Building Tomorrow's Society
Bâtir la Société de Demain

Fredericton, Canada
June 13 – June 16, 2018/ *Juin 13 – Juin 16, 2018*

MARKOV CHAIN-BASED LONG-TERM PREDICTION OF EQUIPMENT USAGE FROM HISTORICAL DATA

Chang Liu[1], Mostafa Ali[1], and Simaan AbouRizk[1,2]
[1]University of Alberta, Canada
[2]abourizk@civil.ualberta.ca

Abstract: The construction industry relies heavily on the use of equipment with fleet management playing a critical role in optimal project delivery, particularly for general contractors. Reliable prediction of equipment usage can enhance acquisition or disposal decisions and, in turn, project performance. This study proposes a methodology for the long-term prediction of equipment usage, in consideration of potential variation in market conditions, using historical data from various sources. Regression models capable of predicting usage of individual equipment categories are developed. For longer-term predictions, a Markov chain, used to simulate market fluctuations, is combined with the regression models. The proposed method can inform equipment managers of future fleet requirements for improved project planning and delivery.

1 INTRODUCTION

Management of equipment, as a major resource, is vital to the success of every construction project. In Canada, equipment costs can account for up to 60% of the total cost of many construction projects (University of Toronto 2001). Accordingly, equipment management decisions, including the acquisition and disposal of equipment, have a considerable impact on daily construction operations. To properly manage equipment fleets, decision-makers must be able to reliably predict future equipment demands and equipment costs far enough in advance to allow for the implementation of decisions in practice. Currently available long-term, empirical prediction methods, however, are not analytical, often resulting in unreliable estimates and, in turn, in poor fleet management decisions. This study proposes an analytical prediction methodology to estimate long-term equipment usage, using historical data currently available in most construction companies, for improved fleet management.

2 BACKGROUND

2.1 Equipment Management in Construction

In general, acquisition or disposal decisions are made based on economic principles and the consideration of a variety of factors, including equipment costs, maintenance costs, residual values, and market values. Cumulative cost models, which provide numerical and graphical analysis of maintenance costs, have been proposed and gradually improved upon with more sophisticated cost models, life-to-date repair costs, and period-cost-based models (Vorster 1980, Mitchell et al. 2010, Bayzid 2014). In conditions where sufficient historical data were available, various forms of regression models have also been developed to estimate maintenance cost time series and to forecast maintenance cost intervals instead of point values (Yip et al. 2014, Duncan 2015, Bayzid et al. 2016). With equipment transactions becoming more common, understanding residual and market value of equipment is becoming increasingly important for acquisition

or disposal decisions. Auction records retrieved from online construction equipment databases have been introduced into studies and have become major sources for equipment residual value analysis (Lucko 2003). Spatial cost analysis has been further developed using residual value regression models (Ponnaluru et al. 2012). Additionally, advanced data mining methods for predicting equipment costs have been proposed (Fan et al. 2008). Changing economic conditions have also been considered in quantitative research to tackle incongruous economic data (Lucko 2011).

Understanding equipment costs, however, is not sufficient for optimal management of equipment. Decision-makers must also investigate and consider future equipment demand. Indeed, poor equipment management may result, for example, in a once rapidly growing company that now, as a consequence of an economic down-turn, is required to sell equipment at lower-than-purchase prices. While short-term (monthly/quarterly) usage prediction is achievable (and more accurate), it may not provide practitioners with enough time to make adjustments to existing fleets due to the lengthy approval and implementation process of large companies. Long-term predictions will not only provide more lead time for decision making but can also provide a more reliable overview of fluctuations in equipment usage. While methods capable of reliably forecasting long-term equipment usage have the potential to improve project delivery, analytical prediction methods, particularly for long-term predictions, have yet to be reported.

2.2 Prediction Modelling

The time series prediction problem is the prediction of future values based on previous and current values of the time series (Hamilton 1994). One-step ahead predictions are referred to as short-term predictions, while multi-step ahead predictions are known as long-term prediction problems. Unlike a short-term time series prediction, a long-term prediction is faced with growing uncertainties arising from various sources. Two major prediction strategies have been described: (a) a recursive prediction strategy, which divides the prediction term into smaller sections, with the same short-term prediction models being calculated repeatedly to achieve a final number and (b) a direct prediction strategy, which only has one model targeting the end of the prediction term. While the latter strategy is more complex, direct prediction strategies can achieve more accurate results (Sorjamaa et al. 2007). To improve accuracy, this study employs a direct prediction strategy.

