European Journal of Forest Research (2020) 139:1029–1043
https://doi.org/10.1007/s10342-020-01304-5

ORIGINAL PAPER

Separating the impact of work environment and machine operation on harvester performance

Lari Melander[1] · **Risto Ritala**[1]

Received: 27 February 2020 / Revised: 15 June 2020 / Accepted: 26 June 2020 / Published online: 4 July 2020

Abstract

In mechanized logging operations, interactions between the forest machines and their operators, forest resources and environmental conditions are multifold and not easily detected. However, increased computational resources and sensing capabilities of the forest machines together with extensive forest inventory data enable modeling of such relationships, leading eventually to better planning of the operations, better assistance for the forest machine operators, and increased efficiency of timber harvesting. In this study, both forest machine fieldbus data and forest inventory data were acquired extensively. The forest inventory data, acquired nationwide, was clustered to categorize general tree and soil types in Finland. The found forest categories were applied when the harvester fieldbus data, collected from the forest operations in the North Karelia region with two similar harvesters, was analyzed. When the performance of the machine and the operator, namely the fuel consumption and log production, is studied individually for each forest cluster, the impact of working environment no longer masks the causes based on the machine or the operator, thus making the observations from separate forest locations comparable. The study observed statistically significant differences in fuel consumption between the most general tree and soil clusters as well as between the harvester-operator units. The modeling approach applied, based on multivariate linear regression, finds such reasons for the differences that have clear interpretation from machine setup or operator working style perspective, and thus offers a feasible method for assisting the operators in improving their working practices and thus the overall performance specifically at forest of given type.

Keywords Forestry · Data fusion · Machine learning · Forest data · Fieldbus data · Harvester · Performance

Introduction

Forests resources are being digitalized throughout the world. Remote sensing in its many forms (see, e.g., Holopainen et al. 2014; Dash et al. 2016; White et al. 2016; Talbot et al. 2017) has been widely applied to provide tree and topographic data of forests, enabling better planning of forest operations. This is often referred to as precision forestry or Industry 4.0 in wood supply (Holopainen et al. 2014; Mason et al. 2016; Müller et al. 2019). Related to this trend, forest inventory data is collected worldwide, in particular in Europe, Canada, USA, Russia, Brazil, China and New Zealand (Tomppo et al. 2010). Furthermore, at least in the Nordic countries, the effort is to make forest data public (Kangas et al. 2018). In Finland, most of the forest inventory data collected with public resources have been made publicly available, and the latest effort is to gather all forest related data sources accessible via a single service (Venäläinen et al. 2015; Hämäläinen 2016; Rajala and Ritala 2016). The key aspect in this service is to fuse the heterogeneous data into constant grid cells (16 m × 16 m in Finland). Openly accessible and aggregated forest inventory data is now enabling all stakeholders to develop new applications for supporting forest operations. For example, forest machine manufacturers are collecting a vast amount of forest data with their machines but are currently not utilizing existing forest data when developing new products or optimizing the current machines.

Cut-to-length (CTL) forest machines dominate the market in the Nordic countries, as almost all the logging

Communicated by Eric R. Labelle.

✉ Lari Melander
lari.melander@tuni.fi

[1] Automation Technology and Mechanical Engineering, Tampere University, Korkeakoulunkatu 10, 33720 Tampere, Finland

is performed with CTL systems in Finland, Sweden and Norway (Lundbäck et al. 2018). This is due to a long tradition in cross-cutting the stems already in the forest for easier transportation (Gellerstedt and Dahlin 1999). The numerous demands set for the timber harvesting in the CTL system, such as the ability to respond to precise cutting specifications of sawmills and to work with minimal forest floor impact, have developed CTL forest machines to intelligent systems capable for extensive sensing of the forest environment and information processing (Lindroos et al. 2015; Olivera and Visser 2016). Efficient timber harvesting necessitates careful planning of the forest operations, as appropriate machines should be chosen to do the correct work at the right time. Forest inventory data, as depicted above, is being widely used for better planning and scheduling of the operations in this general level. However, from the perspective of a single forest machine carrying out a specific operation the forest environment is dynamic: while the machine travels through the intended harvesting route, the environmental conditions and forest types are changing, affecting both the forest machine's actual and maximum achievable performance (Suvinen and Saarilahti 2006; Ala-Ilomäki et al. 2012; Obi and Visser 2017; Melander et al. 2019). Olivera et al. (2016) have pointed out that the type of forest, i.e., the properties of the trees, affects the productivity. Therefore, the operator has to decide the actual route of the harvester, taking into account, for example, the bearing capacity of the forest floor at the given season, while simultaneously keeping in mind the correct harvesting density and the resulting width of the logging road. Due to complex dynamical relationships of the environment with respect to the machine and goals of the operation, an operator can hardly have exhaustive understanding of the optimal actions in the ongoing forestry operation. Today, CTL forest machines assist the operator in many tasks, for example by optimizing the cross-cutting points for each tree, but mostly the operator relies to his own experience and skills while working (Häggström and Lindroos 2016). Furthermore, the operator has a possibility to adjust forest machine settings, which have a considerable effect on the performance of the machine (Prinz et al. 2018). For helping the operator in these many adjustments, the impact of the environment to the optimal machine settings needs to be understood. Such understanding can only be developed by analyzing forest machines in a variety of environments. However, this requires that the effects of the environment and the effects of the machine operation can be separated. The future of the forest machines is foreseen to be increasingly autonomous (Hellström et al. 2009; Ringdahl 2011; Ringdahl et al. 2011), removing the variation caused by the operator, but the requirement for separating the effects of the forest machine and the environmental factors continues to be highly relevant.

Until now, research on Big Data solutions suitable to forest operations has resulted in a rather limited number of publications, as recently pointed out by Rossit et al. (2019). The existing research is mostly concentrated on modeling the processed trees in the forest operations (see, e.g., Lu et al. 2018; Shan et al. 2019) or evaluating the productivity from the production records (see, e.g., Olivera et al. 2016; Eriksson and Lindroos 2017; Rossit et al. 2019). However, analysis of the interactions between the forest machine and its environment necessitates large amounts of data collected from the machines, in particular, fieldbus data in addition to the production records. Data collection of this extent inevitably creates challenges for the data warehousing and communication capabilities while working in remote locations. The solution lies in the machine learning and data mining algorithms, which detect patterns and structures in large data sets (Murphy 2012). Most of these algorithms can be divided into categories of predictive (supervised) and descriptive (unsupervised) learning. In predictive learning, a regression or classification model is constructed between known inputs and outputs available in the dataset. Descriptive learning is used for revealing unknown relationships and structures without any prior knowledge of the data, and it has typically applications in clustering and dimensionality reduction. Use of such algorithms for data in forestry has been depicted, for example, by Rossit et al. (2019) and Melander et al. (2019). By exploiting pre-trained machine learning models and the computational power of the forest machine, it is feasible to analyze most of the data while at the harvesting site, reducing the need for vast data transmission between the forest machine and data warehouses.

The current paper is based on the idea and the early results in Melander et al. (2019), where forest inventory data and machine fieldbus signals were fused for revealing machine–environment relationships. This paper concentrates on explaining performance differences of two harvesters, similar with each other, found in a long-term collection of fieldbus data. This paper expands our earlier work, firstly in that it analyzes the performance both while the machine is cutting and in motion, secondly in that it shows how to separate the performance differences due to the forest type and the operation of the machine, and thirdly that the forest and soil categorization is based on nationwide analysis of forest inventory data. The main contribution of this paper is the methodological basis for fleet-wide analysis of the performance of the forest machine and its operator in relation to the working environment.

Materials and methods

This study analyzes two datasets: firstly, fieldbus and production data collected from two similar harvesters in a recording of 20 working days, and secondly a set of forest inventory data systematically sampled from the database covering whole Finland. In addition, detailed forest inventory data with whole areal coverage was acquired for the areas where the machines were working, allowing fusion of machine and environmental data. Therefore, forest inventory data is used at two levels: the sampled inventory data from the database covering whole Finland for learning generalizable categories of Finnish forests, and the non-sampled forest inventory data for each grid cell where the machines have visited for comparing machine performances individually for each forest type. As the two forest machines under consideration were similar, the differences in performance due to operator actions are highlighted in this study. However, some of the parameters in the control system of the forest machine are user-specific and were set differently in the two harvesters, so the comparison in this paper is actually done between the two machine-operator combinations.

