Data-Driven Approach for Analysis of Performance Indices in Mobile Work Machines

Teemu Väyrynen, Suvi Peltokangas, Eero Anttila, and Matti Vilkko
Department of Automation Science and Engineering
Tampere University of Technology,
Korkeakoulunkatu 10, Tampere, Finland
e-mails: firstname.lastname@tut.fi

***Abstract*—This paper presents a data-driven approach for the analysis of performance indices in mobile work machines. Performance analysis and optimisation of mobile work machines has become increasingly important in recent years. The mobile work machine optimisation is performed based on performance measurements. One of the most interesting and potential approach for improving the quality of the performance analysis is the utilisation of Big Data and data-driven analysis methods, such as machine learning. This study utilises a machine learning algorithm, Classification and Regression Trees (CART), in the performance analysis of the mobile work machines. The most significant benefit of the presented method is that it provides a statistical reference of the machine performance for the operators. The method enables operators to compare performance against reference fleet of machines working in similar operating conditions. This feature can lead to more informative and reliable interpretations and analysis of the performance values. The results of this paper demonstrate how the presented method was used to analyse the performance of a mobile work machine fleet.**

Keywords–performance; mobile work machine; regression tree; CART.

I. INTRODUCTION

Performance analysis and optimisation of mobile work machines has become an increasingly important trend within the industry in the recent years [1], [2]. Both, the mobile work machine manufacturers as well as the operators have started to pay more attention to the performance optimisation of the machines. Optimising the performance of the mobile work machine results in increased productivity and efficiency. However, the optimisation of the mobile work machine is difficult if the performance of the machine cannot be measured and analysed accurately. The importance of the performance analysis is the main motivation for this work.

The objective of this work is to present a data-driven approach that utilises machine learning to assist the operators in the performance analysis of the mobile work machines. The approach is a combination of data preprocessing and Classification and Regression Trees (CART). CART is a supervised machine learning algorithm, that constructs classification and regression trees to model systems [3]. In this work, CART is used to model the relation between the different operating conditions and the performance of the machines based on the data of a mobile work machine fleet. The predictions of the model enable operators to compare the performance against reference fleet of machines working in similar operating conditions. This feature can provide more informative and reliable interpretations and analysis of the performance values.

The performance analysis of a mobile work machine is a challenging task due to the various factors affecting the performance. These factors are, e.g. objectives of the work, operating conditions, skill level of the operator, work load, technical properties of the mobile work machine, and control parameters. Figure 1 describes the factors affecting the performance of the machines. This work focuses on analysing the relation between the operating conditions and the performance of the mobile work machine.

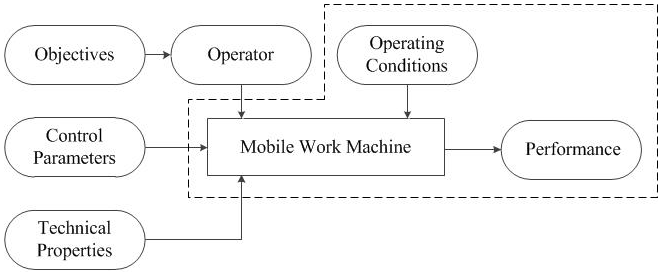

Figure 1. Factors affecting the performance of a mobile work machine. The main focus of this work is delimited by the dotted lines in the figure.

Improved performance analysis enables the operators to optimise the operations of the mobile work machine by tuning the control parameters of the automation system. Depending on the complexity of the machine, the automation system can allow operators to customise hundreds of parameters based on their personal preferences and requirements of the operating conditions. These parameters have a major impact on the operational performance of the mobile work machine in terms of efficiency and productivity.

Conventionally, parameter optimisation has been performed based on the rules of thumb developed by skilled instructors and machine operators. The proper tuning of a mobile work machine is an extremely difficult and time-consuming task especially for an inexperienced operator [4]. Various measurement and performance values can be presented to the operators via the graphical user interfaces of the machines. However, the interpretation of the performance values is most often left to the operators.

