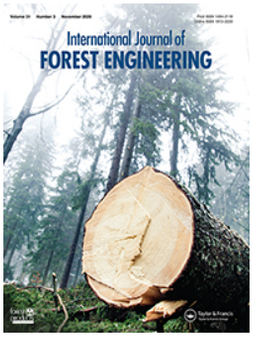

International Journal of Forest Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tife20

Modeling the productivity of mechanized CTL harvesting with statistical machine learning methods

Eero Liski , Pekka Jounela , Heikki Korpunen , Amanda Sosa , Ola Lindroos & Paula Jylhä

To cite this article: Eero Liski , Pekka Jounela , Heikki Korpunen , Amanda Sosa , Ola Lindroos & Paula Jylhä (2020) Modeling the productivity of mechanized CTL harvesting with statistical machine learning methods, International Journal of Forest Engineering, 31:3, 253-262, DOI: 10.1080/14942119.2020.1820750

To link to this article: https://doi.org/10.1080/14942119.2020.1820750

Published online: 19 Oct 2020.

INTERNATIONAL JOURNAL OF FOREST ENGINEERING
2020, VOL. 31, NO. 3, 253–262
https://doi.org/10.1080/14942119.2020.1820750

Taylor & Francis
Taylor & Francis Group

Modeling the productivity of mechanized CTL harvesting with statistical machine learning methods

Eero Liski[a], Pekka Jounela[a], Heikki Korpunen[b], Amanda Sosa[c], Ola Lindroos[d], and Paula Jylhä[b]

[a]Department of Natural Resources, Natural Resources Institute Finland, Helsinki, Finland; [b]Production Systems, Natural Resources Institute Finland, Helsinki, Finland; [c]Department of Science, Waterford Institute of Technology, Waterford, Ireland; [d]Department of Forest Biomaterials and Technology, Swedish University of Agricultural Sciences, Umeå, Sweden

ABSTRACT

Modern forest harvesters automatically collect large amounts of standardized work-related data. Statistical machine learning methods enable detailed analyses of large databases from wood harvesting operations. In the present study, gradient boosted machine (GBM), support vector machine (SVM) and ordinary least square (OLS) regression were implemented and compared in predicting the productivity of cut-to-length (CTL) harvesting based on operational monitoring files generated by the harvesters' on-board computers. The data consisted of 1,381 observations from 27 operators and 19 single-grip harvesters. Each tested method detected the mean stem volume as the most significant factor affecting productivity. Depending on the modeling approach, 33–59% of variation was due to the operators. The best GBM model was able to predict the productivity with 90.2% R^2, whereas OLS and the SVM machine reached R^2-values of 89.3% and 87% R^2, respectively. OLS regression still proved to be an effective method for predicting productivity of CTL harvesting with a limited number of observations and variables, but more powerful GBM and SVM show great potential as the amount of data increases along with the development of various big data applications.

ARTICLE HISTORY
Received 2 April 2020
Accepted 4 September 2020

KEYWORDS
Productivity; cut-to-length; harvester; machine learning; gradient boosted machine; support vector machine; regression model

Introduction

In Northern Europe, the mechanized cut-to-length (CTL) system dominates wood harvesting. In the CTL system, a harvester cuts and limbs the trees, crosscuts the stems into assortments and places the logs into piles to be picked up by a forwarder. In Finland, for instance, more than 99% of commercial roundwood harvesting is mechanized, and monthly numbers of harvesters and forwarders used in harvesting operations have varied between 3300 and 4500 machine units during the past two years (Natural Resources. . .2020).

Productivity is defined as the rate of product output per time unit for a given production system (e.g. Björheden and Thompson 1995). Practitioners use various models describing the productivity of harvesting in operational planning and setting work rates, for example. They are also used in the initial, theoretical phases of development and evaluation of new machines, work methods or entire production systems (e.g. Kärhä et al. 2004; Nuutinen 2013; Prinz et al. 2019).

Forests are a complex environment for production. The main factors affecting the productivity of mechanized cutting include environment (tree and terrain characteristics, climate), machine features (incl. bucking instructions), and the operator's mental and physical capacities and working technique (Ovaskainen 2009; Häggström and Lindroos 2016; Lindroos et al. 2017). Until recently, the availability of data has limited the construction of accurate productivity models. Manual data collection is expensive and normally limited to covering a limited number of potential influencing factors. Moreover, the tendency to change behavior when being monitored (the Hawthorne effect, Mayo 1933) may affect operators' performance (e.g. Lindroos 2010; Strandgard et al. 2013; Eriksson and Lindroos 2014; Manner 2015). The on-board computers of modern forest machinery continuously record large amounts of standardized data on machine functions, production, fuel consumption, etc. In most cases, the data is extracted and communicated using the StanForD protocol (Skogforsk 2019).

Automatically collected machine data have already been utilized in several studies about the productivity of CTL harvesting, often complemented with additional variables. The dataset of Eriksson and Lindroos (2014) covered more than 20 million cubic meters (under bark) of harvested wood, and they used 30 variables for predicting the productivity of harvesting with ordinary least square linear regression (OLS). Also, Purfürst and Erler (2011) and Gerasimov et al. (2012) applied OLS analyses to their machine data. Olivera et al. (2015) developed a linear mixed effect model for predicting the productivity of CTL harvesting, using the maximum likelihood method, while Rossit et al. (2019) applied the decision-tree technique. The accuracy of the data used in the studies above has varied from aggregated stand-level information to work elements of handling individual stems. Productivity models are typically constructed for delay-free time. Of the studies above, however, Purfürst and Erler (2011) constructed their models based on machine time also containing downtimes less than 15 minutes per occasion (E15 or G15).

Regression methods, in particular OLS, is a common procedure in forest works studies. In it, an equation represents the

CONTACT Eero Liski eero.liski@luke.fi Natural Resources Institute Finland, Helsinki, Finland

relationship between a response variable (in forest engineering typically time consumption or productivity) and one or more predictor variables (e.g. mean stem volume) (Bergstrand 1987; Magagnotti et al. 2012). Dummy variables are often used to include discrete factors, such as harvester's size class (e.g. Eriksson and Lindroos 2014). OLS provides an easy inference of the result, and it works well with a limited number of variables and when the underlying assumption of linearity holds at least reasonably well (Bishop 2006; Hastie et al. 2008). Linear regression is often made more flexible via basis expansion and the use of interaction terms. Deriving a good model with high-dimensional datasets and/or uncertain relationships between predictor and response variables is difficult with a parametric linear regression model, in which a somewhat strict linear functional form is assumed *a priori*. According to Costa et al. (2012), a partial least square regression (PLS) analysis allows the production of models that better fit the original data, allows handling collinear variables and facilitates the extraction of sound models from large amounts of field data from forest operations. This can lead to more robust models, but these regression analyses are not as easy to conduct, and they produce models that are less user-friendly than the OLS regression models, which can be expressed as equations to be used by practitioners for their specific conditions.

