CITC GLOBAL
Construction in the 21st Century

The Thirteenth International Conference on Construction in the 21st Century (CITC-13)

Arnhem, The Netherlands| May 8 – 11, 2023

Scientometric Analysis and Machine Learning as Tools for Predicting Construction Equipment's Economical Factors

Kleopatra Petroutsatou[1], Ilias Ladopoulos[1], Iason Kosmidis[1]

[1] Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
kpetrout@civil.auth.gr

Abstract

This study is based on previous research about estimating the residual market value (RMV) of excavators, using machine learning (ML) techniques. Boosted by the promising results of the initial study, the current one seeks to expand the prediction algorithm on different types of earthworks equipment, also by enriching the equipment's database. The conducted scientometric analysis revealed the remaining untapped knowledge concerning ownership, operation and maintenance cost data, that is still been gathered by the global construction equipment (CE) manufacturers, owners or dealers. The current study attempts to highlight the perspectives that machine learning offers, when coming to predict several economic factors from the utilization of different types of CE. The equipment's RMVs were collected from global equipment resellers and auctions. The prediction model was developed by using RapidMiner Studio software, while the scientometric analysis was performed using VosViewer software. The ultimate goal of this approach was to develop a user-friendly platform, supported by RapidMiner Studio, to predict RMVs through certain values, which represent the equipment's financial and operating condition. The algorithm concluded that for any CE, the most dominant factor for predicting its residual market value is its initial purchase.

Keywords

Construction equipment, machine learning, residual market value, prediction, scientometric analysis

1. Introduction

A scientometric analysis was initially performed, with a view to investigate research gaps, scientific trends and interrelationships between the most common used scientific terms.

According to Vorster (2009) [1], the RMV of a machine when sold at any point in its life is an unknown that depends on many factors. Previous multidiscipline studies resulted that one of the most unstable, fluctuating but dominant factor, is the market itself. For this is the reason a machine might have different residual values in different years depending on many factors with the most dominant one that of the construction output of the year under search. If, for example, there is a high demand for a specific model then its residual value is high, whereas for the same model one year before its market value might have been less due to low market interest. To comprehensively cover a wider range of CE types, this study collects the resale and auction prizes of 1000 excavators, 1000 loaders and 977 dozers from Caterpillar. Their manufacturing year ranged between 1980 and 2020. A representation of those equipment data is depicted on Table 1.

Table 1. CEs Values Range

	CE Type	Hours of use	Manufacturing Year	Residual Market Value (€)
1	Excavator	0 – 63.788	2001 - 2020	104 – 914.112
2	Dozer	10 – 275.226	1981 - 2020	20.393 – 1.252.233
3	Loader	8 – 12.6045	1980 - 2020	14.855 – 1.202.365

The equipment data were sourced from the most popular internet sites for sales and auctions, such as constructionequipmentguide.com, euroauctions.com, machinerytrader.com and catused.cat.com. Equipment records concerned the global market for the 1st quarter of year 2021. Those records were collected and classified according to their type, size, manufacturing year, country of reselling (equals the country of purchasing), effective hours (hours of use – reflecting the condition of the machine) and their current RMV, resulting a database consisted of almost 1000 records for each equipment type.

This study attempts to move one step further, by calculating the initial purchase price of those equipment, with the assumption that the year of purchase equals the manufacturing year. For this conversion, equation (1) was applied.

$$Present\ Value = \frac{Future\ Value}{(1+i)^n} \qquad (1)$$

where:
i = the interest bank rate of the country in which the equipment was purchased
n = period of time

The interest rate of each country for every specific year of purchase was extracted by the World Bank's website (www.data.worldbank.org).

Several publications were studied, to reveal the relations between ML and CE financial management, as shown in Table 1. ML was implemented using RapidMiner Studio software. Among various ML capabilities, RapidMiner offers the Auto Modelling ability, where the user provides the necessary data and defines the prediction attribute. The main goal of this study is to identify the patterns under which the CEs' RMVs are evolving, under demanding market rules, and to present robust estimations of RMVs.