These interpretations are often made without proper understanding about the relation between the operating conditions and the performance of the machine. Also, due to the restricted performance analysing capabilities of the human operators, the results of the analysis might be incorrect. However, by utilising reference data and advanced data analysis methods, the

 ISBN: 978-1-61208-423-7

operators gain valuable information to support their analysis of the machine performance.

Originating from the described situation, the research problem of this work focuses on improving the analysis of the performance values in mobile work machines. Derived from the identified problem, the research question of this work is: How can the analysis of performance values be improved in mobile work machines?

The rest of this paper is organized as follows. Section II addresses the state of the art in analysis of performance values and describes the requirements set for the solution. Section III introduces data preprocessing and the CART algorithm. Section IV describes the design of experiment and the results. Section V sums up the work and proposes future research topics.

II. State of the art and requirements

This section introduces the state of the art in the analysis of the performance values in mobile work machines. The requirements set for the solution method are also introduced in this section.

A. State of the art

A wide range of methods have been applied to analyse the performance values of mobile work machines. These methods vary from simple monitoring of measurement values to more sophisticated and holistic analysis. The requirements of the data analysis and the application specific features of the data determine which method is most suited to the given application.

Data-driven performance analysis methods, which require domain expert knowledge, have been presented for mobile work machines and industrial processes [1], [2]. Various other research papers have addressed the problem of analysis and optimisation of performance values in mobile work machines [5]–[9]. These studies have used such methods as statistical data analysis, modelling, root-cause analysis, and optimisation to improve the operational performance of the mobile work machines. Previous research with machine learning algorithms has provided promising results also in the fields of agriculture and industry [10]–[12]. Due to privacy policy of the mobile work machine industry, it is difficult to find up-to-date information about data analysis methods utilised in the analysis of performance values in mobile work machines.

B. Requirements set for the data analysis method

The requirements set for the data analysis method is derived from the objectives of this work. The identified requirements are:

- The method should be able to predict typical performance values for machines in different operating conditions.
- The method should enable easy updating of the model as more measurement data is acquired.
- The model structure should be easy to interpret and utilise in the performance analysis and optimisation.
- The method should select the most relevant input variables for modelling, without extensive prior knowledge about the data.

In the present study, we selected a combination of data preprocessing and CART algorithm to analyse the performance values. CART was selected for this study because it meets the described requirements and also provides features such as nonparametric modelling, robust handling of outliers, computational speed, etc. [3]

III. Methods

This section introduces the data preprocessing and CART method utilised in this work. We also adress the regression tree complexity selection.

A. Data preprocessing

Among the most important factors affecting the performance of machine learning algorithms are the quality and the quantity of the data. In order to create accurate and reliable models from the data, the amount of irrelevant, erroneous, and redundant data should be low. The main goal of data preprocessing phase in this work is to provide a high-quality data set for the machine learning algorithm. [13] There is no standard method for data preprocessing; instead, a set of general guidelines and procedures have been proposed. The requirements for data preprocessing are set by the characteristics of the data and the objectives of the data analysis. [13]

Factors that need to be considered while performing data preprocessing include variable selection, detection and removal of outliers, missing value handling, discretization, resampling, data normalisation, and dimension reduction. Application-specific knowledge of the data preprocessing requirements is usually required. This knowledge can be acquired from machine operators, mathematical models, or by examining the characteristics of the data. [13]

The quality of the data used with the machine learning algorithms is important [13]. The original data is divided into two subsets: training data and validation data. The training data is used for creating the model and validation data is used to validate the prediction accuracy of the model. In order to model the system comprehensively, the variables in the training data need to have a sufficient amount of variation and scale. The correct and incorrect selections of training data is presented in Figure 2.

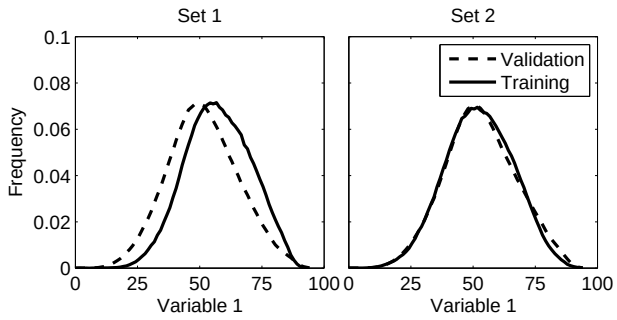

Figure 2. Incorrect (left) and correct (right) selection of training data sets.