Harvester data collection

Fieldbus data was collected from two Ponsse Scorpion harvesters with the same age and with similar equipment, including, for example, the harvester head. Ponsse Scorpion is an 8-wheeled CTL harvester with 210 kW diesel engine, weighing approximately 21 tonnes. Further details on the machine can be found from the datasheet of the machine (Ponsse Plc 2020). The collected data consists of signals from various actuators of the harvester, including signals from the harvester head and boom, transmission system, steering system and the GNSS device. In total there were 48 signals recorded constantly from the fieldbus of a single harvester and eight variables related to the production output. Table 1 shows an overview of the types of the collected signals. The sampling interval of the fieldbus signals was 0.02 s, meaning that in a single 8-h working day approximately 1.5 million rows of fieldbus data was recorded for a single variable (roughly 1 GB of fieldbus data per day for one harvester).

The recording period was from late September to early October 2019, when the outside temperature in the daytime was between 0 and 10 degrees and there were occasional subzero temperatures in the night-time. The soil was free from frost and the mean daily precipitation during the period was 2.9 mm. Data was recorded during harvesters' normal timber harvesting routines, and the data collection was running in the background. No specific operation tests were arranged. Harvesters were operated in one shift, meaning that both harvesters were each operated by a single operator through the whole data collection period. The machine-operator units are referred to as Operator 1 and Operator 2 from now on. The operators had their own custom harvester control system settings, which they kept mostly constant during the data collection period. Both operators had regeneration felling and thinning tasks during the recording period. The datasets were labeled according to these work types. The work sites of both machines were at North Karelia region in Eastern Finland, specifically around municipalities of Kitee and Rääkkylä (Fig. 1).

Finnish forest data for forest categorization

The forest data platform allows user-defined queries with no limitations on the amount of the data or the number of retrieved grid cells. In our earlier study (Melander et al. 2019), the data for representing the Finnish forests was delimited to ten small areas consisting of 100,000 grid cells, and to nine forest parameters. Here, the intention is to take advantage of the limitless data queries of the forest data platform and construct a dataset that genuinely represents Finnish forests, thus enabling search of the underlying structure, i.e., the forest categories of Finland, with unsupervised machine learning methods. However, the total number of grid cells in Finland is of the order of 10^9, so retrieving and analyzing all the grid cells and all the forest parameters would require considerable computational resources. Therefore, a systematic sampling pattern was designed to

Table 1 Overview of harvester signals

Boom and harvester head	Transmission and motor	Orientation and position	Production
Boom rotation control	Fuel consumption	GNSS receiver variables	Number of produced logs and felling cuts
Boom 1st joint control	Speed	Acceleration *X*, *Y*, *Z*	Tree species and assortment
Boom 2nd joint control	Diesel engine RPM and torque	Longitudinal and lateral tilt	Log length
Boom extension control	Hydrostat RPM	Steering control	Log diameter (butt, average, top)
Harvester head rotator control	Hydraulic motor control		Log volume
Saw control	Cooling unit control		
Feed control forward/backward	Working brake		

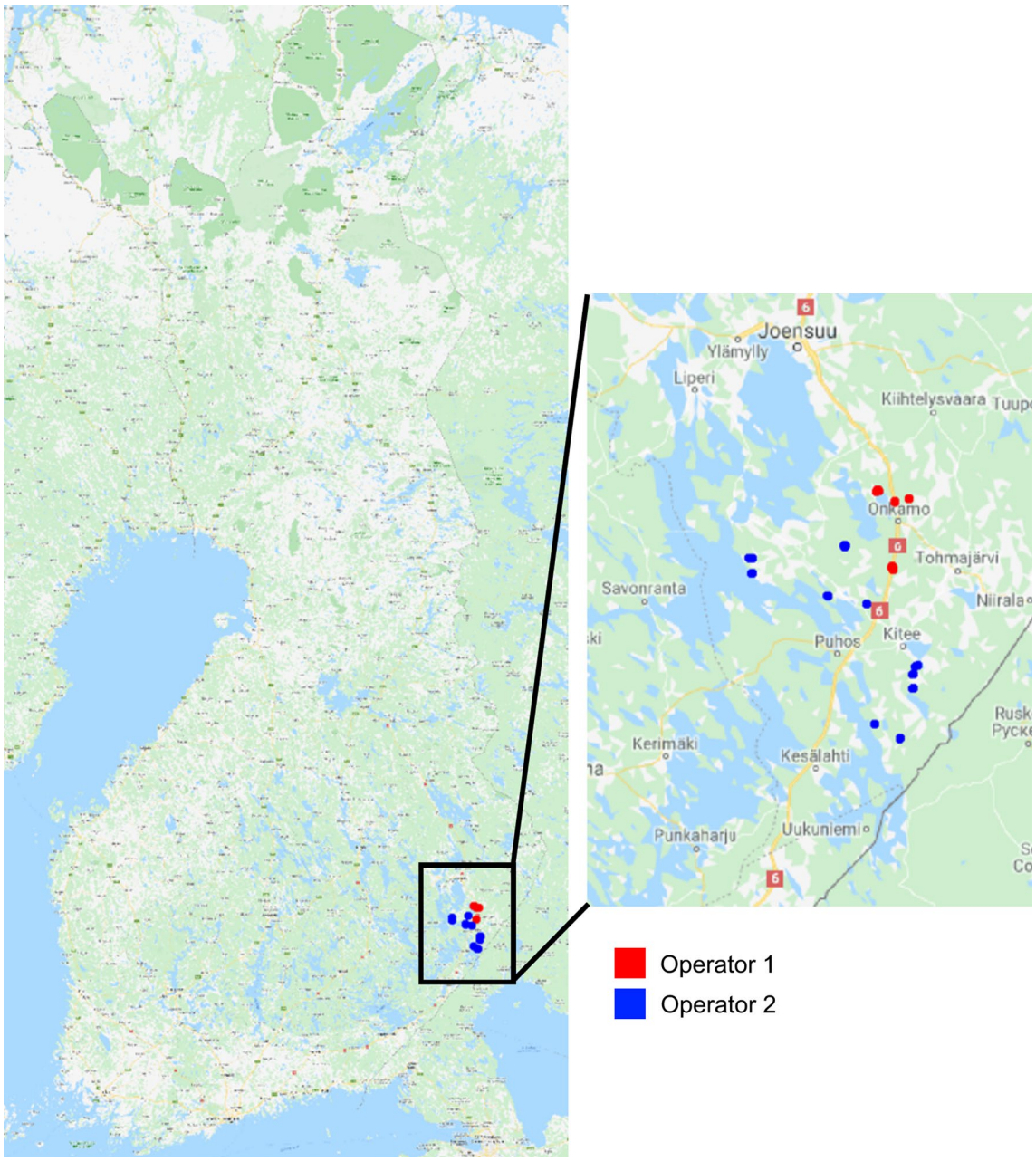

Fig. 1 Recorded forest operation sites

retrieve a representative set of Finnish forest inventory data (Fig. 2a). The sampling pattern consists of 176 square areas of size 12 km × 12 km, containing overall approximately 10^7 grid cells.