Table 3. Previous studies on ML application for RMV estimation

Authors	Year	Title	Contribution
Bertoni et al. [6]	2017	Mining Data to Design Value a Demonstrator in Early Design	They applied data mining algorithms on a dataset build on a wheel loader's performance and contextual and environmental data. They focused their estimation on the fuel consumption of alternative design concepts and estimated the performance variations given different contextual variable.
Fan and Jin [7]	2011	A study on the factors affecting the economic life of heavy construction equipment	They managed to extract rules leading to different cost patterns and therefore different economic life spans of heavy equipment, more effective maintenance strategies and to an accurate comparison among the equipment cost performance from various classes, makes and amount of service during their life cycle.
Spinelli et al. [8]	2011	Annual use, economic life and residual value of cut-to-length harvesting machines	They gathered a large database of second-hand machine sale offers and conducted a statistical analysis. They concluded that the equipment's residual value is strongly related to machine age.
Fan et al. [9]	2008	Assessing Residual Value of Heavy Construction Equipment Using Predictive Data Mining Model	Stressed the importance of predicting the residual value of heavy construction equipment to an acceptable level of accuracy, to maximize the return of this investment. They introduced a data mining-based approach for estimating the residual value of heavy construction equipment.
Lucko et al. [10]	2006	Statistical Considerations for Predicting Residual Value of Heavy Equipment	Identified the factors that affect the residual value of a construction equipment, and they examined them comprehensively by analyzing real market data from equipment auctions, about track dozers.
Lucko [11]	2003	A Statistical Analysis and Model of the Residual Value of Different Types of Heavy Construction Equipment	Through multiple linear regression analysis, he performed a residual value prediction, by using auction sales data and heavy construction equipment manufacturers' publications.

Michell [12]	1998	A Statistical Analysis of Construction Equipment Repair Costs Using Field Data & The Cumulative Cost Model	By using field data on 270 heavy construction machines, he identified a regression model that can adequately represent repair costs in terms of machine age in cumulative hours of use.

2. Methods

2.1 Scientometric Analysis

This study attempts to reveal the void regarding the researched topic, by applying a scientometric analysis, to objectively map the scientific knowledge on this specific field and to identify the research themes and the corresponding challenges based on the scientometric results, with the use of the VosViewer application [2], developed by Van Eck & Waltman (2010) [3]. A four-step process was utilized, to create those scientometric maps, as shown in Figure 1.

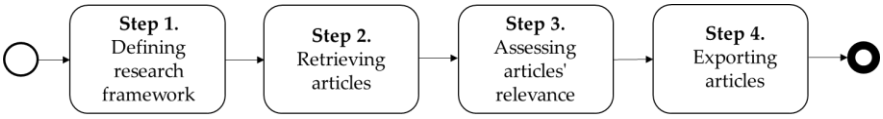

Fig. 1. VosViewer's map creation flowchart

In step one, the research framework is defined, with the scope to identify and set the desired goals. It includes an initial investigation, to separate the relevant from the irrelevant research components. Step two includes the retrieve of the close related articles with the examined topic. Those articles were extracted by Web of Science and Scopus, covering a period from 2001 to 2021, by using the search terms of Table 1. Step three includes the relevance assessment of the extracted publications, to finalize those that are going to be inserted for scientometric mapping into VosViewer. 192 articles were exported for the scientometric analysis.

Table 1. Search Terms Boolean Operators

Boolean operator	Terms	Description
	construction equipment	The term that describes the main topic and the core search rule
OR	equipment	Used for searching all machinery and equipment-based publications, in order to exclude the irrelevant
OR	machinery	
AND	residual market value	Term that specifies the distinctive topic, concerning residual market value
OR	rmv	The abbreviation of residual market value
OR	residual value	Term that specifies an alternative term of residual market value
OR	rv	The abbreviation of residual value
OR	owner*	Term that specifies the distinctive topic, concerning owner, ownership, etc.
OR	cost*	Term that specifies the distinctive topic, concerning cost
AND	engineer*	Term that determines the desired scientific domain
AND	machine learning	Specifies the predicting artificial intelligence method
OR	artificial intelligence	
OR	estimate	Used for searching alternative terms for machine learning
OR	predict*	

Figure 2 presents a significant increase after 2018 for publications and citations covering topics related with the search terms of this study. It signifies the increased scientific interest, following the trend of machine learning integration into many different aspects of the construction industry.