The distributions in Figure 2 describe an example of an operating condition measurement in the data. On the right side of Figure 2, variable in the training data covers the whole scale of the validation data. The incorrect selection of training data is presented on the left side of Figure 2. Variable in the

training data does not cover the same scale as the validation data. Therefore, the model, which is generated by a machine learning algorithm, is expected to lack prediction accuracy as some parts of the validation data are not included in the model.

B. Classification and Regression Trees (CART)

CART is a supervised machine learning method that was originally presented by Breiman et al. [3]. CART constructs a model called a decision tree between input and output variables of the data. Two decision tree types are classification trees and regression trees. If the output variable has discrete and predetermined values (that is, classification problem), CART constructs a classification tree. However, if the output variable has continuous values (that is, the regression problem), CART constructs a regression tree. [3] In this work, CART is used to model the relation between the measurement variables describing the operating conditions and the performance of a mobile work machine.

The first stage of utilising regression tree in modelling is the selection of a data set. The data set consists of input and output variables, where inputs are parts of measurement space and outputs are real-valued numbers. The input variables are also known as predictor or independent variables. The outputs are called response or dependent variables. Regression tree creates a real-valued prediction function between the predictor and the response variables. The prediction function can be utilised in two different purposes: to predict the responses based on new predictor measurements, and to understand the relations between the response and predictor variables. [3]

A decision tree is constructed by splitting the data into subsets that are also known as nodes. The building of the decision tree starts from a root node that contains all of the data. A binary split is performed for the root node in a way that the split minimises the fitting error between response values and the predictions of the model in the two child nodes. The splitting variable (that is one of the predictor variables) and its value are the ones that minimise the fitting error. The splitting is then performed recursively for each child node. [3] An example structure of a decision tree is illustrated in Figure 3.

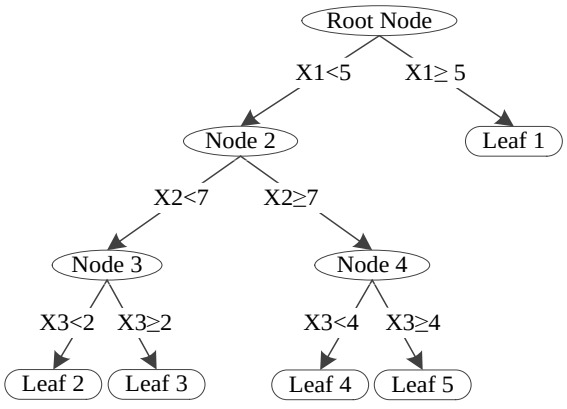

Figure 3. An example of a decision tree structure.

The splitting is continued until the proposed child nodes no longer decrease the fitting error of the regression tree or one of the user-specified stopping criteria is reached. Various stopping criteria can be applied to the splitting, e.g. maximum number of terminal nodes and minimum number of measurement values in each terminal node. The terminal nodes of the regression tree are called leaves. In each leaf, the predicted response value is the average of the response values in the leaf. The following pseudocode describes basic procedure for creating the prediction function with regression tree. [3]

1) SELECT the data used for modelling
2) INSERT the data into the root node of the regression tree
3) SPLIT the data of the node into two child nodes in a way that the fitting error is minimized
4) END IF one of the stopping criteria is met or every node holds only identical response values
5) SELECT the node that has the greatest potential for fitting error reduction
6) CONTINUE from step 3

As presented in the pseudocode, the regression tree continues the splitting of the data until every leaf holds only identical response values or one of the user-defined stopping criteria is met. If the stopping criteria are not used, the result of the modelling is a highly complex and over-fitted regression tree structure. The increased complexity of the regression tree does not necessarily result in improved prediction accuracy with new data. Therefore, while utilising a regression tree in practical applications, a compromise between tree complexity and prediction accuracy is often desired. The required balance between these features is considered to be application-specific. [3]