The platform responds to the query by returning the grid data partitioned to map sheets (size 24 km × 48 km). This data (156 files) was sampled further, so that 10% of grid cells were picked randomly from each map sheet. After sampling, the data consisted of approximately 2.9 million grid cells with 80 continuous and 13 categorical forest variables for each cell. Data was preprocessed to remove any inconsistencies in data (Fig. 2b). All cells containing false or missing values were removed and only cells for which land type was indicated to be forest, were selected. In this process, it was noticed that two of the major data sources having similar forest inventory data, i.e., the forest inventory data maintained by Finnish Forest Center (FFC) (Finnish Forest Center 2019) and the National Forest Inventory (NFI) maintained by Natural Resources Institute Finland (Luke 2019), conflicted at many grid cells. The reason for the conflicts may be due to different inventory instants: if, for example, a forest area is harvested between the inventory

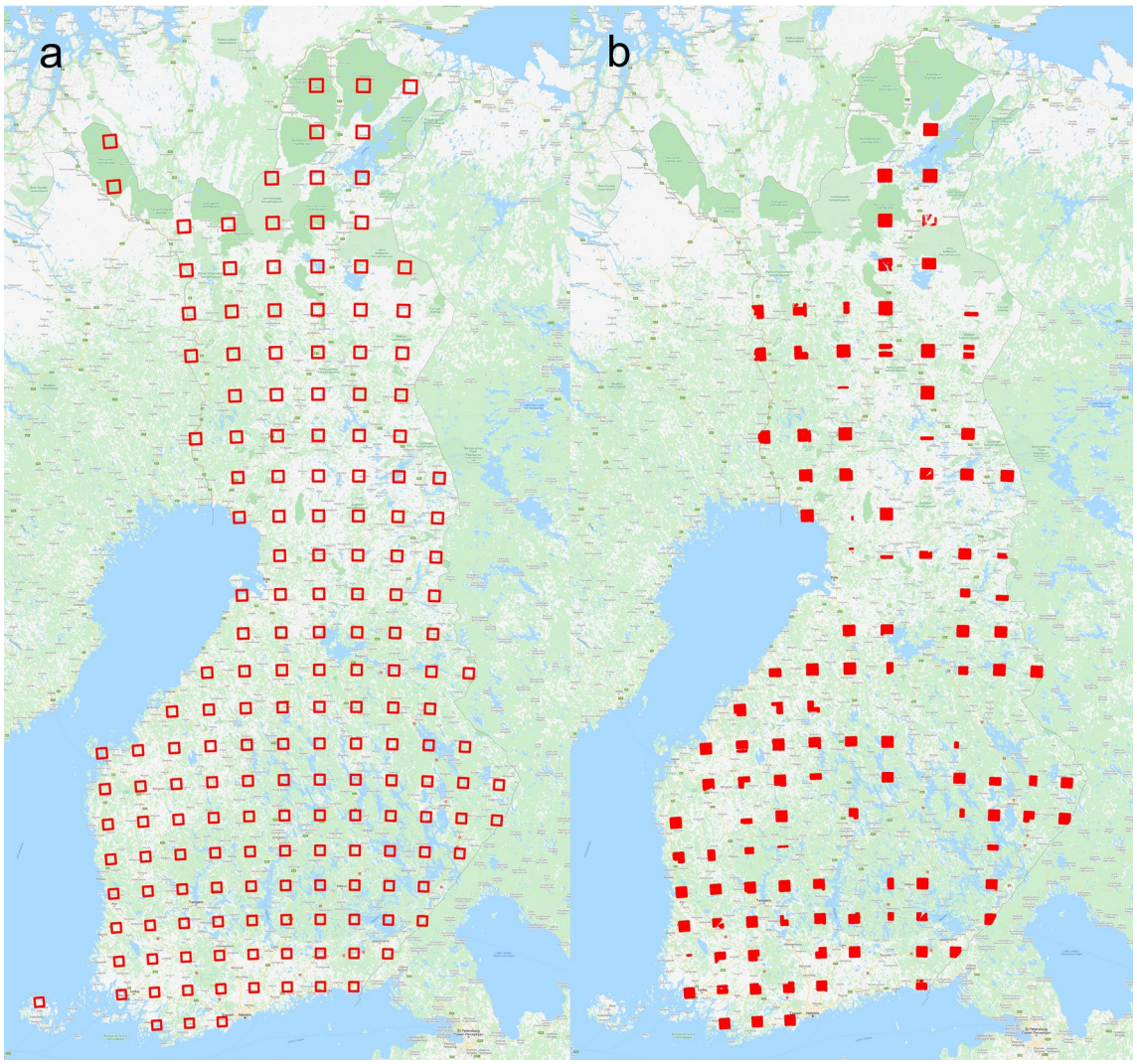

Fig. 2 **a** Constant sampling grid for the forest inventory data retrieval (captured from the forest data platform user interface). **b** Realized samples after data cleaning process

instants, rather different tree dimensions and density are to be expected. Because of the conflicts, only the FFC data was retained in the dataset. After preprocessing, the data comprises approximately 1.5 million grid cells with 28 continuous forest variables. The categorical variables were reduced to four: soil type, harvesting accessibility, fertility class and drainage state. Other categorical variables were rather constant in the final set of grid cells. The final set of variables is collected and given in Table 2.

Fusion of machine and forest data at worksites

The fusion of the forest and machine data closely follows the procedure presented by Melander et al. (2019). The forest inventory data for the data fusion is retrieved for each grid cell along the machine route, and should not be confused with the Finnish forest data collection for forest categorization, presented in the previous section. The difference between the two is that the large-scale sampling presented in the previous section is needed for learning the underlying structure of Finnish forest inventory data by clustering, so that the local forest inventory data on the machine route on every forest operation can then be set in proportion to all other Finnish forests. The fieldbus time series and forest inventory data on grid cells of the machine route are fused according to the position given by the GNSS of the machine. This associates time series of varying lengths to grid cells, according to the period the machine spends in a grid cell. Each repeated visit to a grid cell—if any—associates its own time series to the cell. These grid-positioned time series are further divided according to the working mode of the harvester: driving and processing trees. The division is

Table 2 Forest variables included in the analysis

Variable name	Explanation
Tree-related variables	
Tree age	Mean age of the trees in the grid cell. Includes separate variables for pine, spruce and deciduous
Tree mean diameter	Mean diameter of the trees in the grid cell. Includes separate variables for pine, spruce and deciduous
Tree mean height	Mean height of the trees in the grid cell. Includes separate variables for pine, spruce and deciduous
Tree basal area	Total basal area of the trees in the grid cell. Includes separate variables for pine, spruce and deciduous
Stem count	Stem count in the grid cell, given in stem count per hectare. Includes separate variables for pine, spruce and deciduous
Tree volume	Tree volume in the grid cell, given in cubic meter per hectare. Includes separate variables for pine, spruce and deciduous
Laser height	85% point in the cumulative height distribution of laser observations over two meters in the grid cell
Laser density	Number of laser observations above 2 m in the grid cell divided by the number of all the observations
Soil-related variables	
Topographic wetness index (TWI)	In this study, the continuous-valued TWI is transformed into a categorical variable by dividing its range to 16 equally wide bins
Soil type	Soil type according to the Finnish soil classification standard
Harvest accessibility	Accessibility rating from 1 (always accessible) to 6 (only on wintertime). See Kankare et al. (2019) for detailed classification information
Fertility class	Fertility class describes undergrowth vegetation which is seen to reflect fertility and productivity status of the site. The classification is based on the work of Cajander (1909, 1949) and is widely used in Finland
Drainage state	Drainage state describes whether the area is ditched and the current state of the soil drying

necessary because it is expected that quite different variables are relevant in the two working modes, e.g., forest ground and topography are expected to be important for harvester transmission and orientation, but tree properties are expected to be important for production and boom operation. This division shortens the time series considerably, as the working mode changes frequently inside a grid cell. For gaining statistical robustness, the time series are sampled using a window as many times as is possible without overlapping the windows. The length of the windows was chosen to be 10 s for the moving harvester and 30 s for the working harvester.

In this study, the performance of the harvester-operator combination is evaluated by fuel consumption and total volume of log production per time instance. Both performance indicators are examined together with the forest inventory data, revealing the effects of the forest parameters to the performance.