For parametric models, the relationships between predictors (variables, also known as features) and the response are often derived manually by sequentially testing parametric models' performance against data already used in optimizing their parameters. This easily leads to the overuse of data as several models are fitted to the whole data in order to find the best model. Multiple uses of the same data easily lead to overfitting, and an overly complex model does not predict well due to random noise. In the case of underfitting, a single overly simple model is fitted and treated as the best model (Figure 1, Bishop 2006; Hastie et al. 2008), but the underlying relationship between response and predictors remains unclear. The best model in Figure 1 would have the smallest error for new data, indicating the true relationship between the response and predicting variables.

Finding the best model in the sense of predictive ability requires specific tools, one should be able to assess the model's ability to predict previously unseen data and the model should be flexible enough to find the correct underlying model structure. With complex datasets containing numerous predictor variables, non-parametric machine learning methods usually out-perform their parametric counterparts. Mitchell (1997) describes machine learning as follows: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

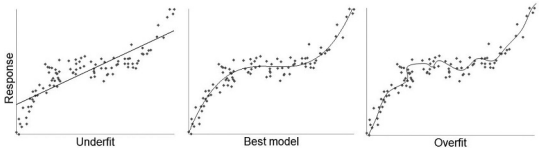

Figure 1. An illustration of the models' fit.

In the present study three widely used machine learning methods – Gradient Boosted Machine (GBM), Support Vector Machine (SVM) and OLS regression – were compared. The aim was to evaluate their potential for improving the accuracy of the prediction of productivity in mechanized CTL harvesting. Due to distinct responses and a predictive agenda, it is a question of supervised machine learning.

Materials and methods

In the present study, where the task (T) was to predict accurately the productivity of CTL harvesting, training data represented experience (E) and root mean square error (RMSE) was used as a performance indicator (P) for the models (cf. Mitchell 1997). The predictive performances of GBM, SVM and OLS regression were compared with identical datasets.

In order to avoid under- and overfitting, the data is usually divided into training, validation and test sets in machine learning (Mitchell 1997). In the present study, however, training and validation sets were combined, because cross-validation was applied. The training set was used to detect relationships between predictor and response variables. The validation set was used to find the optimal complexity parameters of the model, which mitigates, but does not necessarily prevent underfitting or overfitting of the optimally parameterized best model. The models' prediction accuracy with respect to previously unseen data was maximized by minimizing the test error (prediction error), which was chosen as the RMSE over independent test data. Even though RMSE was used as the minimizing criterion, the R^2-value was used as the final performance measure for each method, calculated from the test set.

Data composition

The data were provided by several private wood harvesting enterprises in Finland during 2014–2017 (Jylhä et al. 2019). Used as the response variable, the operator's productivity was calculated for each harvesting block by dividing the operator's total harvested volume (m^3, over bark) by delay-free productive work time (E_0h, (also known as PWH)) registered as processing in operational monitoring files (.drf, Skogforsk 2007, 2012) generated by the harvesters' onboard computers. The data for the computations were compiled using a Microsoft Access–based application developed for reading, transforming and exporting .drf files into Microsoft Excel format. The machine units (n = 19) used in data collection are described in Table 1.

The data originated from 577 thinning blocks and 509 regeneration felling blocks. One block can be composed of several compartments receiving the same cutting treatment. Thinnings from below were selective. Besides conventional clear-cuts, some seed-tree cuttings and strip harvesting sites aiming at natural regeneration were also included in regeneration fellings. In seed-tree cutting, 50–100 trees per hectare are left on site, while in strip harvesting, ca. 25 m wide zones are clear-cut, to be seeded by the edge forest (Äijälä et al. 2019; Finnish Forest Association 2020). The total number of individual operators involved in data collection was 49, and some of them operated more than one machine. However, the

Table 1. The description of the machinery used in data collection. Size classification of machinery according to Eriksson and Lindroos (2014).

Base machine		Harvester head			Number of operators	
Size class	Brand and model	Size class	Model	Number of machine units	Total	Individuals harvested ≥10 blocks
M	Komatsu 901.4	M	340	1	2	
L	John Deere 1170E	M	H413	4	7	
L	John Deere 1170E	M	H460	3	8	
L	John Deere 1170E	L	H414	2	7	
L	Komatsu 901TX.1	L	350	2	3	
XL	John Deere 1270E	L	H414	3	7	
XL	John Deere 1270E	XL	H415	1	6	
XL	John Deere 1270E	XXL	H480	3	10	
All pooled				19	49[1]	28

[1]Some operators used more than one machine.

operators with less than 10 observations were excluded from the computations, because preliminary analyses suggested that the SVM model does not perform well with only a few observations per operator. The remaining 27 operators produced 738 observations from thinnings and 643 from regeneration fellings and harvested in all 270,600 m^3 and 292,300 m^3 of industrial roundwood from these site types, respectively.

Operator-wise mean stem volumes (m^3) for removal, used as a predictive variable, were calculated for each block by dividing harvested volume by corresponding numbers of cut stems. Removal by each operator was categorized based on tree species composition (pine, spruce, birch, other broadleaved and mixed removal). The threshold for being classified as a main tree species was set at a minimum 60% (Kärhä 2007; Kärhä and Keskinen 2011) of operator-wise removal volume in each block; otherwise, the removals were considered mixed. In addition to the machine variables, quarter of year (based on the starting date of each block) and operator codes (IDs) were inserted in the data (Table 2).

Table 2. Variables (features) included in the data.

Variables	Definition	Unit/ Class	Mean	Min–max	SD	n
Y	Productivity of harvesting	m^3h^{-1}	23.2	4.2–69.8	11.7	1381
Th	Thinning	0/1	0.5343	0–1	-	738
RF	Regeneration felling	0/1	0.4656	0–1	-	643
V	Operator-wise mean stem volume	m^3	0.2741	0.033–1.267	0.2034	1381
Harvester size						
HS_1	M	0/1	0.0615	0–1	-	85
HS_2	L	0/1	0.6010	0–1	-	830
HS_3	XL	0/1	0.3374	0–1	-	466
Harvester head size (HH)						
HH_1	M	0/1	0.5394	0–1	-	745
HH_2	L	0/1	0.1904	0–1	-	263
HH_3	XL	0/1	0.1035	0–1	-	143
HH_4	XXL	0/1	0.1665	0–1	-	230
Main tree species in operator-wise removal (MTS)						
MTS_1	Pine	0/1	0.0738	0–1	-	102
MTS_2	Spruce	0/1	0.0260	0–1	-	36
MTS_3	Birch	0/1	0.0094	0–1	-	13
MTS_4	Other broadleaved	0/1	0.0195	0–1	-	27
MTS_5	Mixed removal	0/1	0.8711	0–1	-	1203
Quarter of year (Q)						
Q_1	1 (months 1–3)	0/1	0.2976	0–1	-	411
Q_2	2 (months 4–6)	0/1	0.1013	0–1	-	140
Q_3	3 (months 7–9)	0/1	0.2816	0–1	-	389
Q_4	4 (months 10–12)	0/1	0.3193	0–1	-	441

Modeling

The 1,381 observations described in Table 2 were randomly divided into two groups as follows: 80% for training/validation data (1,104 observations) and 20% for test data (277 observations). For each of the three methods, parameter optimization was performed using 10-fold cross-validation (Kohavi 1995), i.e. by partitioning the training/validation data into 10 disjoint non-overlapping subsets, and RMSE was used as the criterion for the best model. When the set of variables and parameters corresponding to the best model were identified, the best model was fitted to the whole training/validation data. Finally, the predictive ability was evaluated calculating the performance metric R^2-value for the test set.