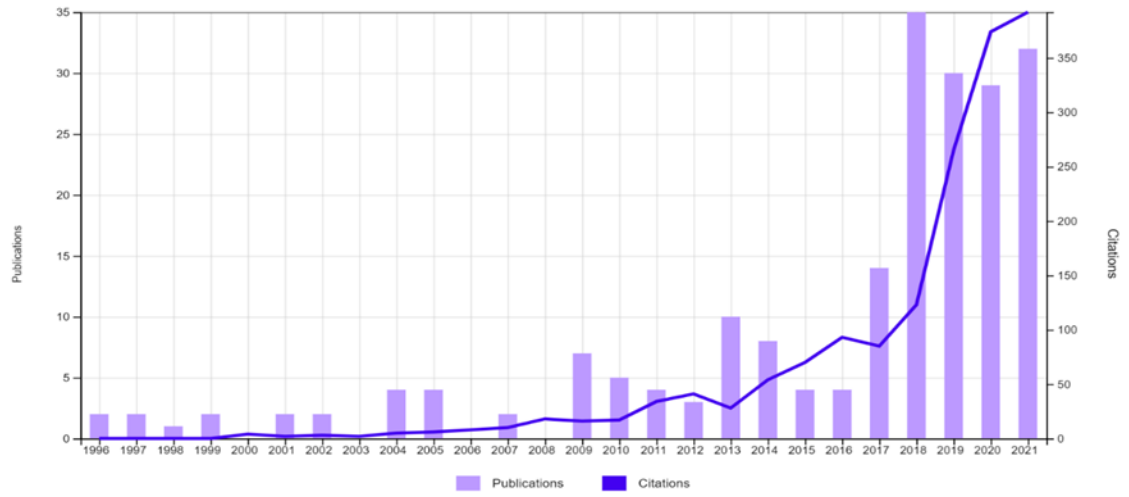

Fig. 2. Total Number of Publications and Citations

The fourth step includes the extraction of the found articles, to be processed by VosViewer. The final product is a comprehensive network comprised by coexisting terms inside the overall publication cluster, where their linkage strength, their occurrences, and their relativity are visible, weighted, and clustered. Every cluster receives a different color, and each color designates a specific research area. The terms inside each cluster are represented by circles, and their size reflects the number of publications in which they were found. The spacing between those circles indicates their relatedness, and their degree of relativity is indicated by the thickness of the curved lines connecting them. The degrees of relatedness between words are indicated by the curved lines [4]. The produced clusters by subject are shown on Table 2, and the keyword co-occurrence network is visualized in Figure 3.

Table 2. Keyword Co-occurrence Clustering

Cluster Number	Main Subject	Color	Terms Included
1	Cost	Red	9
2	Effective hours of use	Green	8
3	Strategy	Blue	8
4	Management	Yellow	7
5	Construction Equipment	Purple	6
6	Useful Life	Light blue	6

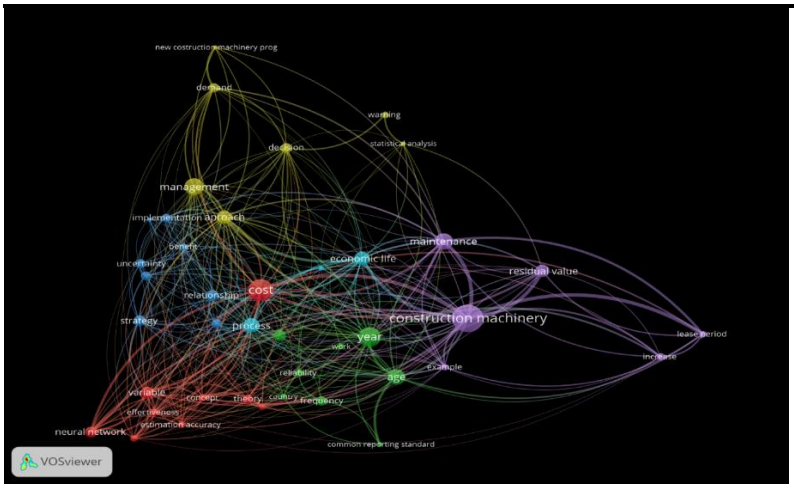

Fig. 3. VosViewer map based on keywords (network visualization)

The analysis based on keywords indicates that the network is constituted by strong links between the terms of "construction machinery", "maintenance", "residual value", "economic life", "cost", "management" and "neural network", even though they belong to different clustering groups. Nevertheless, the "Cost" cluster (red) is located near the "Useful life" cluster (light blue). Specifically, the term "cost" has been putted separately from the rest of the red group, but in a centric position, in order to be closer with the rest of the clusters, indicating their strong relationship. But the most important issue of the network's analysis is the fact that the "machine learning" term is not visible at all. Even the appearance of the "neural network" term, as a ML method, is not considered significantly close (although the thick curved lines indicate a strong term co-occurrence) with the rest of the terms to be able to justify a strong utilization of ML for CE residual value estimation. Thus, its location justifies a "knowledge gap" between ML and the rest of the clusters, a fact that provides a uniqueness to the current study.