With the development of computer power and data mining technologies, many prediction algorithms have been proposed in the past decades as forecasting approaches. Linear regression methods are the most basic methods. Although they are easy to interpret, many researchers have claimed that they are not capable of producing accurate results when used for explaining and capturing non-linear relations of many real-world problems (Pai and Lin 2005, Thissen et al. 2003). “Black-box” methods are being increasingly proposed in the data mining field and have attracted attention owing to their powerful capacities and comprehensive adaptabilities. Artificial neural networks (ANN)—the most famous “black-box” method relying on large training datasets—have become a popular method for solving problems (Santos and Celestino 2008). Support vector machines are also being used in construction for their ability to outperform ANNs and to achieve good generalization by adopting a structural risk minimization (SRM) rather than an empirical risk minimization (ERM) (Behzad et al. 2010). A modified version of SVM, called least-squares SVM (LSSVM), has been proposed and applied for solving construction problems (Suykens and Vandewalle 1999, Sorjamaa et al. 2007, Zhang et al. 2016).

To select the right prediction algorithm, however, one must fully understand the defined problem, as the best algorithm will vary case by case, particularly concerning the size of the available dataset. Linear regression is a simple approach for supervised learning that has been shown to be useful both conceptually and practically. Compared to heuristic methods, which require a large amount of historical data and are not always applicable, a multiple regression model can be used even if the observation number is limited.

2.3 Markov Chain

Applications of Markov models have become widely accepted in the field of economics for market forecasting. Markov switching models have been successfully used to model volatility reduction by treating moderation as a discrete event (Kim and Nelson 1999, Stock and Watson 2002). Given its successful

application within the financial sector, the current study proposes the use of a Markov chain process to model construction market fluctuations.

Markov chains, named after Andrey Markov, are used to model a random process that undergoes transition from one state to another on a state space (Gilks 2005). The Markov process is characterized by a single-step memory. It is a stochastic process that involves both a random variable and a "time" parameter that monotonically increases during the process (Figure 1). Only the most recent step is considered when determining a subsequent step. Although restrictive, the most recent step is usually the most important for determining the next step.

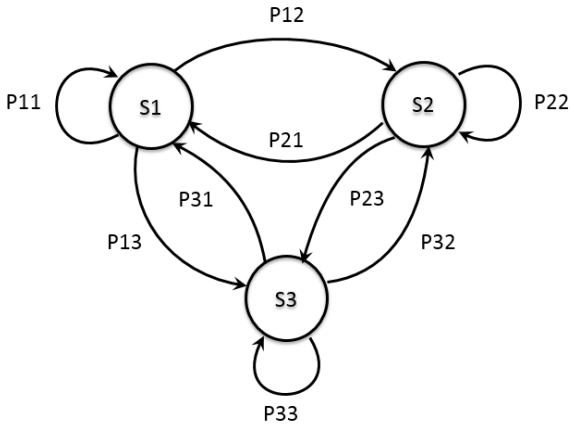

Figure 1: Markov chain process example

3 METHODOLOGY

The objective of this study is to predict long-term equipment usage, in this case over 3-years, from historical revenue data. The proposed methodology is illustrated in Figure 2.