In addition to the best models, respectively, found for GBM, SVM and OLS regression, two additional sets of models were constructed, resulting in the following three sets of models:

(i) The best models for GBM, SVM and OLS regression.
(ii) For each best model for GBM, SVM and OLS regression, operators were added as predictive variables. These models were fitted to the whole training/validation data and the R^2-values for the test set were calculated.
(iii) A predefined set of predictive variables were selected based on domain expert knowledge. GBM, SVM and OLS regression models with this predefined set of variables were fitted to the whole training/validation data and the R^2-values for the test set were calculated. These models with predefined variables can be seen as reference models, and their test set R^2-values can be compared to the the test set R^2-values of the best models (i) to investigate potential improvement of the machine learning approach.

The best models (i) deliver the main results in the present study. They were found by following a machine learning approach, in which the aim was to minimize the test error. This approach leads to sets of best predictive variables for GBM, SVM and OLS regression, respectively. The models constructed in (ii) and (iii) are based on (i). They do not perform any variable selection. Instead, the set of predictive variables in (ii) were found by simply adding operator variables to the best sets of predictive variables found in (i). The set of

predictive variables in (iii) were *chosen a priori* based on domain expert knowledge. Models in (ii) indicate possible improvement in predictive accuracy when operator information is added to the best models. Models in (iii) serve as reference.

Gradient boosted machine (GBM)

The gradient boosted machine (GBM, Friedman 2001; Malohlava and Candel 2017) model is an ensemble regression model that uses multiple decision trees to obtain better predictive performance than could be obtained from any of the constituent models alone. The GBM model gradually improves prediction performance (RMSE) by sequentially applying weak regression algorithms to the incrementally changed data which are boosted to help improve the statistical performance of the decision trees. Boosting filter observations by leaving those observations that the weak learner can handle and then focuses on developing new weak learner to handle remaining difficult observations. The GBM method generalizes tree boosting to increase computational speed. The result is an ensemble model that combines multiple hypotheses to form a (hopefully) better hypothesis, especially appropriate for mining less than clean data with outliers and potential correlated variables. Drop-out method (Vinayak and Gilad-Bachrach 2015) was used to avoid overlearning. The absolute (plus or minus) influence of each variable on harvesters' productivity was estimated using sensitivity analysis (Olden and Jackson 2002) with the best model. The GBM analyses were performed using RapidMiner software (https://rapidminer.com, version Studio Large 9.4.001, Mierswa et al. 2006).

Support vector machine (SVM)

Support vector machines (Vapnik 1995, 1998) are a group of supervised, semi-supervised and unsupervised machine learning methods used for classification, regression, clustering, anomaly detection and distribution estimation for complex data that is difficult to handle with linear functions. SVMs are apt for modeling very high-dimensional datasets (many columns), which may also contain a high number of observations (many rows, samples, also referred to as big data). The harvesters' productivity was predicted using Java version of mySVM with a dot (linear) kernel (Rüping 2000; Mierswa et al. 2006). This model type is based on the optimization algorithm of SVMlight described in Joachims (1999). The absolute (plus or minus) influence of each variable on harvesters' productivity was estimated using feature weights (Lagrange multipliers) of a linear kernel SVM model. SVMs require numerical variables, and therefore, discrete variables were dummy coded (Table 2). The variables were normalized using a zeroed mean with a variance of one, with the aim of avoiding bias caused by very high or very low values of some variables. The emphasis was put on the avoidance of over- and underfitting, by carefully optimizing the complexity parameter C (also called "capacity" and "regularisation" term). Too large C values can lead to overfitting and too small values to overgeneralization. The best model fit (minimum validation error) was sought by optimizing SVM parameters. In this context, the SVM parameters C and insensitivity (also called "slack") parameter ε were estimated using 10-fold cross-validation (Kohavi 1995) applied to sequential grid-search. The variables (feature space) were mostly dummy-coded (Table 2) and selected by expert judgment. The total number of variables (46 with operators, 19 without operators, 8 in the model) was very low compared to the capabilities of the SVM model, and therefore, the feature selection methods were not used. The SVM analyses were performed using RapidMiner software (https://rapidminer.com, version Studio Large 9.4.001, Mierswa et al. 2006).

Linear regression (OLS)

The general linear regression model can be written as follows (Eq. 1) (e.g. Searle 1971):

$$y = \beta' x + \varepsilon \tag{1}$$

where y, x, β and ε denote productivity of cutting, variables, model parameters and the error term, respectively. In addition to the variables shown in Table 2, the second-order polynomial term for mean stem volume and an interaction term between felling type and mean stem volume (with first- and second-order terms) were included in the potential variable set. The grid approach was applied to model selection, and each possible variable combination was evaluated in the cross-validation. The importance of each main-effect was investigated using the lmg method (Grömping 2006), which decomposes R^2 into variable contributions to be summed to the total R^2.

For the OLS regression, predicting variables were selected via grid approach as optimal parameter search. For each candidate model, the model parameters were OLS estimates, but the variable search in itself was an optimization process.

Optimal parameter, weight and variable selection for each model were reached by maximizing predictive accuracy. Reporting p-values for the OLS regression models were not relevant, because specific distributions for parameter samples were not assumed as in the case of hypotheses set in traditional OLS regression modeling. Instead, solely the unbiased OLS parameter estimates are reported and visualized using point estimates.

Since the operators were considered to represent a larger population of operators, they were treated as a sample from a general operator population to which inference can be applied. When modeling variation between operators, they were treated as a random term in the multilevel model (Pinheiro and Bates 2002). First, the multilevel modeling approach was warranted by testing whether there was enough random variation between operators. For that purpose, an intercept-only model was first fitted and compared with a random intercept-only model using log-likelihood. The simple multilevel model containing only operator variables was written as follows (Eq. 2) (e.g. Pinheiro and Bates 2002).

$$y_i = \beta + b_i + \varepsilon_i \tag{2}$$

where y_i is an n_i-dimensional response vector corresponding to operator i, β is a p-dimensional vector of fixed effects, $b_i \tilde{N}(0, \Psi)$ is the q-dimensional vector of random effects, and $\varepsilon_i \tilde{N}(0, \sigma^2)$ is the n_i-dimensional within-group error vector.

