with the users by exposing the data mining models for browsing, analysis and prediction; whereas, the developer mode implements the data mining process with the visual capability of updating the training data, fine-tuning the model parameters, and validating the model.

Table 2. Comparison of ART with ANN and MLE in Prediction Errors

Methods of Prediction	RSE (%)	RRSE (%)	MAE ($)
ART	5.5	23.3	4,248
ANN	22.3	47.2	7,274
MLE	40.4	63.6	11,069

Discussions

Though data mining serves as a unique modeling approach for predictive analysis in construction equipment management, there are several issues which have to be handled carefully to ensure that the derived model truly reflects the inherent relationships or rules in reality:

- Problem definition: The inference capability and storage mechanism of data mining greatly loosens the problem scope definition compared with the traditional statistical methods; however, a data mining problem needs to be defined at an appropriate level to control the model complexity, and model interpretability. This research divides the equipment into major categories for which separate models are built up for prediction. The categorization of equipment can be interpreted as building the first level of a decision tree for all the equipment in a manual process, followed by the automatic tree growing for the remaining part.

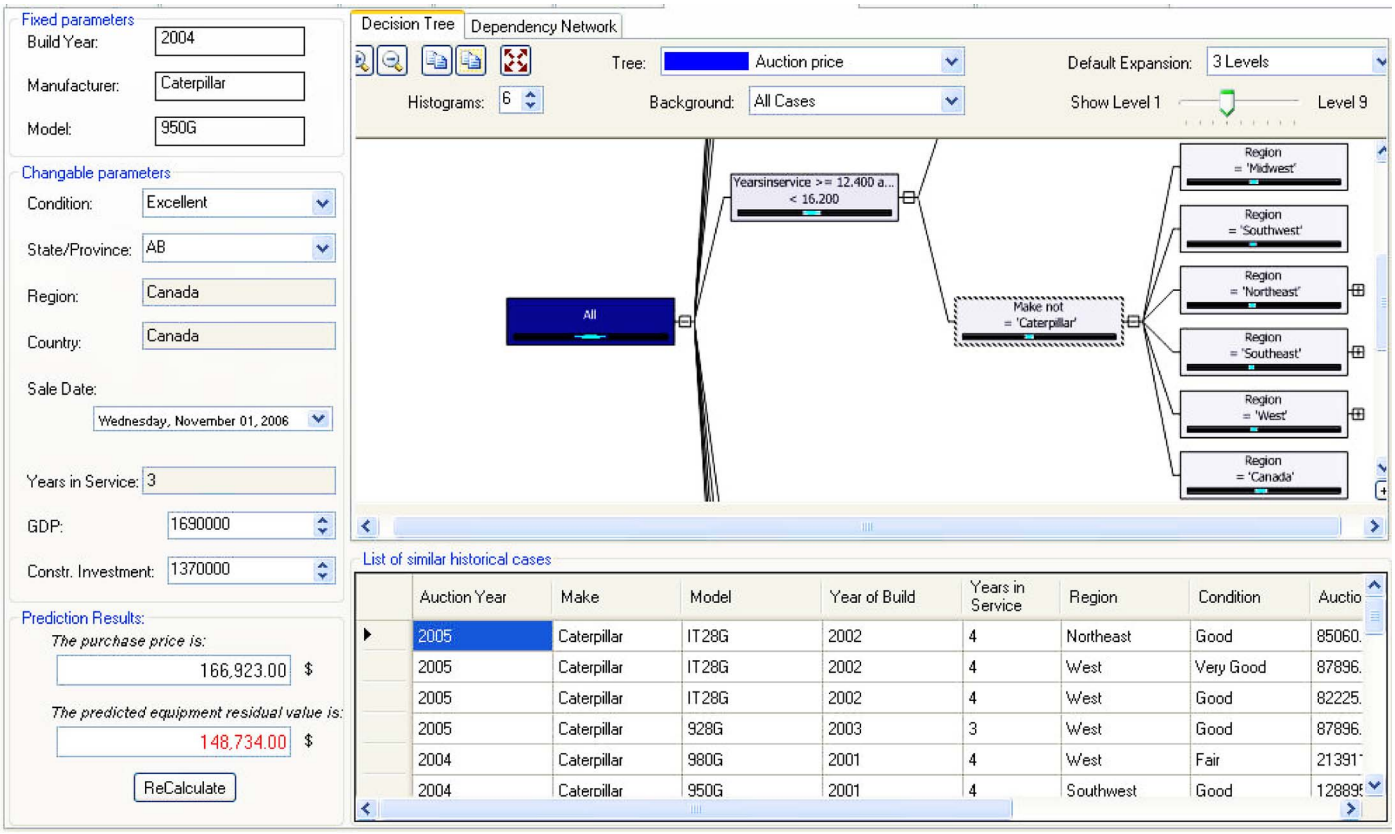

Fig. 4. Screenshot of the equipment management information system (note: The diamond location denotes the mean of the auction price, and its width denotes the variance of auction price on the node)