Forest clustering

For gaining generalizable results in every forest operation, the dimensionality of the Finland-wide forest inventory data was reduced and then clustered. The dimensionality of the continuous forest inventory data, consisting of 26 tree parameters, was reduced with principal component analysis (PCA). PCA necessitates selecting the number of the resulting variables, i.e., the number of principal components, for the model, and this number was chosen to be nine in this study as this preserves 90% of the original variation in the forest inventory data. The nine axes were further clustered with K-means algorithm (Jain 2010) for general tree types of Finland, referred to as tree clusters later in the study. K-means produces a pre-specified number of clusters, but at present there is no prior information about what is the appropriate number of clusters for the Finnish forest inventory data. Therefore, the clustering results, the sum of squared distances and a silhouette score, were evaluated as a function of number of clusters. The silhouette score (Rousseeuw 1987) is a general measure of the quality of the clustering, ranging from − 1 to 1, with higher numeric values indicating better clustering. With the resulting cluster model, each grid cell on the route of the harvesters was labeled by the cluster index based on the forest inventory data in the cells. Categorical forest inventory data variables, describing the forest ground properties, were clustered with K-modes algorithm (Chaturvedi et al. 2001), that is similar to the combination of PCA and K-means for continuous variables. Resulting clusters are referred to as soil clusters later in the study.

Inferring about performance differences

When studying the effect of the soil and tree clusters to harvester performance, it was assumed that the soil conditions affect the most the moving harvester and correspondingly the tree clusters affect the felling operation. The effect of the clusters was examined by grouping the windowed fieldbus signals, such as the fuel consumption, according to those clusters and then identifying the most significant statistical

differences of the signal means between the groups. However, it is highly likely that there are other factors besides environmental ones affecting the performance. The differences caused by the operator actions and machine settings can be examined by studying machine signal features within a single soil or tree cluster so that the harvesters have been operating at similar environment, and then to repeat such analysis to each soil or tree cluster to find cross effects. One approach is to fit a predictive model for finding the function between variables of interest, such as operator control variables, and a performance metric, such as fuel consumption, and study the relative importance of the independent variables given by the model. In this study, linear regression models were fitted for predicting the consumption of the harvester within the time window of samples. 80% of the total dataset (3398 window samples for the driving motion and 5134 samples for the felling work) was used for the fitting and the rest was left for testing the model performance (850 and 1284 samples, respectively). The independent variables for the models were chosen to be either forest data variables or machine variables that are directly controllable by the operator. For example, diesel engine RPM level is a variable set by the operator and thus suited as an independent variable. However, sets of variables that are highly correlated should be avoided in the regression model, and therefore most of the forest variables were not included as independent variables. Furthermore, production volume correlates strongly with the feed control of the tree through the harvester head, so it was not taken into the model. Table 3 presents an overview of the independent variables of the consumption models.

As all independent variables were standardized to zero mean and unit variance, the most important variables with respect to the target variable (fuel consumption) are found simply by examining the coefficients for the independent variables in the model. By weighting the differences in the independent variables (forest data or operator controls) with their coefficients, reasons for the differences can be inferred. For example, the cause for a higher consumption for one machine-operator unit over another can be reasoned by the distinctive usage of the harvester boom control, if the difference in the boom controls, multiplied with the corresponding coefficient, is high.

Table 3 Independent variables selected for consumption models

Working harvester	Driving harvester
Variables in consumption models	
Diesel engine RPM	Diesel engine RPM
Boom rotation control	Speed
Boom 1st joint control	Harvester steering control
Boom 2nd joint control	Inclination (front-rear)
Boom extension control	
Number of tree cuts	
Harvester head feed control (forward and backward)	
Harvester head rotation control	
Tree volume in the grid cell	
Stem count in the grid cell	

Results

General forest clusters

The representative sample of the forest inventory data in the national level was processed separately into most typical tree and soil clusters, to support separate analysis for moving harvester (with soil clustering) and for working harvester (with tree clustering). Figure 3 shows principal component loadings of original tree-related forest inventory variables in the PCA transformation. Figure 4 shows the loadings of variables related to tree species.

To find the best number of clusters, the sum of squared distances between data points and cluster centers and silhouette scores are presented in Fig. 5. The sum of squared distances (Fig. 5a) decreases rather steadily as the number of clusters increases. The silhouette score (Fig. 5b) shows some variation as a function of the number of clusters but is rather low and constant. The K-modes clustering of the categorical soil variables shows similar results (Fig. 6), with the exception that the smaller values for the number of clusters seem to result for better clustering according to the silhouette score. In both cases, the low silhouette score values indicate that there are no clear clusters in the dataset.

Another way to assess the usefulness of the general forest clusters is to apply the clustering to the grid cells of the field tests and examine the variation of the forest inventory data inside every cluster, as the clustering is supposed to minimize this variation inside the clusters. Based on the evaluation shown in Fig. 5 and Fig. 6, number of clusters was chosen to be 45 in the tree clustering and 7 in the soil clustering. With this choice, the five most common clusters in the field tests are next described in detail. The performance of the machine will be compared individually within each of these five clusters. All the cluster centers for the soil clusters are presented in Table 4, with clusters 1, 2, 4, 6 and 7 being the most common. Correspondingly, the five most common tree type clusters in the field tests are described in Table 5 by means of the original forest inventory data inside the cluster.

Figure 7 demonstrates the consistency of the forest inventory data inside clusters by showing the distribution of forest inventory variables inside clustered grid cells along the route of the two harvesters. Two forest parameters, stem count and tree volume, are presented for the five most common clusters in the field tests for both harvesters.

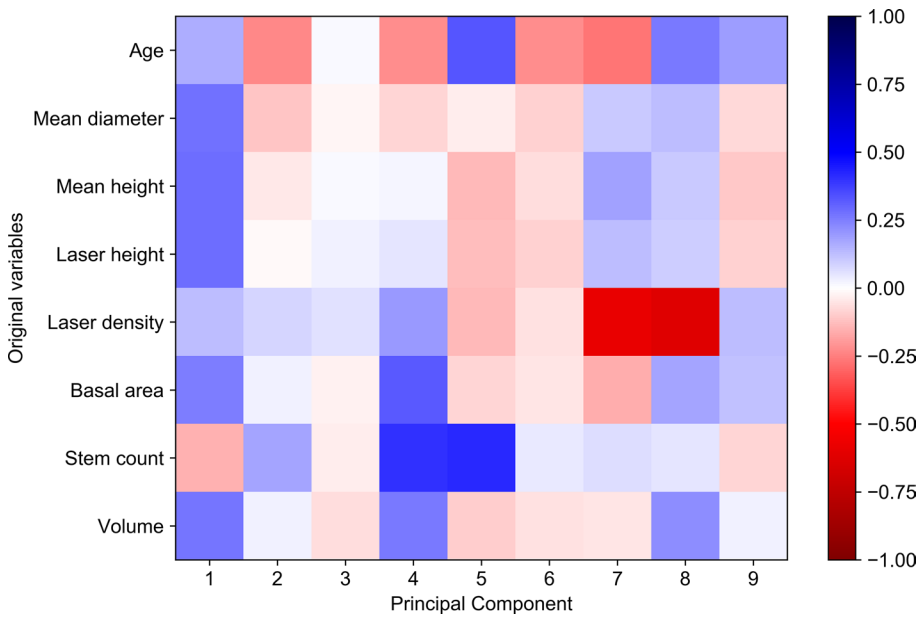

Fig. 3 Loadings of general forest variables in principal components

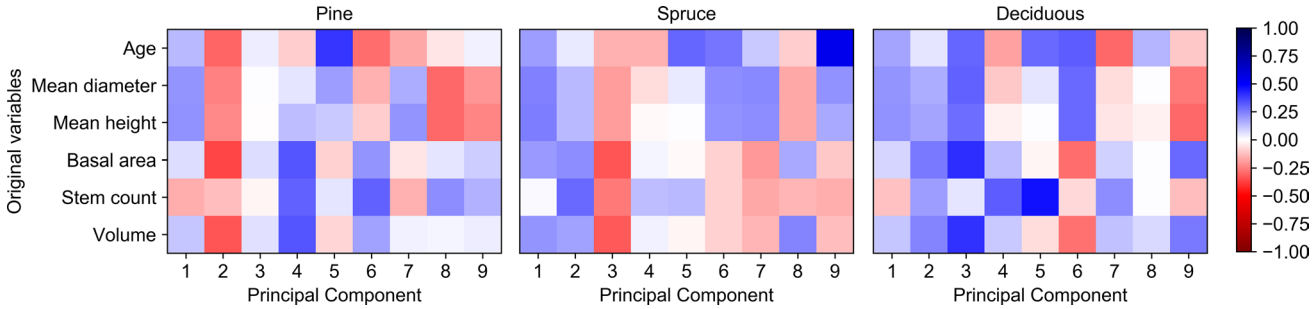

Fig. 4 Loadings of forest variables related to tree species related in principal components