The process of model-building was continued by considering a random intercept term and the set of potential variables, and also a second-order polynomial term for mean stem volume was added to a set of potential predictor variables. One variable at a time was added to the model, and all variables decreasing the AIC criterion (Akaike 1974) were included. The new model obtained in this way was treated as the current best model, after which one variable at a time was again added to the model, minimizing the AIC. When the AIC did not decrease anymore, the process was stopped. A multilevel model with both fixed and random terms was written as follows (Eq. 3):

$$y_i = X_i\beta + b_i + \varepsilon_i, \tag{3}$$

where X_i denotes the fixed effects and the other notation similar to model M2. The overall variance of the predicted productivity was described using the Nakagawa R^2 coefficient of determination (Nakagawa and Schielzeth 2013).

All OLS computations were done using the R software (R Core Team 2018) with packages ggplot2 (Wickham 2016), nlme (Pinheiro et al. 2018), MuMIn (Barton 2018), tidyverse (Wickham et al. 2019), dummies (Brown 2012) and relaimpo (Grömping 2006).

Results

GBM and SVM

Both GBM and SVM models indicated that mean stem volume (V) was by far the most important factor affecting the productivity of CTL harvesting (Figure 2). Also, the type of cutting (regeneration felling (RF) vs. thinning (Th)) played a decisive role, i.e. productivity in regeneration fellings was higher than in thinnings all other factors being equal. The cutting of pine-dominated stands (MTS_1) was more efficient than those with another tree species composition. The use of the M-sized harvester (HS_1), and, on the other hand, the XL-sized harvester head (HH_3) reduced productivity. In both analyses, the use of the XL-sized harvester head had a large negative impact on productivity.

Even though most factors showed similar behavior in the GBM and SVM methods, there were some distinct differences in terms of how levels of factors influenced productivity. Large and XXL-sized harvester heads (HH_2 and HH_4), for instance, had only minor influences in GBM, but a rather large positive influence in SVM. Similarly, the year's last quarter (Q_4) had a large positive effect in GBM, but a small negative effect in SVM.

Based on the GBM model, operators explained on average 48% of predicted productivity. The SVM model showed lower operator effect as operators explained on average 42% of productivity. When the GBM and SVM models were run with operator information, the impacts of the other variables were, to a great extent, in the same order as in those without operator information (Figure 3).

Linear regression (OLS)

The best model (Eq. 4) produced by OLS regression consists of the terms felling type, the size of harvester and harvester head, main tree species, quarter of year, mean stem volume, mean stem volume second-order polynomial, interaction between felling type and mean stem volume and interaction between felling type and mean stem volume second-order polynomial (Eq. 4):

$$\begin{aligned} y = {} & \beta_0 + \beta_1 RF + \beta_2 HS_2 + \beta_3 HS_3 + \beta_4 HH_2 + \beta_5 HH_3 \\ & + \beta_6 HH_4 + \beta_7 MTS_2 + \beta_8 MTS_3 + \beta_9 MTS_4 + \beta_{10} MTS_5 \\ & + \beta_{11} Q_2 + \beta_{12} Q_3 + \beta_{13} Q_4 + \beta_{14} V + \beta_{15} V^2 + \beta_{16} RF * V \\ & + \beta_{17} RF * V^2 + \varepsilon, \end{aligned} \tag{4}$$

The model parameter estimates are shown in Table 4. Mean stem volume (V, m^3) had the largest impact on productivity, and the type of felling had the second largest impact (Figure 4). Based on the model (Table 3), the large harvester head size (HH_2) had a slightly higher productivity compared to the medium-sized harvester head size (HH_1), whereas the XXL harvester head size had a lower productivity compared to medium-sized harvester head.

Figure 5 shows the predicted productivity for the best linear regression model for a fixed variable set. In general, the productivity in thinnings was lower than in regeneration fellings. However, this specific model setting did not show a significant difference in productivity between thinnings and regeneration fellings when the mean stem volume was between 0.272 m^3 and 0.461 m^3. Moreover, the thinning productivity peaked around an approximate stem volume of 0.750 m^3, after which it started to decline.

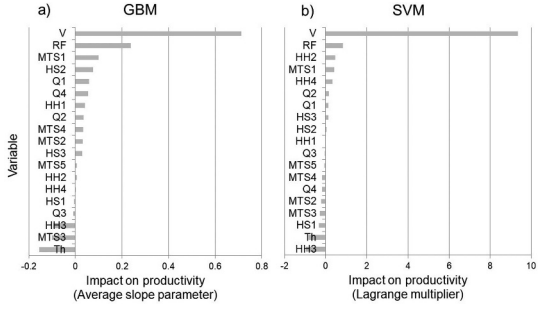

Figure 2. The impact of non-anthropogenic variables (operator variables in the model) on productivity based on the a) GBM and b) SVM models. Note the different units in the x-axes; slope parameter in GBM = slope parameter; Lagrange multiplier in SVM.

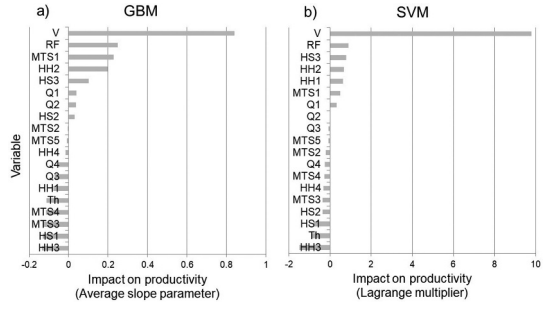

Figure 3. The impact of non-anthropogenic variables (operator variables excluded from the model) on productivity based on the a) GBM and b) SVM models. Note the different units in the x-axes; slope parameter in GBM = slope parameter; Lagrange multiplier in SVM.

Table 3. The parameter estimates of the best linear regression model.

Parameter	Estimate	RMSE
Full model	-	4.57
Intercept[1]	5.188	
RF	5.445	
HH_2	0.261	
HH_3	-6.168	
HH_4	-2.241	
HS_2	1.775	
HS_3	4.504	
MTS_2	-3.574	
MTS_3	-5.159	
MTS_4	-3.496	
MTS_5	-2.224	
Q_2	-0.424	
Q_3	-0.948	
Q_4	-1.215	
V	89.39	
V^2	-58.62	
RF×V	-24.12	
RF× V^2	36.65	

[1]The intercept includes the baseline for each categorical variable, in which the following factors were assumed: Thinning, Harvester head size M, Harvester size M, Main tree species pine and Quarter 1.