The selection of tree complexity can be performed with previously described stopping criteria and with pruning methods. Pruning methods can be used to simplify complex regression trees. The basic principle of the regression tree pruning is to decrease the number of leaves in the regression trees. The number of splits in the regression tree is pruned starting from the split that has the least effect on the fitting error. The level of pruning is selected subject to the desired complexity of the regression tree. [3]

The prediction accuracy of the regression tree can be estimated with the following procedure. First, the regression tree is created with a training data set and it is pruned to a desired level. Regression tree enables the estimation of prediction accuracy with resubstitution error and cross-validation error. The prediction accuracy of the regression tree is also validated with a validation data set. The validation data is measured from the same system as the training data, but it is not used in the creation of the regression tree. Comparing the original responce values of the validation data and the predictions of the regression tree, one can estimate the prediction accuracy of the model. [3]

IV. RESULTS

This section is divided into two subsections. The first subsection describes the design of the experiment and the utilised data. The second subsection introduces the results of the experiment.

A. Design of experiment

The purpose of the experiment is to test how the performance of a mobile work machine fleet can be analysed with

 ISBN: 978-1-61208-423-7

TABLE I. THE METRICS OF THE EXPERIMENT DATA

Type of metrics	*Amount*	*Description*
Machines	17	Preclassified machines, 10 training and 7 validation machines
Training data	2,254,901	Approximately 66 per cent of data for training and 34 per cent for validation
Validation data	1,178,485	
Predictor variable	8	Operating condition measurements
Response variable	1	Performance measurement

the regression tree. The scope of the experiment is focused on modelling the relation between the operating conditions and the performance values of the mobile work machines, as presented in Figure 1.

In this work, the regression tree is used to perform three consecutive actions. The following steps demonstrate an example of an approach that could be used in the performance analysis of a mobile work machines.

1) Model the relation between the operating conditions and the performance based on the data of a mobile work machine fleet.
2) Assign typical performance values for each operating condition.
3) Utilise the typical performance values in the performance analysis of an individual mobile work machine.

In order to test the proposed method, a mobile work machine data base was collected. The data for the experiment was acquired from a global mobile work machine manufacturer. A data set including 17 mobile work machines was selected for this work. The machines were selected based on preclassification criteria which were the same machine model and same country of operations. This kind of preclassification was performed to decrease the undesired variations in the data. These variations are caused by the different performance standards and operating conditions between the countries. Table I presents the metrics of the data set used in the experiment.

The data set used in the experiment was generated by combining data from measurement data bases. Additional data preprocessing methods applied to the data set were resampling, missing value handling, and data normalization. The data was then divided into training and validation sets as presented in Section III. Approximately 66 per cent of the data was used as training data and 34 per cent as validation data. Due to the privacy policy of the company that provided the data, all of the variable names are changed and values are normalised in this work.

B. Evaluation of results

The regression tree was applied to the preprocessed data and the model between the different operational conditions and the performance of the mobile work machines was created. Data preprocessing and analysis were performed with MATLAB software. Figure 4 presents the structure of the regression tree after the pruning procedure. The pruning of the tree was performed as presented in Section III. The original tree was constructed of 22,697 nodes, and then pruned to 41 nodes

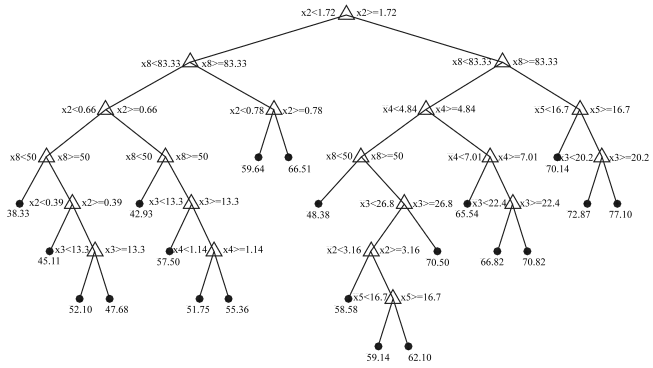

Figure 4. The regression tree constructed with CART.

– this was a compromise between prediction accuracy and complexity.