2.2 Machine Learning Model

The CE's RMV fluctuation was performed using a machine learning model. This model was developed with RapidMiner Studio software that provides the Auto Modelling ability, which offers sufficient capabilities to accelerate the ML process and analyze the results. It develops different types of ML models, depending on the imported data type, allowing the user to select the most efficient one and proceed to further analysis [5]. The constructed database includes eight distinctive attributes with records from each equipment: a. CE type, b. Effective hours (hours of use), c. Manufacturing year, d. Country of purchase, e. Residual Market Value, f. Interest Rate (by the year and country of purchase) and, g. Purchase Initial Value (PIV). PIV was calculated by applying equation (1) transformed in equation (2).

$$PIV = RMV \times (1+i)^{(2021-Year\ of\ manufacture)} \quad (2)$$

RapidMiner evaluates the imported database for its quality, characteristics, and diversity of the data. It also presents an analysis for which of the data are suitable to be exploited for ML and prediction and which are not. It uses specific quality indicators for the data evaluation: a. Correlation, for data categories very close to the prediction goal or for those that are irrelevant, b. ID-ness, for attributes that all their data are completely different, c. Stability, for attributes that have a great percentage of the same data or values and d. Missing, for data categories with absence of values. At the final stage, the software proposes several prediction models. The user can compare the performance of each model and decide which is the best, depending on the relative error, the standard deviation and the training time.

The imported database can be processed by three types of ML: prediction, clustering and finding outliers. This study focuses on RMV prediction. The machine learning techniques used by RapidMiner are: a. Generalized linear model, b. Decision tree analysis, c. Deep learning, d. Random forest, e. Gradient Boosting trees and f. Support vector machine. Overall, one of the main advantages of all the machine learning tools is that these models lead to an effective visual analysis, since are similar to that of a human's neural network, and thus they can process unorganized data [6].

The results for each type of CE are presented on a table form summary (Table 4) and the best method is been proposed, according to the best performance (due to the minimum correlation), the minimum scoring type and the minimum total time (scoring + training), for each of the five types of error: a. for the Mean Squared Error (MSE), a risk function and a measure of the estimator's quality, which measures the average of the squares of the errors, that is, the average squared difference between the estimated and the actual value, b. for the Root Mean Squared Error (the root of MSE), which is a measure of accuracy, to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power, by comparing forecasting errors of different models for a particular dataset, c. for the Absolute Error, which sums up the absolute values of the differences between labels and predictions and divides this sum by the number of examples, d. for the Relative Error, which is the average of the absolute deviation (error) of the prediction from the actual value divided by the actual value, and e. for the Correlation, which describes the strength of relationship between two numeric attributes in a data set and also the direction of this relationship (positive, negative or none). The bold values in Table 4 represent the proposed ML model with the best performance (minimum error) for each type of CE.

Table 4. Error comparisons

Model	CE Type	Root Mean Squared Error	Absolute Error	Relative Error	Mean Squared Error	Correlation
	Dozers	56.531,416	38.772,172	18%	3.320.836.660,315	0,947