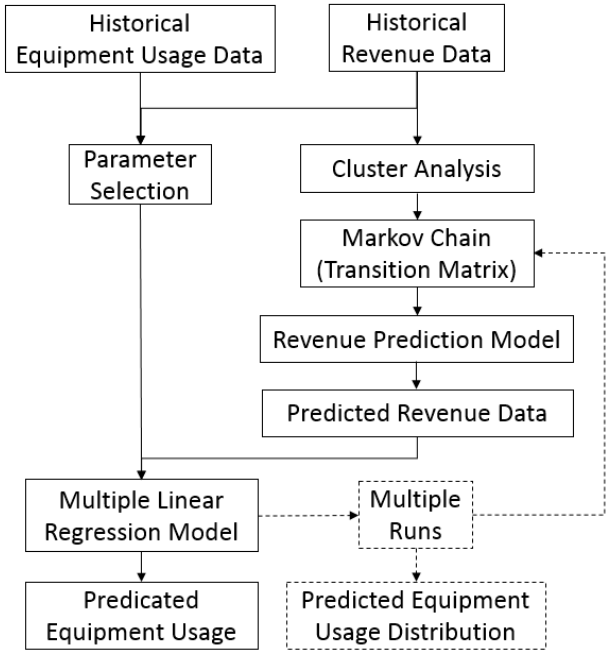

Figure 2: Long-term prediction methodology flowchart

For many general contractors, key performance indexes of each department, including project duration and cost, are recorded in an internal system. Each department has diverse equipment needs and, therefore, these data have an underlying relationship with total equipment hours, companywide. In the proposed model, a Markov chain is used to predict revenue data from this historical data. A direct prediction model is then established to match results with market conditions of the prediction year. From the revenue data, the prediction model for the total equipment usage can be applied to predict monthly equipment hours. Due to the dynamic nature of Markov chains, the model can be simulated for multiple runs to achieve various prediction values (as indicated by the dashed lines in Figure 2). For each run, one random seed is assigned, and the prediction value of equipment usage is calculated. After collecting and combining values from multiple runs, a distribution of the predicted equipment usage is generated.

3.1 Regression Model and Parameter Selection

To obtain a more explanatory model with enhanced predictive accuracy, the multiple non-linear regression model, shown in Equation [1], was considered most suitable for this type of study.

[1] $Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \ldots + \beta_k X_i^k + \varepsilon$

where $i = 1,2,\ldots,n$ and usually $\varepsilon \sim N(0, \sigma^2)$

Definition of factors, or so-called predictors, is crucial for regression model development, yet in many situations, many predictors are available. Inclusion of too many predictors, however, may result in overfitting and an increase in the workload required for collecting and processing data. To keep the current model simple and easy to interpret, the law of parsimony, economy, or succinctness is followed to ensure only the most important predictors are included in the final model. A best subset selection method is applied to reduce the number of predictors from a predefined list. Here, regression models including all possible parameter combinations are established, using Equation 1, and a single best model is selected based on certain criteria. Commonly used criteria include C_p, adjusted R^2, and Bayesian Information Criteria (BIC). C_p, defined in Equation 2, is the statistic that penalizes larger models.

[2] $C_p = (n-p-1)\cdot SSE(q)/SSE(p) - (n-2(q+1))$

where q is the number of predictors in the model, n is the observation number, p is the number of predictors in the model, and SSE is the sum of squared residuals.

Adjusted R^2, defined by Equation 3, is calculated as follows:

[3] $R_{adj}^2 = 1- (n-1)/(n-q-1)\cdot SSE(q)/SSE(p)$

where q is the number of predictors in the model, n is the observation number, p is the number of predictors in the model, and SSE is the sum of squared residuals.

The calculation for BIC is defined by Equation [4].

[4] $BIC = -2l(y) + \log(n)\cdot(q+1)$

where q is the number of predictors in the model, n is the observation number, and l(y) is log-likelihood of y.

In most cases, the three parameter selection criteria will lead to similar results. If not, parameters are subjectively selected based on the analytical selection criteria.

3.2 Cluster Analysis

To apply the Markov chain, a set of states must first be defined. In many cases, data have too many categories and must be simplified. In other cases, data may be numerical and must be converted into categorized data. In the proposed method, data are first divided into groups that are meaningful and useful through cluster analysis, which is sometimes referred to as unsupervised classification. In our study, K-means cluster analysis, a prototype-based partitional clustering technique that attempts to find a user-

specified number of clusters (K) represented by centroids, is applied. K-means clustering attempts to categorize data into groups where within-cluster variation is minimized, as shown in Equation [7].

[7] $\min_{C1,..,Ck} \Sigma^{k} \{WCV(C_k)\}$

where $WCV(C_k)$ is within-cluster variation for cluster C_k.

States of the Markov chain are selected when K-means iterations reach a state in which no points are shifting from one cluster to another.

3.3 Markov Chain

Due to fluctuations in construction market conditions and the seasonality of the construction industry, business revenue, and, in turn, equipment usage varies. A Markov chain process is used to model and predict future revenue data. Each Markov chain contains a set of states, $S = \{s_1, s_2, \ldots, s_r\}$. The process begins in one of these states and moves successively from one state to another. Each move is called a step, and each step is related only to the previous step. The transition matrix is also comprised of probability p_{ij}, defined as the probability at which state s_i moves to state s_j,, which are referred to as transition probabilities. Notably, for regular Markov chains, long-range predictions are independent of the starting state. In general, if a Markov chain has r states, then probability p_{ij} can be calculated using Equation [5].

[5] $p_{ij}^{(2)} = \Sigma^{r} p_{ik} \cdot p_{kj}$

If **P** is the transition matrix of a Markov chain and u is the probability vector that represents the starting distribution, the probability that the chain is in state s_i after n steps is the ith entry in the vector as shown in Equation [6].