By looking at the structure of the regression tree, the most significant predictor variables in terms of modelling capability can be found in the upper nodes of the tree. In this work the variables $x2$, $x4$, $x5$, and $x8$ were identified to be the most important predictor variables. Additional predictor variables can be found in the lower nodes of the tree. The predictor variable significance in the regression tree structure was very well in line with the knowledge of the experienced mobile work machine operators. Also, the predictor variables with minor significance on the performance are not used for splitting in the pruned regression tree.

The prediction capability of the regression tree was first analysed by comparing how well the model is fitted to the training data. The resubstitution error of the tree is 19.83 and the 10-fold cross validation error of the tree is 16.29. Figure 5 presents the performance values of the mobile work machines in the training data and the predictions of the regression tree model. As Figure 5 illustrates, the predictions of the model and the original performance values correlate well on machines 1, 4, 6, 7, 8, and 9. The differences between the measurements and the predictions of the other machines are most likely caused by the pruning of the tree and the affects of non-operating condition related factors, such as the tuning of control parameters of the machines. Especially, the machine number 5 outperforms the typical performances of the machines mainly due to the advanced tuning of control parameters.

The model is then used to predict the reference performance values for the machines in the validation data. Figure 6 presents the performance values of the mobile work machines of the validation data and the predictions of the model. Based on the measurements of the operating conditions, the model predicts different performance values for the machines. These predictions are regarded as typical performance values for specific operating conditions. If the performance value of an individual mobile work machine is greater than the prediction, the machine has outperformed same types of machines working in the similar operating condition, and vice versa.

The following information can be observed from Figure 6: The measured performance values of the machines 12, 15, 16, and 17 are mostly similar to the predictions of the regression tree, which indicates average performance in given operating

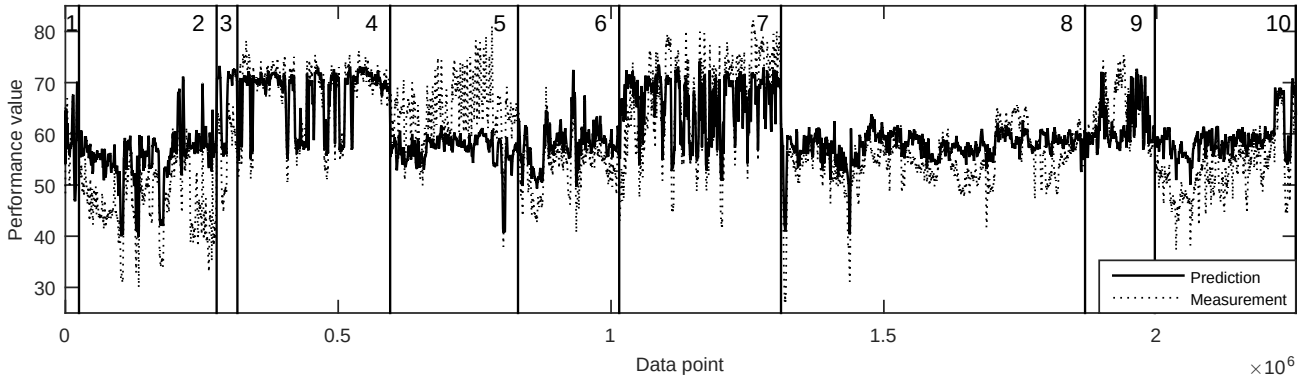

Figure 5. Performance predictions and the measured performance values of the training machines. Data is filtered for visualization.