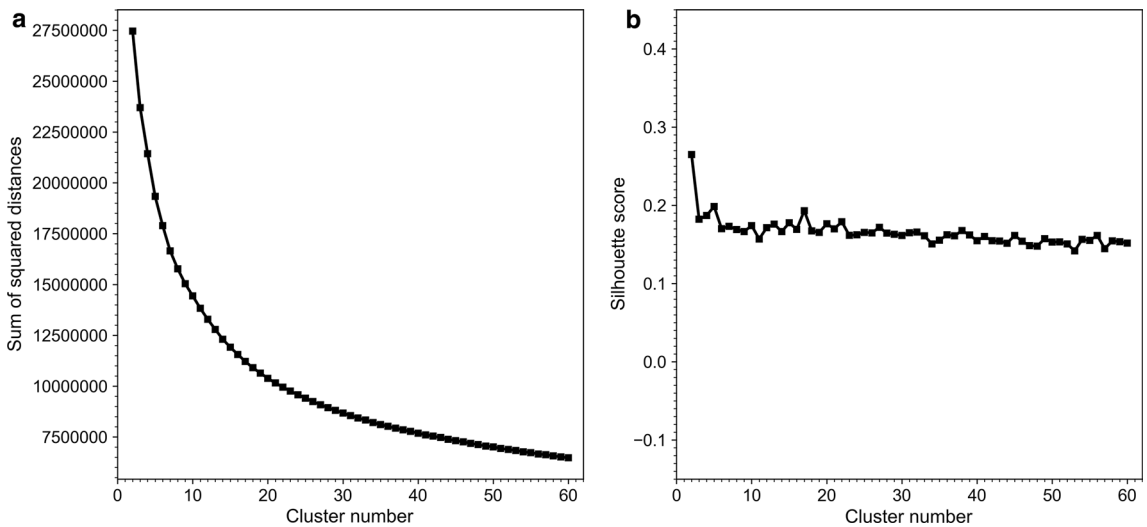

Fig. 5 Scores of tree clustering with cluster numbers from 10 to 60

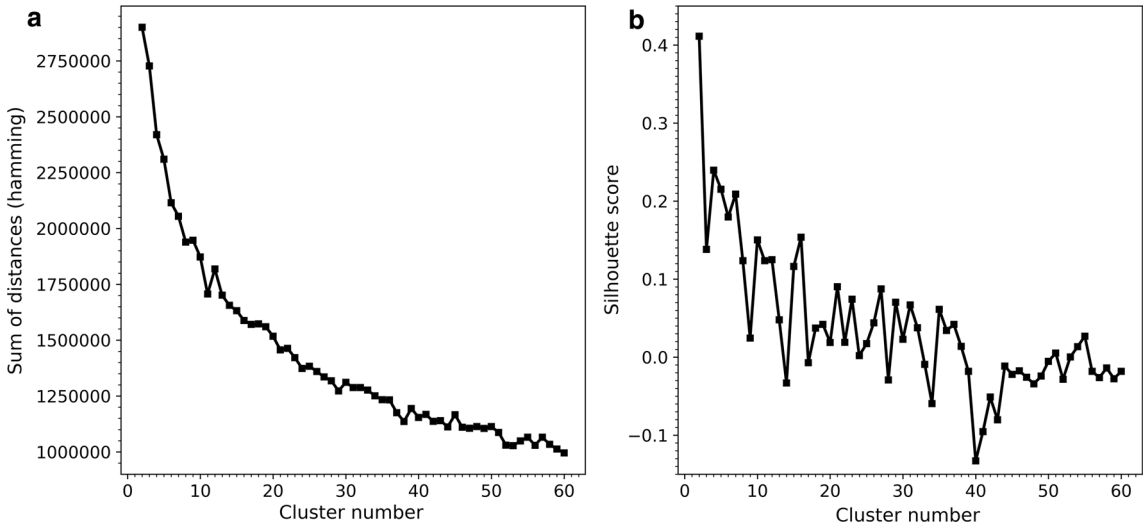

Fig. 6 Scores of soil clustering with cluster numbers from 10 to 60

Table 4 Cluster centers for the soil clusters

Cluster number (share of the grid cells in field tests)	Soil type	Harvesting accessibility	Drainage state	Fertility class	TWI class
1 (34.2%)	Coarse moorland	Mineral soil, accessible during summer	Unditched moorland	Fresh moorland or corresponding wetland	2
2 (3.8%)	Coarse moorland	Mineral soil, accessible during summer	Unditched moorland	Dry moorland or corresponding wetland	2
3* (0.7%)	Peatland	Wetland, accessible during summer	Natural state wetland	Dry moorland or corresponding wetland	2
4 (9.2%)	Coarse moorland	Mineral soil, accessible during dry summer	Unditched moorland	Fresh moorland or corresponding wetland	1
5* (2.1%)	Peatland	Mineral soil and wetland, accessible during winter time	Natural state wetland	Rough moorland or corresponding wetland	3
6 (44.5%)	Fine moorland	Mineral soil, accessible during dry summer	Unditched moorland	Moorland with rich grass-herb vegetation or corresponding wetland	2
7 (5.4%)	Coarse moorland	Mineral soil, accessible on summer time	Unditched moorland	Dry moorland or corresponding wetland	1

*Denotes a cluster not included in field test analysis

Table 5 Description of the five most common tree clusters in the field tests

Cluster number (share of the grid cells in field tests)	Dominant tree species	Mean age (years)	Mean diameter (cm)	Mean height (m)	Stem count (pcs/ha)	Tree volume (m^3/ha)
15 (7.3%)	Spruce	52	23.6	18.9	682	175.5
17 (9.2%)	Spruce, pine, deciduous	66	25.6	19.9	466	133.7
22 (11.4%)	Spruce	60	25.5	20.6	771	262.8
33 (8.0%)	Deciduous	65	25.3	21.0	929	250.9
41 (12.5%)	Spruce	79	30.2	24.1	673	378.5

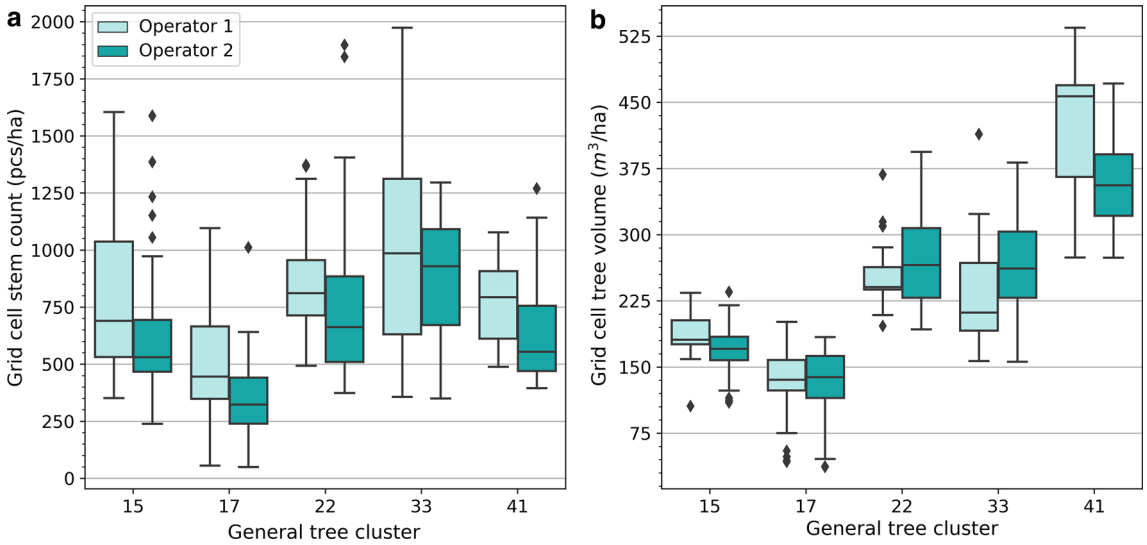

Fig. 7 General tree clusters applied to the grid cell stem count (a) and volume (b) of the visited cells in the field tests

Factors affecting harvester fuel consumption

The most important factors affecting the fuel consumption of the harvester were studied separately for the moving and for the working harvester. The data for the working harvester was limited to cases of regeneration felling, as the consumption between thinning and regeneration operations are considerably different and thus should be studied separately. In the analysis of working harvester, only fieldbus data windows containing at least one cut of a log according to the production records were included. In the analysis of moving harvester, data was restricted to cases where harvester inclination in the direction of traversal was less than 5 degrees. Figure 8a and b shows the fuel consumption (mean of time series samples) for the driving motion in the most common soil clusters and for the felling work in the most common tree clusters. The significance of the cluster and machine-operator unit in the differences between group means were evaluated using two-way ANOVA, see Table 6. According to the test, fuel consumption has statistically significant differences between the machine-operator units and the forest cluster groups with a significance level of 0.05.