Table 4. The parameter estimates of the reference linear regression model. Abbreviations are presented in Table 2.

Parameter	Estimate	RMSE
Full model	-	4.86
Intercept	8.451	
RF	2.682	
MTS_2	-4.062	
MTS_3	-6.828	
MTS_4	-3.901	
MTS_5	-2.737	
V	71.87	
V^2	-31.35	

[1]The intercept includes the baseline for each categorical variable, in which Thinning and Main tree species pine are assumed.

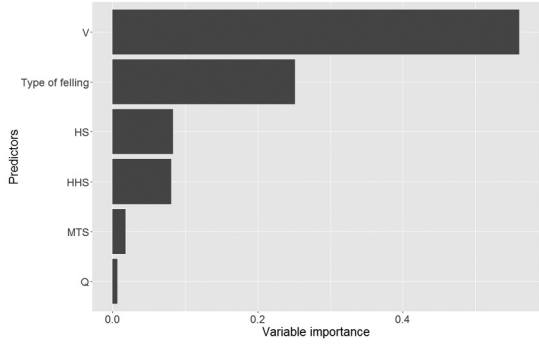

Figure 4. Relative variable importance (percent of the total R^2) of each main-effect term for the best linear regression model. The y-axis notation is the following: HS = harvester's size class, HHS = size class of harvester head, MTS = main tree species, Q = quarter of year.

The reference model can be written as follows (Eq. 5):

$$y = \beta_0 + \beta_1 RF + \beta_2 MTS_2 + \beta_3 MTS_3 + \beta_4 MTS_4 + \beta_5 MTS_5 + \beta_6 V + \beta_7 V^2 + \varepsilon \quad (5)$$

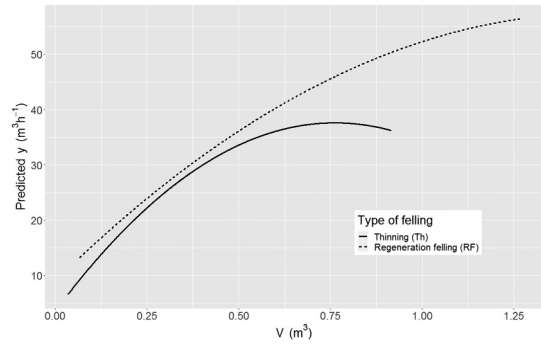

Figure 5. An example of predicted productivity (of y) against the mean stem volume (V) by felling type for linear regression, with fixed values MTS_5 (Mixed species removal), HS_2 (Harvester size class L), HH_1 (Harvester head size M) and 4th quarter of the year (months 10–12).

The parameter estimates for the reference model are presented in Table 4.

The log-likelihood test between the intercept-only model and the random intercept-only model resulted in a small p-value (<0.05), indicating that there was enough random variation between operators for multilevel modeling. Nakagawa R^2-value of 33.0% indicated that operator information explained about one-third of productivity variation when being used as the sole predictor.

The final multilevel model included the following variables: type of felling, mean stem volume (V), second-order polynomial term for mean stem volume (V^2), size of harvester head, quarter of year and main tree species. For the final model, Nakagawa R^2 values of 80.9% for fixed and 88.8% for both fixed and random terms were obtained. Consequently, when fixed variables were already included in the model, the operator variable increased the model's predictive performance by 7.9 percentage points.

Comparison of modeling approaches

The best variable subsets based on linear regression, linear kernel SVM regression and GBM models with their hold-out set performances are shown in Table 5.

For GBM, SVM and OLS regression, the best model with operators gave the highest test set R^2 value, i.e. the highest predictive accuracy (see Table 5). This is not surprising, since more variable information is used in prediction. The reference models gave the lowest predictive accuracy. When applying the models with operator information to new data, the best model with operators must always include the same operators included in the present dataset. The best models without operators and the reference models do not have this restriction.

Table 5. The best level of prediction (R^2) for the best OLS regression (OLS R), linear kernel SVM regression and GBM models over the three model levels.

		The best model	
Method	Reference model	Without operator	With operator
GBM	83.7	85.6	90.2
SVM	80.9	82.9	87.0
OLS Regression	82.2	84.6	89.3

Discussion

Automatically recorded forest machine data are considered to be a reliable source of information for predicting the productivity of CTL harvesting (e.g. Nuutinen 2013; Brewer et al. 2018). The predictive performance of the machine learning models of the present study was high compared to linear regression models constructed using much larger datasets based on forest machine data (e.g. Purfürst and Erler 2011; Eriksson and Lindroos 2014). In the present study, the level of prediction (R^2) varied from 81% up to 90%. Rossit et al. (2019) reached accuracies greater than 90% with the decision-tree technique applied to stem data. However, there are uncertainties associated with these comparisons due to variation in methods, metrics and data quality. The present approach of maximizing the predictive accuracy is, however, quite strict, because the overfitting of models is unlikely due to the partitioning of data and cross-validation.

As shown by traditional regression models (e.g. Eliasson 1998; Kärhä et al. 2004; Nurminen et al. 2006; Purfürst and Erler 2011), stem size and operator were important factors affecting the productivity of CTL harvesting. Also, the decision-tree method applied by Rossit et al. (2019) indicated that tree size is the most significant factor predicting the productivity of CTL harvesting, but tree species and operator also show significant influences. A larger number of tree species combinations would theoretically result in larger predictability, but this would require more data. Based on the studies of Kärhä (2007) and Kärhä and Keskinen (2011), we ended up using a somewhat rough mixed tree species class for stand-level data. Nevertheless, the tree species distribution was comparable with the tree species distribution of the total roundwood removals in Finland between the years 2014 and 2017 (Natural Resources…2020). The results are also parallel with a Swedish follow-up study (Eriksson and Lindroos 2014) based on millions of cubic meters of harvested wood.

Predictor variable representativeness is connected to the model's ability to generalize to the population. However, evaluating predictor variable representativeness is difficult and cannot be ensured due to lack of detailed statistics. Large discrepancies of relative frequencies between data and population might cause prediction bias. Assuming the data and population have similar predictor variable distributions, our models are likely to give generalizable results.

The operator's influence is a complex phenomenon to analyze. Besides the numerous factors affecting actual performance and their interaction (e.g. Purfürst and Erler 2011), the result is also dependent on the modeling approach. In the present study, the operator alone explained ca. 30–60% of productivity variation of CTL harvesting, and the operator's effect was greater in regeneration fellings than in thinnings. Also, earlier studies have indicated that stand conditions affect the operator effect – the more difficult the conditions, the greater the human factor will have on productivity (Väätäinen et al. 2005; Kariniemi 2006; Purfürst and Erler 2011). Kariniemi (2006) found out that the differences between operators increased along with an increase in stem volume, which may explain the differences in the operator effect in thinnings and regeneration fellings. In addition, in the present data, the average number of assortments was higher in regeneration fellings (Jylhä et al. 2019). Previous studies (e.g. by Nurminen et al. 2006; Eriksson and Lindroos 2014) have indicated that increasing the number of timber assortments directly increases the time consumption of harvesting. Also, the present study confirms the conclusion of Purfürst and Erler (2011) about the need to include the operator in productivity models. Väätäinen et al. (2005) have estimated that 10–15% of the differences in the performance of harvester operators are due to differences in work technique, 20–30% due to better crane and generator control and the remaining 50–55% result from competence in planning and decision-making.