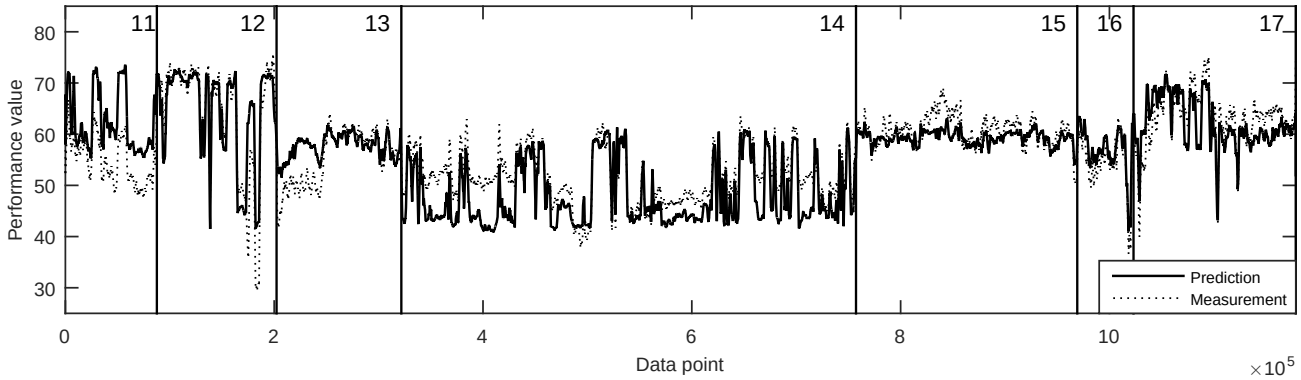

Figure 6. Performance predictions and the measured performance values of the validation machines. Data is filtered for visualization.

conditions. During the measurement periods, there are also better and worse performances compared to the predictions with the previously mentioned machines. These occasional variations can be considered normal, due to various factors, such as unmeasured variations in the operating conditions, temporary decrease in the technical condition of the machine, performance and skill level variations of the operators, etc.

As presented in the Figure 6, the machine number 11 has worse performance compared to the prediction during most of the measurement period. If comparison information about the performance would have been available for the operator, the declined performance could have been spotted and actions taken to improve the performance of the machine. Also the machine number 13 has at first worse performance than the predictions, but then due to actions taken by the operator, the machine reached an average performance in given conditions.

On the other hand, the machine number 14 has on average better performance compared to the references. However, the machine number 14 operates alternately in two different operating conditions. This can be noticed since the value of the performance prediction changes between two main levels. While operating in the environment that typically results in higher performance values, the measurement and the prediction are similar. However, while operating in the other operating condition, the measured performance is higher than the prediction. This is caused by the better control parameter selection.

The utilisation of regression tree enables more detailed analysis of the performance values. Figure 7 presents the performance distributions of the training data and machine number 13, for a specific operating condition. For demonstrative purposes, the selected operating condition is the one where the machine number 13 lacks performance compared to the training data. The distributions illustrate that the performance values of the machine number 13 are concentrated on the leftmost section of the original training data distribution.

Comparison information such as that presented in Figure 7 can be used to assist the machine operators to analyse the performance of the mobile work machine in different operating conditions. With the reference information about similar machines, the operator can analyse the performance of the machine with increased accuracy and reliability. There can be various reasons for the similarities and differences in the performance values between the machines. Depending on the work objectives of the machines, the differences can be explained with logical reasons, e.g. efficiency and productivity priorities set by the machine operators. However, the decreased performance is often a result of declined technical condition of the machine, low skill-level of the operator, or improper control parameter tuning.

 ISBN: 978-1-61208-423-7

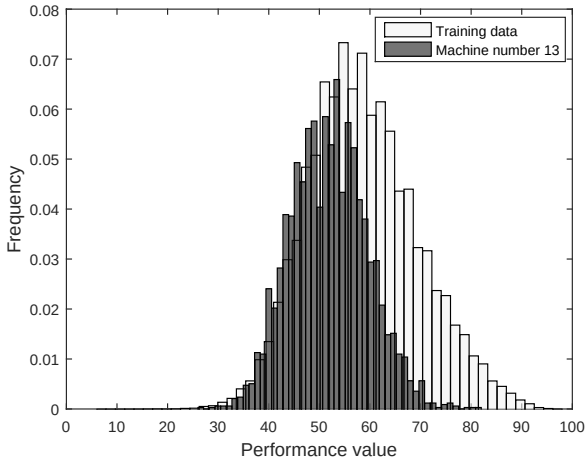

Figure 7. Performance distributions of the training data and the machine number 13, for a specific operating condition.

V. Conclusion and future work

The objective of this work was to research how to improve the performance analysis of mobile work machines. The most significant contribution of this study is the data-driven approach for analysing performances of mobile work machines. The presented approach is a combination of data preprocessing methods and the CART algorithm. The analysis of the performance is executed in three phases: modelling the relation between the operating conditions and performance values, predicting the typical performance values in specific operating conditions, and utilising the performance predictions as a reference values in the performance analysis of an individual mobile work machine.