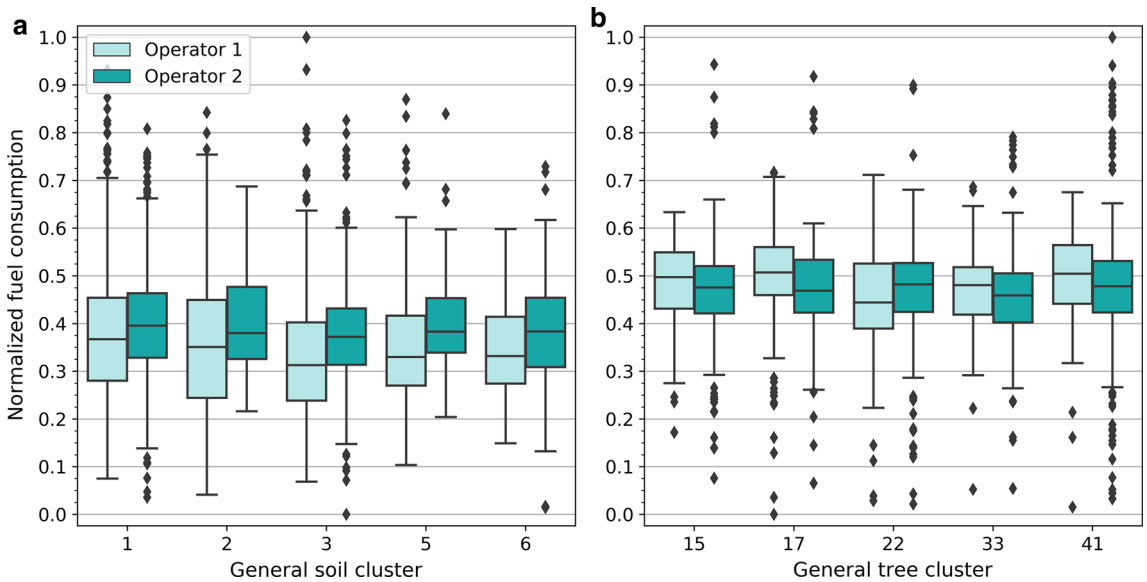

Fig. 8 Distribution of the fuel consumption (mean of time series sample) for moving harvester under general soil clusters (**a**) and for regeneration felling under tree clusters (**b**)

Table 6 Results of two-way ANOVA with soil (driving) and tree (working) clustering for fuel consumption

	Sum of squares	Degrees of freedom	*F*	*p*
Harvester driving				
Operator	3.853407	1	49.21	< 0.001
General soil cluster	1.197522	4	8.70	< 0.001
Operator: general soil cluster	4.743062	4	2.87	0.022
Residual	1.206663	3116		
Harvester working				
Operator	8246.85	1	4.28	0.037
General tree cluster	40317.92	4	5.23	< 0.001
Operator: general tree cluster	25885.93	4	3.36	0.010
Residual	4385958.0	2277		

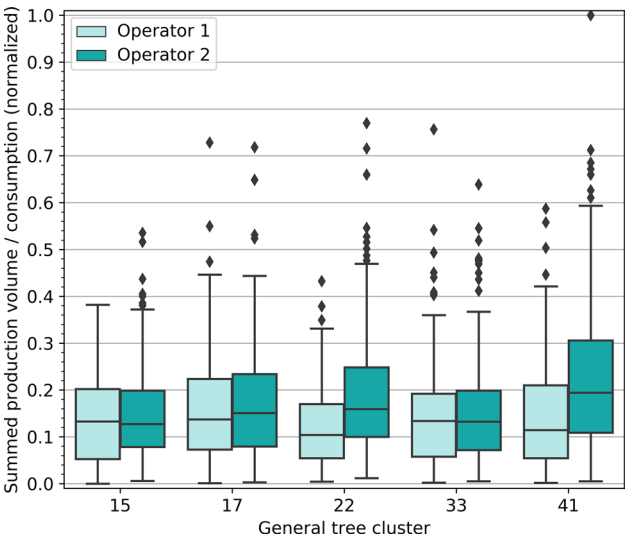

Fig. 9 Summed production volume per mean fuel consumption in the recorded time windows (regeneration felling)

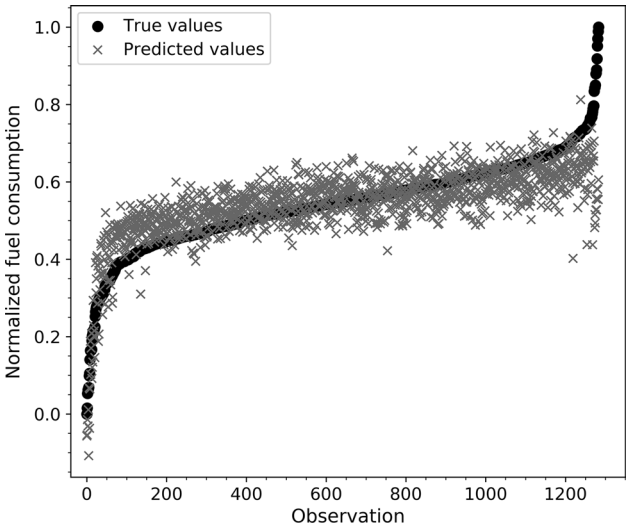

Fig. 10 Test data predictions with a linear regression model of the consumption ($R^2 = 0.61$)

Table 7 Coefficients of the linear regression models with standardized independent variables

Harvester working		Harvester driving	
Linear regression coefficients			
Feed control forward	14.5	Inclination (front-rear)	63.6
Diesel engine RPM	13.6	Speed	45.2
Boom 2nd joint control	8.9	Diesel engine RPM	34.6
Number of cuts	7.2	Harvester steering control	28.6
Feed control backward	5.4		
Boom rotation control	4.6		
Boom 1st joint control	3.6		
Boom extension control	3.1		
Tree volume in the grid cell	1.6		
Stem count in the grid cell	1.4		
Harvester head rotation control	1.1		

Analyzing fuel consumption of the two harvesters separately for each cluster (Fig. 8) means that the effect of differences due to forest properties has been removed, but differences due to their production rates still affect the fuel consumption. Figure 9 depicts the produced volume per fuel consumption for revealing the efficiency differences between the machine-operator units.

Figure 10 shows the performance of a linear regression model trained to predict the fuel consumption while the harvester is working, based on the signals listed in Table 3. Similarly, a model was trained for the moving harvester. The R^2 value for the working harvester model was 0.61 and 0.62 for the moving harvester model. The coefficients of the models are reported in Table 7.

Based on the weighted differences in the signals, three most important reasons for higher consumption between the machine-operator units at each cluster are presented in Table 8. The Operator 1 had higher fuel consumption while working, except for the tree cluster 22, whereas the Operator 2 had higher fuel consumption while driving in all of the presented soil clusters.