The present study indicated that matching the size of the base machine and harvester head with stem volume is worth considering when balancing the efficiency and unit cost (per cubic meter) of harvesting. From the modeling viewpoint, the combination of harvester head size and the size of the base machine could theoretically have been taken into consideration, for example, by including an interaction term between these variables. However, most of the combination classes in the present dataset had zero frequency, and therefore, it was not possible to include such an interaction term in the models. In practice, all combinations of harvester head size and the size of the base machine are not technically or economically feasible. As pointed out by Eriksson and Lindroos (2014) only a part of the variables used in this study can be considered as good models of the underlying factors, and most of them are rather indications of areas where there is a need to develop improved productivity models. Such areas include machine, type of operation, stand complexity and environmental factors.

Also, Purfürst and Erler (2011) used operational monitoring data (.drf files). They emphasized that their parameter estimates may contain unknown errors, and they recommend the use of data related to single trees rather than the whole stand (or larger units as in the case of the present study). In general, the results of various studies are only valid in conditions similar to those under which the data were generated. In the present study, information related to environmental conditions was scarce as only a few additional variables were inserted in the .drf data. Furthermore, the variables were aggregated at the operator and block (harvesting unit) level. The variation in model parameters within one block can be large, as one block can be composed of one or more compartments harvested during several shifts. However, data storage and transmission capacities can limit the use of more detailed data.

In the present study, only the time registered as delay-free time (E0) registered as processing time was considered, while also delays shorter than 15 minutes were included in the data of Purfürst and Erler (2011). Inclusion of delays exceeding the main filtering time likely further increased the heterogeneity of their data. The data of the present study were also used in the follow-up study of Jylhä et al. (2019), in which E0 processing time constituted ca. 82% of E15 time and 86% of production time. When comparing the models based on .drf files to those based on manual timing, one should note that E0 times can include short downtimes below the minimum filtering time (StanForD default value 15 s, Skogforsk 2019). Such short delays are excluded from manual timing to a great extent.

The main filtering time (default value 120 s) defines how main work phases based on CAN-bus data (processing and terrain travel) are registered in .drf files. Therefore, processing sessions normally also include moving between processing points. These phases were combined in the data of Eriksson and Lindroos (2014), which was collected using vibration sensors installed on the machines. Judging from equal productivity levels, the timing principles are comparable.

The GBM had the best predictive performance of the three compared methods. This is logical as the model derives its predictive power from an ensemble of multiple, overlapping regions of variable values, without assuming any *a priori* specified value distribution. A general reason for using ensemble models (also random forests, Breiman 2001) is to reduce uncertainty and stability on predictions when compared with a single model, such as a generalized linear model or a generalized additive model. Dietterich (2000) considers statistical, computational and representational factors as the reasons for the high predictive power of ensemble models. Firstly, when the training set is small (as in the reference model setup with only three variables), a learning algorithm can typically find several models (functions) in the hypothesis space and, by averaging several models, ensemble models may reduce the risk of choosing the wrong hypothesis. Secondly, an ensemble of individual models built from many different starting points may provide a better approximation of the true unknown function than one using any of the single models. By combining several models in an ensemble, it is possible to expand the space of representable functions and obtain a better model of the true function. The shortcoming of GBM is that it may be computationally "expensive" to derive the contribution of variables (plus-minus importance) with very high-dimensional datasets. Due to the overwhelming number of variable combinations to be examined (Olden and Jackson 2002), it is common to vary each feature from its minimum to maximum value while keeping all other variables constant at a certain summary measure (e.g. mean, min and max).

The Linear SVM model expects equal importance for correlated variables, and it does not suffer from the "winner takes all" phenomenon typical of parametric regression models. In a linear kernel SVM, a highly correlated variable does not heavily affect the weights of other variables, which results in smooth and balanced weights (plus-minus importance). Linear SVM weights are widely used in prediction competitions for weighting variables with high-dimensional datasets (Chang and Lin 2008; Guyon et al. 2008), for example, for selecting the most important variables (feature selection; Guyon and Elisseeff 2003) or for deriving inference of all predicting variables without selection. The model does not require the derivation of interaction terms for variables. However, the dataset used in the present study was small and hence the linear SVM did not perform as well as the two other model types based on test set performance. With a small dataset with only a few variables, intuitively the winner-takes-all approach or ensemble models are more powerful.

Data from harvesting operations are often incomplete (e.g. Purfürst and Erler 2011; Eriksson and Lindroos 2014), which limits the utilization of productivity models. The behavior of machine learning models with incomplete data and missing values depends on the model type. For example, the GBM model tolerates missing data, but SVM and the OLS regression require imputation or the omission of missing values or variables. Prominent examples of missing data imputation methods are pattern removal, conditional mean/mode approach and k-nearest neighbor, whose k-values can be optimized using n-fold cross-validation. The dataset used in the present study did not have variables with missing values.

The practical application of the results of the present study is limited by the fact that GBM and SVM do not produce parametric equations. The OLS regression models' equations can be utilized using non-specialized software (e.g. MS Excel). For GBM and SVM, one option for practical applicability would be to serve the GBM and SVM models in the cloud, where anyone could apply them to their own dataset. However, applying all the models would require data with the same distribution as in the dataset used in the present study. Another solution to apply to GBM and SVM would be to follow the procedures described in this article and apply them to another dataset. For this, however, specific software is needed (e.g. R, R Core Team 2018; RapidMiner, Mierswa et al. 2006; Knime, Python and Weka, Eibe et al. 2016). They all include specialized packages (extensions) for various tasks. They can be used interactively; for example, RapidMiner has extensions for R, Python and Weka. So a fraction of a process can be run concurrently in another software using the same (one) script of the parent software. Currently available machine learning software has more or less the same mainstream functionalities with minor differences in the implementation of algorithms.

In the deployment phase, a machine learning model is typically saved in a repository to make predictions with streaming (continuously updated and probably bigger) data. It is also common to apply a set of competing machine learning models, built on the same or similar data sets, where one model is active, and the remainder are challengers. The performance of the models may change over time, alerting of drift and bias, because the input (training/validation) data were not representative or because the new streaming data drifted. If the measured drift is significant, one may want to rebuild (re-parameterize) models. In addition, deployment models are usually shared by a group collaborating on a common project with web services, so that it can be integrated with other software.