The proposed method utilises a data-driven modelling approach. All of the operating conditions and performance values in the model are measured from machines working in various operating conditions. As more measurement data is collected, the regression tree can be updated to include new operating conditions and to increase the reliability of the performance predictions. One of the most important benefits of the presented method is the ability to model systems without extensive knowledge of the input-output variable relations in the data.

The results of this study indicate the potential of the presented method in the performance analysis of mobile work machines. However, further research is still required in data preprocessing, modelling, and analysis phases of the topic. Interesting topics related to data preprocessing are dimension and redundancy reduction. Further research topics concerning the modelling phase, include adding new input variables to regression tree modelling and testing new data analysis methods. The analysis phase is the most interesting part, since it enables the development of many practical applications designed for optimisation and root-cause analysis of the mobile work machine. The next step of the research is to increase the volume of the mobile work machine data and to evaluate the performance analysis in practice.

Acknowledgment

The research work was funded by Tekes (D2I – Data to Intelligence) and Academy of Finland (HOPE – Human Operator Modelling And Performance Evaluation In Human-machine Interaction) research programs.

References

[1] V. Hölttä, M. Repo, L. Palmroth, and A. Putkonen, "Index-based performance assessment and condition monitoring of a mobile working machine," in Proceedings of the 2005 ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Long Beach, California, USA: ASME, IEEE, 2006, pp. 615–622, the 2005 ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Long Beach, California, USA, September 24-28, 2005.

[2] V. Hölttä, "Plant performance evaluation in complex industrial applications," Ph.D. dissertation, Helsinki University of Technology, Espoo, Finland, 2009. [Online]. Available: http://lib.tkk.fi/Diss/2009/isbn9789522480927/isbn9789522480927.pdf

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks, 1984.

[4] L. Palmroth, K. Tervo, and A. Putkonen, "Intelligent coaching of mobile working machine operators," in Proceedings of the IEEE 13th International Conference on Intelligent Engineering Systems. Barbados: IEEE, 2009, pp. 149–154, iEEE 13th International Conference on Intelligent Engineering Systems 2009 - INES 2009, Barbados, April 16-18, 2009.

[5] R. H. Macmillan, "The mechanics of tractor-implement performance: theory and worked examples: a textbook for students and engineers," 2002, p. 166.

[6] T. Jokiniemi, H. Rossner, J. Ahokas et al., "Simple and cost effective method for fuel consumption measurements of agricultural machinery," in Agronomy Research, vol. 10, no. Special Issue I. Estonian Research Institute of Agriculture, 2012, pp. 97–107.

[7] S. Park, Y. Kim, D. Im, and C. Kim, "An assessment of eco driving system for agricultural tractor," Journal of Agricultural Science and Technology, 2011, pp. 906–912.

[8] F. Inns, Selection, Testing and Evaluation of Agricultural Machines and Equipment: Theory, ser. FAO agricultural services bulletin. Food and Agriculture Organization of the United Nations, 1995, no. 115.

[9] Z. Ismail and A. Abdel-Mageed, "Workability and machinery performance for wheat harvesting," Misr J. Ag. Eng, vol. 27, no. 1, 2010, pp. 90–103.

[10] J. Lu, Y. Liu, and X. Li, "The decision tree application in agricultural development," in Artificial Intelligence and Computational Intelligence. Springer, 2011, pp. 372–379.

[11] T. Waheed, R. Bonnell, S. O. Prasher, and E. Paulet, "Measuring performance in precision agriculture: CART – A decision tree approach," Agricultural water management, vol. 84, no. 1, 2006, pp. 173–185.

[12] M. Li, S. Feng, I. K. Sethi, J. Luciow, and K. Wagner, "Mining production data with neural network & CART," in Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003, pp. 731–734.

[13] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data preprocessing for supervised leaning," International Journal of Computer Science, vol. 1, no. 2, 2006, pp. 111–117.

 ISBN: 978-1-61208-423-7