Table 8 Most important reasons for higher consumption of a machine-operator unit within a tree cluster

	Reason 1	Reason 2	Reason 3
Soil clustering			
1	Higher speed	Increased steering movement	More inclined route (front-rear)
2	Higher speed	More inclined route (front-rear)	–
4	Higher speed	Increased steering movement	–
6	Higher speed	Increased steering movement	–
7	Increased steering movement	Higher speed	–
Tree clustering			
15	Higher diesel engine RPM setting	Increased 2nd boom joint movement	Increased boom extension movement
17	Higher diesel engine RPM setting	Increased stem count in the grid cell	Increased harvester head rotator movement
22	Increased feed control forward	Number of cuts	Increased feed control backward
33	Higher diesel engine RPM setting	Increased 2nd boom joint movement	Increased harvester head rotator movement
41	Higher diesel engine RPM setting	Increased tree volume in the grid cell	Increased boom extension movement

Discussion

The availability of comprehensive forest inventory data in Finland enables new possibilities for analyzing forest machine performance automatically. Without the forest data, performance metrics of the harvesters are not comparable between the stands as the soil and tree properties affect the forest machine routing and production. The comparability of the performance metrics is important when instructing the operators or tuning the machine parameters, as these actions need to be tailored for the current forest environment. This study presented ideas for creating a machine learning pipeline capable for the generalization of the forest environment and detection of the reasons behind the measured differences.

The pipeline for generating the general tree clusters started with a dimensionality reduction of forest inventory data with the PCA. The contributions of the 28 forest inventory variables in the PCA transformation are presented in Figs. 3 and 4. The meaning of each nine principal component is not obvious, but when looking at the contributions of the original forest inventory variables in the figures, explanation in forestry perspective can be found rather easily for at least for the first four components. The first component after the PCA transformation describes the sturdiness of the trees in the grid cell, regardless of the tree species. Original inventory variables that are related to the higher tree mass increase and decrease together, only exception being the stem count, which acts in the opposite way. Thus, the value of the first principal component is high in grid cells containing old and massive trees rather sparsely and low in grid cells containing young trees densely. The next two components mainly describe the tree species: the second component separates forests where pine is the dominant species from the rest, and the third component distinguishes spruce and deciduous dominant forests. The fourth component seems to react to the age of the trees in the grid cell, also separating forests with pine as the dominant tree species. Although PCA interpretations are logical and clear, the number of clusters for principal component scores is not evident for the K-means clustering. The quality of clustering as a function of number of clusters was evaluated with two techniques. Based on the stabilization of sum of squared distances (Fig. 5), the number of clusters was chosen to be 45, but there seems to be no unique number of clusters that would lead to superior clustering results over other choices. Furthermore, low silhouette score values indicate that the tree data shows no clear clustered structure. However, the five most common clusters in our field tests turned out to be useful: the stem count and the tree volume, given as an example of the forest inventory variables in the grid cells, were similar in the grid cells having the same cluster index in the route of the two harvesters (see Fig. 7). Furthermore, the fuel consumption showed statistically significant differences between the clusters and the machine-operator units according to two-way ANOVA. Such findings signify that the clustering succeeded in the standardization of forest inventory data: the two machines experienced similar forest conditions in the grid cells with the same cluster so the performance indicators under the conditions of the cluster are comparable. However, clustering of forest inventory data needs further research as the K-means algorithm did not return strong clusters for the PCA transformed data.

Similar clustering, although with a K-modes algorithm suitable for categorical variables, was performed to find general soil clusters in Finland. In this study, all the soil-related variables showing reasonable variation in the sampled areas and available from the forest data platform were included in the clustering. It should be noticed that some of the variables, such as trafficability, are originally derived partly from the other included variables. The results indicate that with the selected variables, the best number of clusters can be

found from the range of 2–15 clusters. This is a rather low number considering the number of possible combinations of the variable classes but can be partly explained with the rather strongly correlated categorical variables. The clustering into seven clusters shows significant differences in the fuel consumption between the clusters and the clusters with higher consumption have cluster centers which suggest more moist soil conditions. Unfortunately, in the field tests the most wet conditioned clusters (3 and 5) were scarce. In such conditions, the fuel consumption would have been expected to be at its highest based on earlier research (Melander et al. 2019).

Labeling machine signal samples according to the nationwide grid cell clusters and analyzing each cluster label separately revealed differences in the fuel consumption and production of the machines-operator units (Fig. 9). Based on the presented summaries of the consumption, it seems that Operator 2 managed to work more efficiently in most of the tree type clusters: with less consumed fuel and with a higher volume of logs produced. Finding such differences is in itself important, but even more valuable knowledge is the reasons behind the differences. Therefore, a linear regression model was fitted for predicting the fuel consumption based on all the recorded forest inventory data and the operator controllable variables. The resulting R^2 values, 0.62 for the moving and 0.61 for the working harvester, denote that the independent variables are not enough for explaining the entire variation in the fuel consumption. However, most of the variation is explained, and the learned coefficients for the variables seem reasonable regarding the preconceptions of the factors affecting the harvester fuel consumption. As seen in Table 8, the model predicts the RPM level being the most important reason for the higher consumption of Operator 1 while working under most of the tree type clusters. This is not surprising, as it was known that the operator had higher RPM setting for the regeneration felling and higher motor RPM is known to cause higher fuel consumption. However, the higher RPM seems not to have justification as the production done by the Operator 1 is lower. When driving forward, the higher consumption of Operator 2 was explained mostly by the higher speed, indicating that, for example, the inclination of the route, the most influential cause according to the model, was not significantly different between the harvested stands of the two operators.

Linear models are particularly well suited for problem settings where contributions of single features to the target variable needs to be understood. In this study, the importance of the operator control signals with respect to fuel consumption were characterized by multiplying the mean differences of the signals between operators with the learned model coefficients, in a certain tree type cluster. Generalizing this simple linear importance assessment to a nonlinear one has several options for importance evaluation, such as permutation importance. However, even this method would result only in the importance without indication of sign of the effect, thus reducing the possibilities to infer about the differences in the performance.

In any future fleet-wide application, comparing operators pairwise is probably not sensible or even feasible. If data—signal means labeled according to forest type—would be continuously collected, comparing each operator against a common forest-type specific distribution would be a more fruitful approach toward improving timber harvesting efficiency. The methods presented in this paper would serve as the key functions of such a system. When the system would identify exceptionally low performance values, it could assist the operator to either tune the appropriate machine parameters or to change the detected non-efficient working routines.

Conclusion

In this study, large datasets of both, the forest inventory data and the machine fieldbus data were collected in order to reveal separately the impacts of the forest environment and of the way the harvester is operated on the machine performance. The performance of two machine-operator units was compared by investigating the differences in fuel consumption and log production after the variation in forest inventory data in individual forest locations was taken into account. The variation was successfully managed by clustering the forest inventory data in Finland with unsupervised machine learning methods, thus finding general forest types that apply for the whole country. Such clustering model is valuable when collecting data from individual forest operations, as the collected machine data will be comparable within all other locations in Finland sharing the same clustering group, i.e., the same forest type. Inside the clustering group, meaning all the grid cells having the same cluster index, the forest environment can be considered to be constant, enabling fair comparisons between machine-operator units. Additionally, this paper proposed linear regression model for predicting the fuel consumption of the machine based on the operator input and the forest inventory data. Such model can explain the actual reasons behind the detected differences in performance, if considered inside the clustered forest type. The statistically significant results suggest that the differences between the forest environments, operator actions and machine settings need to be closely monitored when evaluating the performance of separate forest operations.

The forest data processing methods presented in this paper are aimed for managing and taking advantage of the increasing amounts of data produced in forestry operations. The presented pipeline of data fusion and machine learning methods was designed to enable continuous data collection of machine

fieldbus data in respect to the forest inventory data. Once the clustering model for the forest inventory data is trained, as shown in this paper, the data fusion process in individual forest machines necessitates only local forest inventory data for associating the machine fieldbus data to the general forest types. This decreases the need for high capacity communication to data servers from the forest machines, as the data fusion results can be calculated in the on-board computer before transmission of the results. In addition, the suggested approach enables performance analysis to be constrained to very specific situations, for example to sawing of the logs, machine climbing uphill or single crane movement situations. Such automated methods are required for full-scale, commercial applications for fleet-wide performance improvement.