Generalized Linear Model	Loaders	31.975,959	11.718,0	7,55%	1.447.867.413,916	0,969
	Excavators	**7.084,824**	7.457,271	**4,3%**	**52.353.904**	0,997
Deep Learning	Dozers	**34.293,262**	19.153,017	11,2%	**1.266.307.185,587**	**0,984**
	Loaders	25.457,175	12.753,6	7,3%	710.876.496,955	0,982
	Excavators	12.496,611	**3.872,302**	8,6%	181.146.105	0,996
Decision Tree	Dozers	35.964,155	**15.164,809**	**6,3%**	1.529.876.859,582	0,978
	Loaders	19.833,866	8.134,19	3,4%	528.486.793,713	0,983
	Excavators	20.937,937	7.216,24	4,5%	481.976.300	**0,998**
Random Forest	Dozers	41.611,137	19.723,803	8,2%	1.926.634.217,607	0,978
	Loaders	**10.693,513**	**4.584,70**	**3,4%**	**182.614.092,579**	**0,997**
	Excavators	21.023,473	9.149,153	6,9%	483.738.082	0,989
Gradient Boosted Trees	Dozers	47.663,163	17.708,397	6,9%	2.606.364.645,869	0,957
	Loaders	27.674,42	9.970,41	4,7%	977.796.280,210	0,967
	Excavators	17.062,412	6.673,099	5,2%	338.436.652	0,992
Support Vector Machine	Dozers	155.200,944	96.717,353	32,8%	24.513.375.194,25	0,708
	Loaders	85.182,299	44.436,5	20,2%	7.897.000.448,640	0,58
	Excavators	105.446,357	53.850,22	28,1%	11.506.444.351	0,557

3. Results

Further analysis of Table 4 reveals the best ML method for predicting the RMV of each CE. So, when coming to dozers, Deep Learning appears to have the best performance, since this ML method present the less errors. For loaders, Random Forest performed better than any other method and for excavators, the Generalized Linear Model was dominant.

After every prediction, RapidMiner provides also a weighting table for each attribute, as shown in Table 5, where, for every CE, the purchase – initial value was the most crucial factor, when coming to predict their RMV.

Table 5. Attribute Weighting

Attribute	Weight		
	Dozers	**Loaders**	**Excavators**
Purchase – Initial Value	**0.871**	**0.867**	**0.885**
Manufacturing Year	0.204	0.180	0.157
Country	0.084	0.058	0.053
Effective Hours	0.082	0.039	0.031
Type	0.016	0.029	0.019
Country's Interest Rate	0	0	0.003

For the purpose of this study the simulation ability of RapidMiner was also utilized, to identify the accuracy of the ML model and if the predicting results are reflecting the market reality. In the selected example the user can choose between different types of Caterpillar loaders, eg. Cat 980M Wheel Loader and insert the depicted in Figure 4 parameters. In this way he will be able to know how much he should sell the equipment after five years of use, with 4.231 operating hours and with an initial purchase price of 313.932€, according to the market rules. The algorithm predicts that he should sell the equipment for the price of 278.127€, which is close with the average price of 265.000€ in the selling market of 2021. In the same way, the user who just purchased a CE could also predict its residual market value after a certain number of years and with certain operating hours.

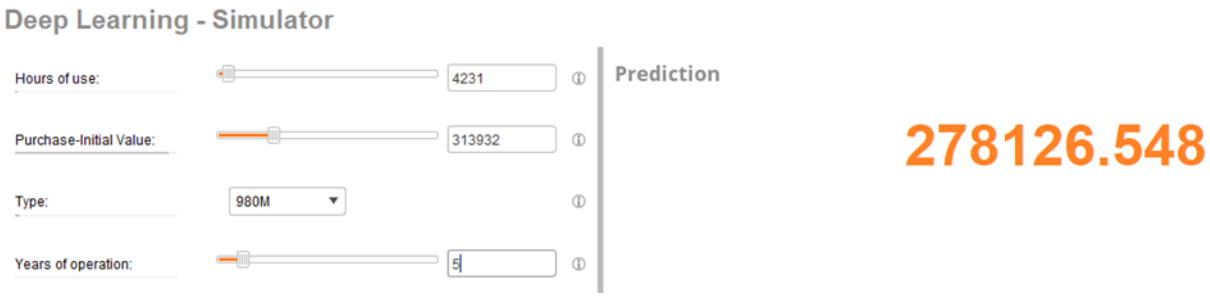

Fig. 4. RMV Prediction Simulator

4. Discussion

The idea of utilizing ML methods to estimate CE's economic indicators was a challenge worth to be considered also by previous studies, but they were only examining the prediction problem from the statistical perspective or by applying specific ML methods. However, no research has been found to systematically summarize different ML methods and provide the chance to select the one that best performs. Furthermore, this study dives deeper into the ML possibilities, to develop a software assisted simulation tool for predicting the CE's RMV, taking into account strict market rules. Moreover, this analysis is based on the assumption that all the machines are maintained adequately accordingly the manufacturers' manuals depending on hours worked and site conditions.