- Data source selection: The data mining algorithms infer the model structure out of data, hence, the data source used for mining should contain unbiased, fully represented information on the data mining problem. The data sources should be sufficient and reliable, with a fair level of data quality.
- Data preparation: As demonstrated in this research, the data used for the model inference should be appropriately represented by a set of predictive features and also quality insured. Common methods for preparation are feature transformation and preselection, data validation, evaluation of data representation on both input features and output results, etc.
- Model inference: Select a data mining algorithm which fits into the data mining problem, is easy to use and simple to interpret. In this application, neither neural network nor CART algorithms were selected because the former generates a model that is difficult to interpret, whereas the latter only predicts a ranged value for the response variable.
- Model validation: The data mining model must be validated in order to be effective. Out-of-sample cases should be used for the validation of the model inferred from in-sample data. The multifold cross validation is used in this research to verify that the generated data mining model is stable, representative, and accurate.

Finally, incorporating the prior domain knowledge into the data mining process is both necessary and important, in spite of the fact that data mining is an automated or semiautomated process of discovering knowledge from data. The feature selection and model validation in this research demonstrates the importance of domain knowledge for data mining. It would be difficult to obtain an unbiased accurate predictive model if model features are not represented by the data or the generated model is not validated based on domain knowledge.

Summary and Conclusions

This paper presents a data mining approach for the prediction of construction equipment residual value and its deployment in a construction equipment information system. In summary, the data mining-based solution provides some benefits which cannot possibly be achieved with the current rule-of-thumb or statistical methods:

- The data mining model is capable of capturing the relationships, patterns, and rules that exist in a dynamic and complex environment. The prediction of equipment residual value involves a large number of influential factors which are subject to changes over time. Statistical regression tackles this problem by inferring the statistical relationship between the residual value and a few meticulously selected predictors and cannot easily adapt to changes.
- The data mining model is primarily data driven and less dependent on personal experiences. The data mining algorithm searches over the data space to infer a model structure that reflects the relationship between the equipment residual value and its influential factors. Prior knowledge can be incorporated into a data mining process, but the model inference is based more on recorded facts, less on individual experience.
- Many data mining models are transparent and interpretable. Many data mining models, such as the decision tree, and the

Bayesian inference, can be visualized for human judgment and analysis, with the reasoning method and process also explained. For a few unpredictable cases, because historical related cases are missing from the training data, or the cases with a high level of deviation, the user is always informed or has an opportunity of making further investigation. Therefore the "white box" data mining models help to make informed decisions in the prediction of equipment residual value.
- The data mining model can be deployed in the equipment management information system with an automated process of modeling and updating.

Using the predictive modeling for the wheel loader residual value as an example, this paper explains how a typical data mining algorithm, the AutoRegressive Tree, infers knowledge from data. The entire process of data mining from data preparation, model generation, and model validation is illustrated by using the ART model of wheel loaders. Multifold cross validation and boxplots are used for testing the stability and accuracy of the generated data mining model. The paper also demonstrates the advantages of data mining enabled applications in construction equipment management by deploying the predictive models of equipment residual value in an equipment management information system.

In summary, predictive data mining provides a more accurate, flexible, and interpretable approach for assessing the residual value of heavy construction equipment. Using the embedded predictive modules for equipment residual value, the sellers can determine the best time to sell their machines, the buyers can determine the best time to purchase their required machines, and the equipment owners can perform life cycle analysis on equipment to make decisions on equipment repair, overhaul, disposal and replacement. Other data mining applications in construction equipment management, including outlier mining for problem identification (Fan et al. 2007), and time series forecasting for budget planning are also underway in this research. The data mining captures the complexity and dynamics of construction equipment management by making inferences out of data, which could not be realized by using conjectured mathematical or statistical models. The application of data mining in construction equipment management makes it possible for the management team to gain insight into the large amounts of data collected in construction equipment operations and management, and to make proactive decisions.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (Grant No. CRD 226956-99) and Yonsei Research Grant. The writers wish to express their sincere appreciation to Dale Tillapaugh, for his participation and contribution to this research.

References

Alexander, W. P., and Grimshaw, S. D. (1996). "Treed regression." *J. Comput. Graph. Stat.*, 5(2), 156–175.

American Society of Agricultural Engineers (ASAE). (2003). "Agricultural machinery management data." *ASAE D497.4*, American Society of Agriculture and Biological Engineers, St. Joseph, MI.

Anderson, J. A. (1995). *An introduction to neural networks*, MIT Press, Cambridge, Mass.

Arditi, D., and Pulket, T. (2005). "Predicting the outcome of construction litigation using boosted decision trees." *J. Comput. Civ. Eng.*, 19(4), 387–393.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*, Wadsworth, Inc., Belmont, Calif.

Burges, C. J. C. (1998). "A tutorial on support vector machines for pattern recognition." *Data Min. Knowl. Discov.*, 2(2), 121–167.

Chickering, D. M., Meek, C., and Rounthwaite, R. (2001). "Efficient determination of dynamic split points in a decision tree." *Proc., IEEE Int. Conf. on Data Mining 2001*, San Jose, Calif., 91–98.

Chipman, H. A., George, E. I., and Mcculloch, R. E. (2002). "Bayesian treed models." *Mach. Learn.*, 48(1–3), 299–320.

Cross, T. L., and Perry, G. M. (1995). "Depreciation patterns for agricultural machinery." *Am. J. Agric. Econom.*, 77(1), 194–204.

Dumler, T. J., Burton, R. O., and Kastens, T. L. (2000). "Management implications of farm tractor depreciation methods." *J. ASFMRA Am. Soc. Farm Managers and Rural*, 63(1), 3–10.

Fan, H., Kim, H., AbouRizk, S., and Han, S. H. (2007). "Decision support in construction equipment management using a nonparametric outlier mining algorithm." *Expert systems with applications*, Elsevier, New York.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From data mining to knowledge discovery: An overview." *Advances in knowledge discovery and data mining*, AAAI Press, Menlo Park, Calif., 1–30.

Karalic, A. (1992). "Linear regression in regression tree leaves." *Proc., ISSEK '92 (International School for Synthesis of Expert Knowledge)*, Bled, Slovenia.

Kim, H., Rauch, A. F., and Haas, C. T. (2004). "Automated quality assessment of stone aggregates based on laser imaging and a neural network." *J. Comput. Civ. Eng.*, 18(1), 58–64.

Lee, M., Hanna, A. S., and Loh, W. (2004). "Decision tree approach to classify and quantify cumulative impact of change orders on productivity." *J. Comput. Civ. Eng.*, 18(2), 132–144.

Lucko, G., Anderson-Cook, C. M., and Vorster, M. C. (2006). "Statistical considerations for predicting residual value of heavy equipment." *J. Constr. Eng. Manage.*, 132(7), 723–732.

Lucko, G., and Vorster, M. C. (2004). "Predicting the residual value of heavy construction equipment." *Proc., Towards a Vision for Information Technology in Civil Engineering, 4th Joint Int. Symp. on Information Technology in Civil Engineering*, ASCE, Reston, Va.

Lucko, G., Vorster, M. C., and Anderson-Cook, C. M. (2007). "Unknown element of owning costs—Impact of residual value." *J. Constr. Eng. Manage.*, 133(1), 3–9.

McNeill, R. C. (1979). "Depreciation of farm tractors in British Columbia." *Can. J. Agric. Econ.*, 27(2), 53–58.

Meek, C., Chickering, D. M., and Heckerman, D. (2002). "Autoregressive tree models for time-series analysis." *Proc., 2nd SIAM Int. Conf. on Data Mining*, Arlington, Va.

Perry, G. M., Bayaner, A., and Nixon, C. J. (1990). "The effect of usage and size on tractor depreciation." *Am. J. Agric. Econom.*, 72(2), 317–325.

Perry, G. M., and Glyer, J. D. (1990). "Durable asset depreciation: A reconciliation between hypotheses." *Rev. Econ. Stat.*, 72(3), 524–529.

Prism Business Media Inc. (2005). Last Bid auction results [online], ⟨https://www.equipmentwatch.com/Marketing/LB_overview.jsp⟩ (Mar. 11, 2007).

Quinlan, R. J. R. (1992). "Learning with continuous classes." *Proc., AI'92*, Adams and Sterling, World Scientific, Singapore, 343–348.

Quinlan, R. J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, Mateo, Calif.

Reid, D. W., and Bradford, G. L. (1983). "On optimal replacement of

farm tractors." *Am. J. Agric. Econom.*, 65(5), 326–31.
Shannon, C. E. (1948). "A mathematical theory of communication." *Bell Syst. Tech. J.*, 27, 379–423, 623–656.
Stewart, L. (2006). "Giants 2006." *Constr. Equip.*, Boston, 109(9), 47.
Unterschultz, J., and Mumey, G. (1996). "Reducing investment risk in tractors and combines with improved terminal asset value forecasts." *Can. J. Agric. Econ.*, 44(1996), 295–309.
Vorster, M. C. (2004). "How to estimate market value." *Constr. Equip.*, 107(6), 64–65.
Wu, J., and Perry, G. M. (2004). "Estimating farm equipment depreciation: Which functional form is best?" *Am. J. Agric. Econom.*, 86(2), 483–491.