The clustering proposed in this paper offers a first version of the general forest types in Finland, but further cooperation of forestry researchers is needed for developing it further. The results regarding the impact of the environment to the harvester performance in this paper seem to be consistent with our earlier results in Melander et al. (2019), although the clustering of the soil-related data is somewhat different. Furthermore, in both studies, mean values of signals in the fieldbus data windows were used when analyzing the datasets. A future research is suggested on using more versatile signal features of the fieldbus data after the data fusion.

Acknowledgments Ponsse Oyj, especially Mr. Kalle Einola and Mr. Simo Tauriainen, are acknowledged for their great effort in arranging the field test data collection for the two harvesters and helping with the interpretation of the recorded signals. Metsäteho Oy and its partners are thanked for providing access to the Forest Data Platform and thus enabling automatic retrieval of forest data.

Funding The author(s) received no specific funding for this work.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

Availability of data and material The forest data used in this study is public and downloadable from the eService of Metsäkeskus (see https://www.metsaan.fi/en/briefly-english). The collected fieldbus data that support the findings of this study are available from Ponsse Oyj, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Code availability The code for the algorithms in this study is open source (Python libraries), and the documentation can be found from the references of each cited algorithm.

References

Ala-Ilomäki J, Lamminen S, Sirén M, et al (2012) Using harvester CAN-bus data for mobility mapping. In: Spec issue abstr int conf organ by LSFRI silava coop with SNS IUFRO, vol 25, pp 85–87

Cajander AK (1909) Über Waldtypen. Acta For Fenn 1:1–175

Cajander AK (1949) Forest types and their significance. Acta For Fenn 56:1–71

Chaturvedi A, Green PE, Carroll JD (2001) K-modes clustering. J Classif 18:35–55. https://doi.org/10.1007/s00357-001-0004-3

Dash J, Pont D, Brownlie R et al (2016) Remote sensing for precision forestry. NZ J For 60:15–24

Eriksson M, Lindroos O (2017) Productivity of harvesters and forwarders in CTL operations in northern Sweden based on large follow-up datasets. Int J For Eng 25:179–200. https://doi.org/10.1080/14942119.2014.974309

Finnish Forest Center (2019) eServices for forest owners and service providers. https://www.metsaan.fi/. Accessed 2 Feb 2020

Gellerstedt S, Dahlin B (1999) Cut-to-length: the next decade. J For Eng 10:17–24

Häggström C, Lindroos O (2016) Human, technology, organization and environment—a human factors perspective on performance in forest harvesting. Int J For Eng 27:67–78. https://doi.org/10.1080/14942119.2016.1170495

Hämäläinen J (2016) Kohti puuhuollon digitalisaatiota—Forest Big Data—project. In: Metsätehon tuloskalvosarja 11/2016. http://www.metsateho.fi/kohti-puuhuollon-digitalisaatiota/. Accessed 2 Feb 2020

Hellström T, Lärkeryd P, Nordfjell T, Ringdahl O (2009) Autonomous forest vehicles: historic, envisioned, and state-of-the-art. Int J For Eng 20:31–38. https://doi.org/10.1080/14942119.2009.10702573

Holopainen M, Vastaranta M, Hyyppä J (2014) Outlook for the next generation's precision forestry in Finland. Forests 5:1682–1694. https://doi.org/10.3390/f5071682

Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recognit Lett 31:651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kangas A, Astrup R, Breidenbach J et al (2018) Remote sensing and forest inventories in Nordic countries–roadmap for the future. Scand J For Res 33:397–412. https://doi.org/10.1080/02827581.2017.1416666

Kankare V, Luoma V, Saarinen N et al (2019) Assessing feasibility of the forest trafficability map for avoiding rutting—a case study. Silva Fenn 53:1–9

Lindroos O, Ringdahl O, La Hera P et al (2015) Estimating the position of the harvester head—a key step towards the precision forestry of the future? Croat J For Eng 36:147–164

Lu K, Bi H, Watt D et al (2018) Reconstructing the size of individual trees using log data from cut-to-length harvesters in Pinus radiata plantations: a case study in NSW, Australia. J For Res 29:13–33. https://doi.org/10.1007/s11676-017-0517-1

Luke (2019) National Forest Inventory (NFI). http://www.metla.fi/ohjelma/vmi/info-en.htm. Accessed 2 Feb 2020

Lundbäck M, Häggström C, Nordfjell T (2018) Worldwide trends in the methods and systems for harvesting, extraction and transportation of roundwood. In: 6th international forest engineering conference. Rotorua, New Zealand

Mason EG, Morgenroth JA, Bown HE (2016) Precision forestry research project—final report. In: University of Canterbury Research Repository. https://ir.canterbury.ac.nz/handle/10092/13431. Accessed 2 Feb 2020

Melander L, Einola K, Ritala R (2019) Fusion of open forest data and machine fieldbus data for performance analysis of forest machines. Eur J For Res. https://doi.org/10.1007/s10342-019-01237-8

Müller F, Jaeger D, Hanewinkel M (2019) Digitization in wood supply—a review on how Industry 4.0 will change the forest value chain. Comput Electron Agric 162:206–218. https://doi.org/10.1016/j.compag.2019.04.002

Obi OF, Visser R (2017) Influence of the operating environment on the technical efficiency of forest harvesting operations. Int J For Eng 28:140–147. https://doi.org/10.1080/14942119.2017.1357391

Olivera A, Visser R (2016) Using the harvester on-board computer capability to move towards precision forestry. NZ J For 60:3–7

Olivera A, Visser R, Acuna M et al (2016) 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 27:15–28. https://doi.org/10.1080/14942119.2015.1099775

Prinz R, Spinelli R, Magagnotti N et al (2018) Modifying the settings of CTL timber harvesting machines to reduce fuel consumption and CO2emissions. J Clean Prod 197:208–217. https://doi.org/10.1016/j.jclepro.2018.06.210

Rajala M, Ritala R (2016) Data platform promoting forest data utilization through uniform access to heterogeneous data. In: Metsäteho Rep. 240. http://www.metsateho.fi/data-platform-promoting-forest-data-utilization/. Accessed 2 Feb 2020

Ringdahl O (2011) Automation in forestry: development of unmanned forwarders. Umeå University, Sweden

Ringdahl O, Lindroos O, Hellström T et al (2011) Path tracking in forest terrain by an autonomous forwarder. Scand J For Res 26:350–359. https://doi.org/10.1080/02827581.2011.566889

Rossit DA, Olivera A, Viana Céspedes V, Broz D (2019) A Big Data approach to forestry harvesting productivity. Comput Electron Agric 161:29–52. https://doi.org/10.1016/j.compag.2019.02.029

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shan C, Bi H, Watt D, Li Y (2019) A new model for predicting the total tree height for stems cut—to—length by harvesters in *Pinus radiata* plantations. J For Res. https://doi.org/10.1007/s11676-019-01078-6

Suvinen A, Saarilahti M (2006) Measuring the mobility parameters of forwarders using GPS and CAN bus techniques. J Terramech 43:237–252. https://doi.org/10.1016/j.jterra.2005.12.005

Talbot B, Pierzchała M, Astrup R (2017) Applications of remote and proximal sensing for improved precision in forest operations. Croat J For Eng 38:327–336. https://doi.org/10.5281/zenodo.890539

Tomppo E, Gschwantner T, Lawrence M, McRoberts RE (2010) National forest inventories: pathways for common reporting. Springer, New York

Venäläinen P, Räsänen T, Hämäläinen J (2015) Potential business models for forest big data—data to intelligence (D2I) project report, Task 3.2. In: Metsäteho Rep. 235. http://www.metsateho.fi/potential-business-models-for-forest-big-data-data-to-intelligence-d2i-project-report-task-3-2/. Accessed 2 Feb 2020

White JC, Coops NC, Wulder MA et al (2016) Remote sensing technologies for enhancing forest inventories: a review. Can J Remote Sens 42:619–641. https://doi.org/10.1080/07038992.2016.1207484