Conclusions

Any machine learning method is restricted by the quantity and quality of available data. The OLS regression implemented using machine learning performed well when compared with the more flexible GBM and SVM approaches. Considering the data properties (number of observations and available variables/features), the underlying phenomenon most likely follows rather closely a parametric form captured by a linear regression model with a simple basis expansion – or at least a more flexible representation cannot predict much more accurately with the present dataset. The potential of SVM and GBM is well known and their ability to outperform OLS regression would increase as the number

of variables and observations increase. Implementing machine learning methods shows great potential in forest engineering as various applications for collecting and analyzing real "big data" are under development (Tech4Effect 2019; Koneyrittäjien Datapankki 2019). However, the questions related to data ownership and regulation of data protection (The European Parliament and the Council 2016) limit the utilization of big data.

In the future, the amount of harvesting data is likely to increase greatly. Existing and new machine learning methods will be useful tools for finding nuanced relationships between variables and performing more accurate predictions. For applicability, models should be served on a website or in a cloud platform. Data pipelines would feed data to a model that would continuously serve predictions. The private sector already builds these systems, and the academic world should follow suit.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Northern Periphery and Arctic Programme 2014–2020; Natural Resources Institute Finland; The Ministry of Employment and the Economy.

ORCID

Amanda Sosa http://orcid.org/0000-0001-9887-3974
Ola Lindroos http://orcid.org/0000-0002-7112-4460

References

Äijälä O, Koistinen A, Sved J, Vanhatalo K, Väisänen P, editors. 2019. Metsänhoidon suositukset. Tapion julkaisuja [Best practices for sustainable forest management]. Helsinki. p. 252. Finnish.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Automat Contr. 19(6):716–723. doi:10.1109/TAC.1974.1100705.

Barton K. 2018. MuMIn: multi-model inference. R package version 1.42.1. https://CRAN.R-project.org/package=MuMIn.

Bergstrand K. 1987. Planning and analysis of time studies on forest technology. The Forest Operations Institute of Sweden. Report 17. p. 58.

Bishop CM. 2006. Pattern recognition and machine learning. New York: Springer; p. 315, 520.

Björheden R, Thompson MA. 1995. An international nomenclature for forest work study. In: Field DB, editor. Proceedings, IUFRO 1995 S3:04 subject area: 20th World Congress; Tampere, Finland, August 6–12; University of Maine. p. 190–215.

Breiman L. 2001. Random forests. Mach Learn. 45(1):5–32. doi:10.1023/A:1010933404324.

Brewer J, Talbot B, Belbo H, Ackerman P, Ackerman S. 2018. A comparison of two methods of data collection for modelling productivity of harvesters: manual time study and follow-up study using on-board-computer stem records. Ann For Res. 61(1):109–124.

Brown C. 2012. Package 'dummies': create dummy/indicator variables flexibly and efficiently. p. 8. https://cran.r-project.org/web/packages/dummies/dummies.pdf.

Chang YW, Lin C-J. 2008. Feature ranking using linear SVM. J Mach Learn Res. 3:53–64.

Costa C, Menesatti P, Spinelli R. 2012. Performance modelling in forest operations through partial least square regression. Silva Fennica. 46 (2):241–252. doi:10.14214/sf.57.

Dietterich TG. 2000. Ensemble methods in machine learning. In: Multiple classifier systems. Berlin: Springer; p. 1–15.

Eibe F, Hall MA, Witten IH, Pal JC. The WEKA workbench. In: Pitts T, editor. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Cambridge, Massachusetts USA: Morgan Kaufmann Publishers; 2016

Eliasson L. 1998. Analyses of single-grip harvester productivity [doctor's dissertation]. Acta Universitates Agriculturea Sueciae, Silvestria 80. Umeå: Swedish University of Agricultural Sciences. p. 24.

Eriksson M, Lindroos O. 2014. Productivity of harvesters and forwarders in CTL operations in northern Sweden based on large follow-up datasets. Int J For Eng. 25(3):179–200. doi:10.1080/14942119.2014.974309.

Finnish Forest Association. 2020. Not just clearcutting – finnish state forests are harvested by twelve different methods. [accessed 2020 May 14]. https://forest.fi/article/not-just-clearcutting-finnish-state-forests-are-harvested-by-twelve-different-methods/#3140b969.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. Ann Stat. 29(5):1189–1232. doi:10.1214/aos/1013203451.

Gerasimov Y, Senkin V, Väätäinen K. 2012. Productivity of single-grip harvesters in clear-cutting operations in the northern European part of Russia. Eur J For Res. 131:647–654. doi:10.1007/s10342-011-0538-9.

Grömping U. 2006. Relative importance for linear regression in R: the package relaimpo. J Stat Softw. 17(1):1–27. doi:10.18637/jss.v017.i01.

Guyon I, Aliferis C, Cooper G, Elisseeff A, Pellet J-P, Spirtes P, Statnikov A. 2008. Design and analysis of the causation and prediction challenge. J Mach Learn Res. 3:1–33.

Guyon I, Elisseeff A. 2003. An introduction to variable and feature selection. J Mach Learn Res. 3:1157–1182. doi:10.5555/944919.944968.

Häggström C, Lindroos O. 2016. Human, technology, organization and environment - a human factors perspective on performance in forest harvesting. Int J For Eng. 27(2):67–78.

Hastie T, Tibshirani R, Friedman J. 2008. The elements of statistical learning: data mining, inference and prediction. New York, Springer.

Joachims T. Making Large-scale SVM Learning Practical. Advances in Kernel Methods—Support Vector Learning, Scholkopf B, Burges C, and Smola A (ed), 169-184, Cambridge MA: MIT Press, 1999.

Jylhä P, Jounela P, Korpunen H, Koistinen M. 2019. Koneellinen hakkuu. Seurantatutkimus. [Mechanised cutting. Follow-up study]. Luonnovara- ja biotalouden tutkimus 11/2019. Natural Resources Institute Finland; p. 53. Finnish.

Kärhä K. 2007. Ensiharvennusten korjuuolot vuosina 2000–2006 [Harvesting conditions of first thinnings in 2000–2006]. Metsätehon Tuloskalvosarja 17/2007. Finnish.

Kärhä K, Keskinen S. 2011. Ensiharvennukset metsäteollisuuden raaka-ainelähteenä 2000-luvulla [First thinnings as a raw material source of Finnish forest industries in the 21st century]. Metsätehon Tuloskalvosarja 2/2011. Finnish.

Kärhä K, Rönkkö E, Gumse S-I. 2004. Productivity and cutting costs of thinning harvesters. Int J For Eng. 15(2):43–56. doi:10.1080/14942119.2004.10702496.