[6] $u^{(n)} = u\mathbf{P}^{n}$

Using the results of the cluster analysis, the Markov chain is established and run. Predicted revenue data is fed into the multiple non-linear regression model. Outputs of the models are then combined, run for multiple iterations, and fit into a distribution of predicted equipment usage hours.

4 CASE STUDY

To test its functionality, the proposed methodology was applied at a construction organization in Edmonton, Alberta. Historical data from the last four years were collected from the contractor's internal equipment and project management systems.

4.1 Parameter Selection

Parameter selection was used to condense the list of considered factors. Initially, 20 parameters including month, project number counts, revenue for each discipline, company-wide total revenue, and utilization rates of major equipment categories were identified for parameter selection. Using the various model selection methods previously described, five parameters, namely month, total revenue, revenue for building projects, revenue for industrial projects, and revenue for infrastructure projects, were ultimately selected.

Various parameter numbers in the model were compared (Figure 3). Low C_p values, low BIC values, and high R^2 values indicate improved prediction accuracy of a model. Lowest C_p values, using both backward and forward C_p methods, were obtained when five parameters were considered. Consistent with this, the lowest BIC values and highest R^2 values were obtained with the five parameter model.

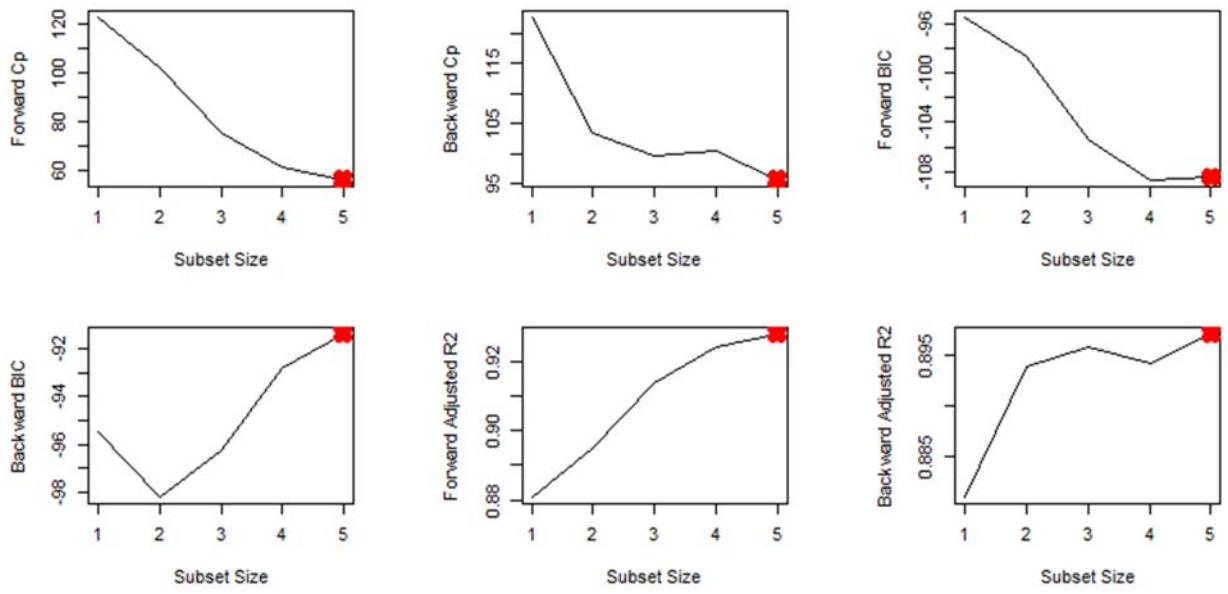

Figure 3: Parameter selection output

4.2 Multiple Non-linear Regression Model

Using the selected parameters and the historical data collected from the contractor's internal equipment and project management systems, the prediction model was established as shown in Equation [8].

[8] Total Equipment Hour = $-42990*Month^3 - 44690*Month^2 + 35420*Month + 0.001985*R_{Building} + 0.002190*R_{Infrastructure} + 0.002063*R_{Industrial} - 0.001995*R_{Total} + 92910$

Where Month is the calendar month ranging from 1 to 12, $R_{Building}$ is the revenue of the building department in the prediction month, $R_{Infrastructure}$ is the revenue of the infrastructure department in the prediction month, $R_{Industrial}$ is the revenue of industrial department in the prediction month, and R_{Total} is the total revenue of different departments in the prediction month.

4.3 Markov Chain and Simulation Model

Revenue data of the case company clustered into two groups. Based on the similarity of the groups to real-world construction market behavior, the two groups were named "Expansion" or "Recession." A simulation model was developed and used to encode the proposed long-term prediction approach. The general purpose template of Simphony, a discrete-event modeling environment, was used to establish the simulation models as shown in Figure 4.

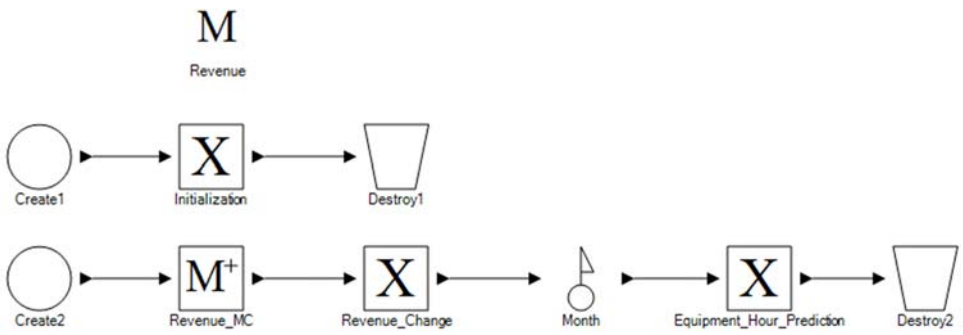

Figure 4: Long-term prediction simulation model embedded with Markov chain

The Markov chain element in Simphony was used simulate the transition between states, and the “Execute” element was embedded with a customized code to export the output into a data file. In this case, the Markov chain model was coded into the “Revenue” and “Revenue_MC” elements. The calculation of revenue prediction was coded into the “Revenue_Change” element. The calculation of equipment hour prediction model and export of calculation results were coded into “Equipment_Hour_Prediction” element.

4.4 Results

The model was used to predict long-term equipment usage hours, as shown in Figure 5, with the blue line and grey range representing the distribution mean and 95% confidence interval, respectively. Cyclical fluctuations in predicted equipment usage were observed. The 10th to 90th percentile of the predicted equipment usage hours was 21,082,486 CAD to 171,438,079 CAD with a period of approximately 11.9 months. As expected, uncertainty accumulated and enlarged as prediction time increased (i.e., the range of prediction values in month 36 is larger than the range for month 12).

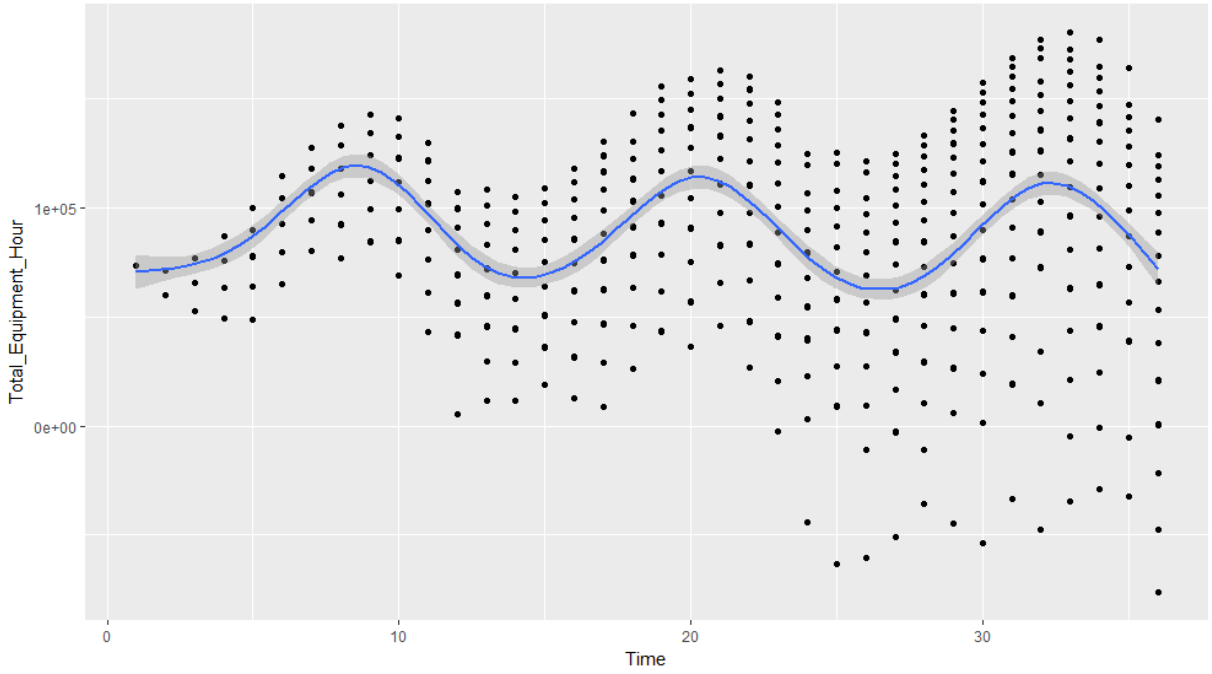

Figure 5: Long-term prediction output

5 CONCLUSION

This research proposes an analytical approach, which combines regression models and Markov chain processes, for the long-term prediction of equipment usage. Using historical equipment and revenue data, the model is capable of predicting future equipment usage, as a distribution, in consideration of predicted company revenue. The functionality of the proposed method was examined following its practical application. The model was found capable of providing managers with a controllable, analytical method for long-term predictions, thereby alleviating the need of managers to rely on subjective experience for critical decision making and on their potentially limited insight of future market conditions. Notably, the proposed model can be easily adapted to predict other usage data, such as labour man-hours, allowing the quantitative comparison of multiple, long-term usage predictions. Nevertheless, the prediction method should be improved and generalized to enhance its adaptability to other, long-term prediction needs. Notably, since the long-term prediction approach cannot be detached from the processed historical data, a need to update the model as more data become available will be required.

Acknowledgements

This research work was funded by Collaborative Research and Development Grant (CRDPJ 492657) from the Natural Sciences and Engineering Research Council of Canada. The authors would like to acknowledge Graham Industrial Services LP for providing data and research support.

References

Bayzid, S.M. 2014. *Modeling Maintenance Cost for Road Construction Equipment*, Doctoral Dissertation, University of Alberta, Edmonton, Alberta, Canada.

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

Behzad, M., Asghari, K. and Coppola Jr, E.A. 2009. Comparative study of SVMs and ANNs in aquifer water level prediction. *Journal of Computing in Civil Engineering*, **24**(5): 408-413.

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.

Fan, H., AbouRizk, S., Kim, H. and 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.

Gilks, W.R., 2005. Markov Chain Monte Carlo. *Encyclopedia of Biostatistics*, John Wiley and Sons, Hoboken, NJ, USA.

Hamilton, J.D. 1994. *Time Series Analysis*. Princeton University Press, Princeton, NJ, USA.

Kim, C.J. and Nelson, C.R. 1999. Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. *Review of Economics and Statistics*, **81**(4): 608-616.

Lucko, G., 2003. *A Statistical Analysis and Model of the Residual Value of Different Types of Heavy Construction Equipment*, Doctoral Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

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.

Mitchell, Z., Hildreth, J. and 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.

Pai, P.F. and Lin, C.S. 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, **33**(6): 497-505.

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

Santos Jr, O.J. and Celestino, T.B. 2008. Artificial neural networks analysis of Sao Paulo subway tunnel settlement data. *Tunnelling and Underground Space Technology,* **23**(5): 481-491.

Stock, J.H. and Watson, M.W. 2002. Has the business cycle changed and why? *NBER Macroeconomics Annual*, **17**: 159-218.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y. and Lendasse, A. 2007. Methodology for long-term prediction of time series. *Neurocomputing*, **70**(16-18): 2861-2869.

Suykens, J.A. and Vandewalle, J. 1999. Least squares support vector machine classifiers. *Neural Processing Letters*, **9**(3): 293-300.

Thissen, U., Van Brakel, R., De Weijer, A.P., Melssen, W.J. and Buydens, L.M.C. 2003. Using support vector machines for time series prediction. *Chemometrics and Intelligent Laboratory Systems*, **69**(1-2): 35-49.

University of Toronto. 2001. *A Guide to Construction Cost Sources*, Prism Economics and Analysis and the Department of Civil Engineering, University of Toronto, Toronto, Ontario, Canada.

Vorster, M.C. 1980. *A Systems Approach to the Management of Civil Engineering Construction Equipment*, Doctoral Dissertation, Stellenbosch University, Stellenbosch, South Africa.

Yip, H.L., Fan, H. and 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.

Zhang, L., Wu, X., Ji, W. and AbouRizk, S.M. 2016. Intelligent approach to estimation of tunnel-induced ground settlement using wavelet packet and support vector machines. *Journal of Computing in Civil Engineering*, **31**(2): 04016053.