The practical applications of this study principally relate to helping respecting stakeholders to better understand the importance of predicting CE's economic factors. More specifically, it offers valuable indicators to: (i) monitor the RMV evolution of any kind of equipment, (ii) reinforce their financial management, iii) predict certain financial indicators and (iv) thus ensure the most efficient and reliable return of investment.

5. Conclusions

Construction equipment represents a significant capital investment for its owner. Thus, its management strategy, which includes monitoring every financial parameter during his operating life, offers the opportunity to gain most of this resource. This study was challenged to exploit the scientific gap concerning the examined topic, in order to offer a user-friendly and accessible method to monitor and predict the residual market value of three different kinds of equipment. Assisted by VosViewer software, the conducted scientometric analysis offered the opportunity to identify scientific trends and gaps, but also to examine the interrelationships between multidisciplinary domains, with the scope to understand the way the CE market evolves and behaves.

This study presents an innovative approach on predicting the RMV of different types of construction equipment, by using advanced ML techniques, through RapidMiner software. The equipment initial purchase value was the most important factor that determined its RMV. This result supports the idea that to ensure a successful financial strategy, the purchasing price should be taken under serious consideration. It signifies the need for applying a predefined and well-planned construction equipment replacement strategy, taking into account current or future market and auction trends.

RapidMiner analyzed a database consisted by 1000 excavators, 1000 loaders and 977 dozers from Caterpillar, trained the ML model, and presented interesting fluctuations of their RMV. This huge data enables more secure predictions if treated well. The software processed six different ML methods and proposed the most efficient ML method for each equipment: i) Deep Learning for dozers, ii) Random Forest for loaders, and iii) the Generalized Linear Model for excavators. For some equipment the model presented great differences between the predicted and the current RMV, mainly because the current RMV is not a mathematically calculated value, but it is determined by the market itself and the law of supply and demand, confirming that the economic situation of the country and the indexes of construction output play a significant role in determining market's trends.

References

1. Vorster M., 2009. *Construction Equipment Economics*. Pen Publications, Blacksburg, USA [1]
2. Tijssen, R., & Van Raan, T. (1994). Mapping Changes in Science and Technology: Bibliometric Co-Occurrence Analysis of the R&D Literature. *Eval. Rev.*, 18, 98–115. [2]

3. Van Eck N.J., & Waltman L. (2010). Software survey: VosViewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. [3]
4. Petroutsatou, K., Ladopoulos, I., & Tsakelidou, K. (2022). Scientometric Analysis and AHP for Hierarchizing Criteria Affecting Construction Equipment Operators' Performance. *Sustainability*, 14, 6836. [4]
5. Petroutsatou, K. Ladopoulos, I., & Polyzos, M. (2020). Construction Equipment's Residual Market Value Estimation Using Machine Learning. *Proceedings of the XIV Balkan Conference on Operational Research (Virtual BALCOR 2020), Thessaloniki, Greece*, 239-243. [5]
6. Bertoni A., Larsson T., Larsson J., & Elfsberg J., (2017). Mining data to design value: A demonstrator in early design. *Proceedings of the 21st International Conference on Engineering Design (ICED17), Vancouver, British Columbia, Canada*, 21-29. [6]
7. Mitchell Z. (1998). A Statistical Analysis of Construction Equipment Repair Costs Using Field Data & the Cumulative Cost Model. *Accession 16244, Doctoral Dissertation*, Virginia Polytechnic Institute and State University, Virginia, United States. Virginia Tech DSpace.
8. Lucko G. (2003). A Statistical Analysis and Model of the RV of Different Types of Heavy Construction Equipment. *Accession No. 14481, Doctoral Dissertation*, Virginia Polytechnic Institute and State University, Virginia, United States. Virginia Tech DSpace.
9. 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.
10. Lucko G., Anderson-Cook C. & Vorster M. (2006). Statistical Considerations for Predicting Residual Value of Heavy Equipment. *Journal of Construction Engineering and Management*. 132(7), 723-732.
11. Fan H. & Jin Z. (2011). A Study on the Factors Affecting the Economical Life of Heavy Construction Equipment. *Proceedings of the 28th International Symposium on Automation and Robotics in Construction (ISARC 2011), Seoul, Korea*, 923-928.
12. Spinelli R., Magagnotti N. & Picchi G. (2011). Annual use, economic life and residual value of cut-to-length harvesting machines. *Journal of Forest Economics*, 17(2011), 378-387.