Kariniemi A. 2006. Kuljettajakeskeinen hakkuukonetyön malli – työn suorituksen kognitiivinen tarkastelu. [Operator-specific model for mechanical harvesting - cognitive approach to work performance]. Helsingin yliopiston Metsävarojen käytön laitoksen julkaisuja; Helsinki, Finland. p. 38. Finnish.

Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI'95 Proceedings of the 14th international joint conference on artificial intelligence; San Francisco (CA): Morgan Kaufmann Publishers. p. 1137–1143.

Koneyrittäjien Datapankki. 2019. [accessed 2019 Aug 22]. https://www.koneyrittajat.fi/pages/etusivu/toiminta/koneyrittaejae-tuotteet/datapankki/datapankin-esittely.php.

Lindroos O. 2010. Scrutinizing the theory of comparative time studies with operator as a block effect. Int J For Eng. 21(1):20–30.

Lindroos O, Häggström C, La Hera P. 2017. Drivers of advances in mechanized timber harvesting – a selective review of technological innovation. Croatian J For Eng. 38(2):243–258.
Magagnotti N, Spinelli R, Acuna M, Bigot M, Guerra S, Hartsough B, Kanzian C, Kärhä K, Lindroos O, Roux S, et al. 2012. Good practice guidelines for biomass production studies. Sesto Fiorentino (Italy): CNR IVALSA; p. 50. ISBN 978-88-901660-4-4.
Malohlava M, Candel A. 2017. Gradient boosting machine with H2O. H20 Booklet. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/GBMBooklet.pdf.
Manner J. 2015. Automatic and experimental methods to studying forwarding work [doctoral thesis]. Acta Universitatis agriculturae Sueciae No. 2015:128. Faculty of Forest Sciences, Swedish University of Agricultural Sciences. p. 70.
Mayo E. 1933. The human problems of an industrial civilization. New York: Macmillan Co; p. 194.
Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. 2006. Yale: rapid prototyping for complex data mining tasks. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '06); New York (NY): ACM. p 935–940. https://rapidminer.com.
Mitchell TM. 1997. Machine learning. McGraw-Hill Science/Engineering/Match; New York, p. 432 p.
Nakagawa S, Schielzeth H. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution. 4(2):133-142. doi: 10.1111/j.2041-210x.2012.00261.x.
Natural Resources Institute Finland. 2020. Statistics database. [accessed 2020 Feb 19]. https://statdb.luke.fi/PXWeb/pxweb/en/LUKE/?rxid=6bdecc27-bb61-47a7-91a6-37394b371061.
Nurminen T, Korpunen H, Uusitalo J. 2006. Time consumption analysis of the mechanized cut-to-length harvesting system. Silva Fennica. 40 (2):335–363. doi:10.14214/sf.346.
Nuutinen Y. 2013. Possibilities to use automatic and manual timing in time studies on harvester operations. Dissertationes Forestales. 156:68. doi:10.14214/df.156.
Olden JD, Jackson DA. 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecol Modell. 154(1–2):135–150. doi:10.1016/S0304-3800(02)00064-9.
Olivera A, Visser R, Acuna M, Morgenroth J. 2015. Automatic GNSS-enabled harvester data collection as a tool to evaluate factors affecting harvester productivity in a Eucalyptus spp. harvesting operation in Uruguay. Int J For Eng. 27(1):15–28. doi:10.1080/14942119.2015.1099775.
Ovaskainen H. 2009. Timber harvester operator's working technique in first thinning and the importance of cognitive abilities on work productivity. Dissertationes Forestales. 79:62. https://dissertationesforestales.fi/pdf/article1862.pdf.
Parliament and Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 26 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Off J Eur Union L. 119:88. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.
Pinheiro JC, Bates DM. 2002. Mixed-effects models in S and S-plus, Corr. 3. print. edn. New York: Springer-Verlag.
Pinheiro JC, Bates DM, DebRoy S, Sarkar D. 2018. R Core Team. _nlme: linear and nonlinear mixed effects models_. R package version 3.1-137. https://cran.r-project.org/package=nlme.
Prinz R, Väätäinen K, Laitila J, Sikanen L, Asikainen A. 2019. Analysis of energy efficiency of forest chip supply systems using discrete-event simu lation. Appl Energy. 235:1369–1380. doi:10.1016/j.apenergy.2018.11.053.
Purfürst T, Erler J. 2011. The human influence on productivity in harvester operations. Int J For Engineering. 22(2):15–22. doi:10.1080/14942119.2011.10702606.
R Core Team. 2018. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. https://www.R-project.org/.
Rossit DA, Olivera A, Viana Céspedes V, Bro D. 2019. A big data approach to forestry harvesting productivity. Comput Electron Agric. 161:29–52. doi:10.1016/j.compag.2019.02.029.
Rüping S. 2000. mySVM-manual. Universität Dortmund, Lehrstuhl Informatik VIII. http://wwwai.cs.uni-dortmund.de/SOFTWARE/MYSVM/.
Searle SR. 1971. Linear models. New York: John Wiley & Sons.
Skogforsk. 2007. Standard for forest data and communication. http://www.skogforsk.se/contentassets/b063db555a664ff8b515ce121f4a42d1/stanford_maindoc_070327.pdf.
Skogforsk. 2012. StanForD variables in numerical order. 106 s. https://www.skogforsk.se/contentassets/b063db555a664ff8b515ce121f4a42d1/allvarno_eng_120418.pdf.
Skogforsk. 2019. [accessed 2019 Aug 22]. https://www.skogforsk.se/english/projects/stanford/.
Strandgard M, Walsh D, Acuna M. 2013. Estimating harvester productivity in Pinus Radiata plantations using StanForD stem files. Scand J For Res. 28(1):78–80. doi:10.1080/02827581.2012.706633.
Tech4Effect. 2019. Efficiency portal. [accessed 2019 Aug 22]. http://www.tech4effect.eu/efficiency-portal/.
Väätäinen K, Ovaskainen H, Ranta T, Ala-Fossi A. 2005. Hakkuukoneenkuljettajan hiljaisen tiedon merkitys työtulokseen työpistetasolla. [The effect of harvester operator's tacit knowledge on work result at the level of processing location]. Metsäntutkimuslaitoksen Tiedonantoja. 937:100. Finnish.
Vapnik V. 1995. The nature of statistical learning theory. New York: Springer-Verlag.
Vapnik V. 1998. Statistical learning theory. New York: Wiley.
Vinayak R, Gilad-Bachrach R. 2015. DART: dropouts meet multiple additive regression trees. In: Lebanon G, Vishwanathan SVN, editors. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS'15). San Diego, California, USA. p. 489–497.
Wickham H, Averick M, Bryan J, Chang W, D'Agostino McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Lin Pedersen T, Miller E, Milton Bache S, Müller K, Ooms J, Robinson D, Paige Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the the tidyverse. J Open Source Softw. 4 (43):1686. doi:10.21105/joss.01686.
Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag.