European Journal of Forest Research (2020) 139:213–227
https://doi.org/10.1007/s10342-019-01237-8

ORIGINAL PAPER

# Fusion of open forest data and machine fieldbus data for performance analysis of forest machines

Lari Melander[1] · Kalle Einola[2] · Risto Ritala[1]

Received: 16 April 2019 / Revised: 3 September 2019 / Accepted: 12 October 2019 / Published online: 21 October 2019


## Abstract

Forest resource data is important in targeting the forestry operations, and it is in the hearth of the precision forestry concept. The forest resource data can be produced with many techniques, and the number of existing forest data sources has increased during the years. In addition to the forest resource data, other data describing the circumstances of the forest site, such as trafficability and weather conditions, are available. In Finland, a forest data platform gathers the data sources under a single service for easier implementation of the precision forestry applications. This data is useful in operations planning, but it also describes the conditions that prevail when the forest machine arrives to the forest site. This study proposes data fusion between fieldbus time series of the forest machine and the forest data. The fused dataset enables explorative statistical analysis for examining the relationship between the machine performance and the forest attributes and provides data for building predictive models between the two. The presented methods are applied into a dataset generated from a field test data. The results show that some fieldbus time series features are predictable from forest attributes with $R^2$ value over 0.80, and clustering methods help in interpreting the machine behavior in different environments. In addition, an idea for generating a new forest data source to the forest data platform based on the fusion is discussed.

**Keywords** Forestry · Data fusion · Machine learning · Forest data · Fieldbus data · Harvester · Forwarder

## Introduction

In precision forestry, a wide range of data sources is utilized to generate accurate information about the state of the forest at a given location. This information is mainly for forest resource management and decision support systems in the wood procurement process for targeting forestry operations (Fardusi et al. 2017; Gülci et al. 2015). Detailed and timely information about forest parameters is collected primarily with remote sensing technologies, such as airborne laser scanning (ALS), aerial photography and satellite imagery methods (Holopainen et al. 2014; Mason et al. 2016; Fardusi et al. 2017). Forest machines, in particular the cut-to-length (CTL) machinery, are able to gather the forest parameters as the tree dimensions are measured during the tree harvesting (Lindroos et al. 2015; Mason et al. 2016; Lu et al. 2018). The environment sensing capabilities of the forest machines at forest site can be extended by equipping them with mobile laser scanners (MLS) or different types of cameras (Melander and Ritala 2018; Salmivaara et al. 2018; Holopainen et al. 2014). Furthermore, the current state estimate of the forest for precision forestry operations can be updated with data from other geographical information systems (GIS), such as soil topography and weather databases (Mason et al. 2016; Salmivaara et al. 2017).

The data sources for precision forestry, described above, are very homogenous in terms of the representation and the availability. In Finland, these data sources are being collected under a forest Big Data platform for easier implementation of new precision forestry applications (Hämäläinen 2016; Venäläinen et al. 2015; Rajala and Ritala 2016). This platform is intended to be publicly available in near future, but currently it is still in a development phase with a restricted access. The platform has a single interface where a client application can retrieve forest-related data originally

Communicated by Eric R. Labelle.

✉ Lari Melander
lari.melander@tuni.fi

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

[2] Ponsse Plc, Ponssentie 22, Vieremä, Finland

residing in many separate databases. The client can freely request any number of data sources for a given forest area in Finland, and the platform returns a single fused dataset containing the data in standardized grid cells. The grid cell size adopted in the platform is 16 m × 16 m. At the time of writing, there was data from 10 open sources available through the platform implementation. The platform includes for example forest inventory data (Finnish Forest Center 2019; Luke Natural Resource Institute of Finland 2019), land survey data (National Land Survey of Finland 2019), weather reports and predictions (Finnish Meteorological Institute 2019), other smaller datasets like forest classification based on changes in the Sentinel-2 satellite images, and metadata like Finnish municipality borders. The geographical area covered is wide; values of many forest attributes are available throughout Finland. The interface is built for automatic retrieval of the forest data, enabling easier implementation of new services in precision forestry framework. In addition, the platform supports custom data sources generated by the client applications. For instance, forest machine data, harvesting locations or forest parameters measured by the harvester could be uploaded to the platform as private or public data.

Harvesters and forwarders, used in the CTL framework, measure extensively parameters related to their own performance and the wood procurement process. The automation layer of the CTL machine consists of a fieldbus system that connects all the related units, such as actuators, sensors and controllers together, forming a distributed control system. The control system constantly produces and processes hundreds of signals related to the vehicle engine, transmission and harvester head performance and control, and to the production. The control system and the human operator interact through the on-board IT system of the forest machine, which also produces standard production and performance data based on the measurements during the work. This data can be transferred from the IT system of the forest machine to other wood procurement IT systems via mobile networks, offering a source of information about the visited forest site and the executed work (Olivera and Visser 2016; Strandgard et al. 2013). The CTL forest machines have an on-board Global Navigation Satellite System (GNSS) receiver that constantly records the geospatial coordinates of the machine, thus enabling the positioning of the recorded data. For precision forestry, the positioned harvester data can be fused with the GIS data sources, for example remotely sensed forest inventories (Lu et al. 2018). With such fused data, the interactions between the machine performance and environment can be better understood, thus expanding the usefulness of harvester data beyond collection of forest parameters. For example, Eriksson and Lindroos (2014) and Obi and Visser (2017) have analyzed the production and the performance data to learn the effect of the working environment on the logging performance. Most notably, they found that the stem size, the forest terrain and the size of the operation significantly affect the harvester productivity. However, detailed analysis of the machine behavior in different environments requires fieldbus data. Although the forest machines have quite good communication capabilities, neither time series of fieldbus variables nor their summaries are automatically recorded for later investigation. As the number of variables available on the fieldbus is high, recording all the variables at a high sampling rate is not feasible due to the limited communications network bandwidth. Hence, the fieldbus time series must be either analyzed and summarized at the on-board computer or collected for a limited test period and possible limiting the number of signals. Examples of the latter exist in literature; for example, in field tests of Suvinen and Saarilahti (2006) and Ala-Ilomäki et al. (2012) variables from the forest machine fieldbus were collected over a limited period of time and then analyzed offline together with the environmental data. These authors focus on modeling the relationship between certain machine variables and the environmental data for trafficability and have found that the motion resistance measured from the fieldbus variables has a connection to the resulting wheel rut depth.

Many studies on the effect of the working environment on the forest machine operation have been published (see, e.g., Han et al. 2006; Häggström and Lindroos 2016; Obi and Visser 2017; Sirén et al. 2019). Recent research shows that the operator can significantly affect the forest machine performance by setting appropriate machine parameter values for the given environment (Prinz et al. 2018). It was found that in particular harvester fuel consumption can be reduced, if the settings of the machine are manipulated correctly. Currently, forest machines calculate locally a set of performance indicators from the fieldbus time series, which can help the operator to tune the machine parameters. However, such performance indicators ignore the effect the environment causes. Fusion of forest machine signals, produced during the normal logging work, and the nation-wide forest platform data would provide automatically *forestry Big Data* that allows data analysis methods and machine learning algorithms reveal useful machine–environment relationships. Should a set of forest parameters be clearly related to a level of a particular machine signal, valuable information for both forest machine development and forest operations planning would be obtained.

The current paper presents how machine data and forest platform data are fused and analyzed in field tests. Initial findings resulting from the automatic data fusion of two completely different data sources in forestry domain are presented. The detailed aims of this paper are: (1) to present an automated data fusion approach that combines forest resource data to fieldbus time series data of the forestry machine; (2) analysis of the forestry equipment and forest

environment interactions in the fused data; and (3) discuss the opportunities of such data fusion and storing summaries of forest machinery data in the platform as a private or public data source.

## Materials and methods

### Fusion of forest data

The forest data platform offers possibility to request only the data sources that are important for the client application. It is in a pilot stage, and new data sources are being added continuously. This study analyzes forest inventory data produced by the Finnish Forest Center, which is available for the field test site. In addition, land survey data were included in the dataset. The client queries the platform in JSON (JavaScript Object Notation), defining the geographic area of the data to be retrieved, the data sources and the data tags within the sources. The area is expressed as a GeoJSON multi-polygon, and the platform returns the requested data for each 16 m × 16 m grid cell intersecting the given polygon. The platform returns the data as CSV files (Comma Separated Values), so that an individual Finnish (ETRS-TM35FIN) map sheet (size 24 km × 48 km) generates one CSV file. In the data, a grid cell is represented as a single row, with an identifying grid number, and values of each forest parameter in its columns.

The fieldbus data of a harvester (Ponsse Scorpion King) and a forwarder (Ponsse Elk) were collected during field tests in Vihti region in Finland in May 2016. The data consists of 22 signal channels (time series) recorded 4.5 h for the harvester and 2.75 h for the forwarder. The tests consisted of harvester and forwarder runs on the same forest path. The harvester first opened the path, conducting an ordinary thinning operation, and the forwarder followed the route with a constant tree load. The forwarder run was repeated several times in selected parts of the test track. The dataset has been the basis of research on rut formation and measurements (see Salmivaara et al. 2018; Sirén et al. 2019), which specifies the test run environment and conditions, not repeated here. In this study, seven fieldbus channels related to the diesel engine, transmission and cooling of the forest machine were analyzed together with forestry data. Hereafter, the fieldbus signals are referred to collectively *'System signals'* in this paper, only the traveling speed and fuel consumption are separately labeled. *System signals 1* and *3* are the rotational speeds of the hydrostatic drive motor and the diesel engine, respectively. *System signal 2* is the torque of the diesel engine, and the signals *4* and *5* are the control signals of the hydraulic motor and the cooling unit. The sampling time throughout the fieldbus data collection was 20 ms. The GNSS (Global Navigation Satellite System) receiver of the forest machine recorded the location for each time instant of the fieldbus data, although with a higher sampling time of one second. Such data will be referred to as *positioned time series*.

The data fusion between the forest data and the fieldbus signals is initialized by requesting the forest data for the area where the machine has been operating. The correspondence between the forest and the fieldbus data points is determined based on the location information given in the platform for the grid cells and by the forest machine positioned time series data. The forest data provided by the platform is static in the time scales of forest machine runs. As the forest data is given as 16 m × 16 m grid cells, the fieldbus time series data must be mapped on the same cells. Therefore, the second task in the data fusion operation is to label each time instant in the fieldbus dataset with the grid identification number associated to the recorded location at that time instant. This is referred to as *grid-positioned time series*. Thus each grid cell is associated with *grid episodes*, which represent machine operation at the cell. A grid episode is a subsequence of the total fieldbus time series, and the length varies from cell to cell depending on the duration the machine has spent in that cell (Fig. 1). Note that a grid cell may have several grid episodes associated to it, if the forest machine returns to a cell that was already visited earlier.

A grid episode includes all executed machine operations, such as driving forward or cutting a tree. To analyze the effect of the environment to a particular operation, the grid episodes need to be divided further. In this study, the driving motion of the machine was separated from the in-place work of the harvester and the forwarder based on a fieldbus signal that indicates whether the driver has enabled a brake for the in-place work. This breaks the grid episodes into several *sub-episodes*, as there can be

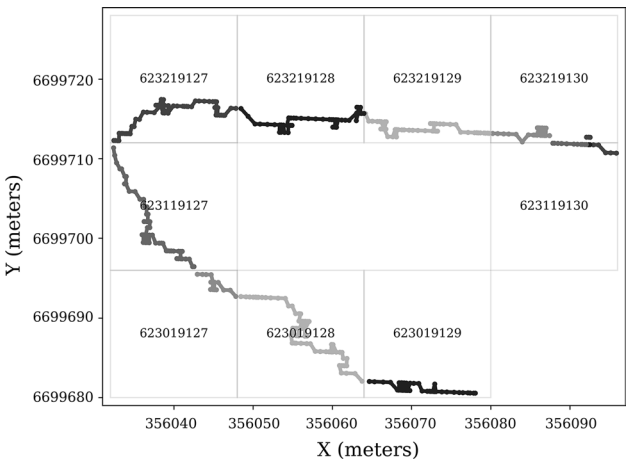


**Fig. 1** A forest machine route divided according to the grid cells of the forest data. Change in color indicates change in the grid episode

multiple in-place working points within a grid cell, and the average length of sub-episodes is naturally shorter. Figure 2a shows an example of the sub-episodes resulting from the driving motion. Here, each sub-episode holds a subsequence of fieldbus time series of varying length. Due to the effect of forest canopies to the performance of the GNSS receiver, the route seems slightly erratic and intermittent. Obviously, formation and selection of the sub-episodes is according to the research question addressed and Fig. 2a serves only as an example.

To generate useful datasets for statistical analysis and machine learning algorithms, the varying-length sub-episodes are sampled several times with a constant time window that is shorter than the original sub-episode. In this process, the starting point of each *constant-length sub-episode* (CLSE) is chosen uniformly randomly within the duration of the sub-episode, leading to a set of constant-length sub-episodes that can be overlapped. In this study, the number of the CLSE's per sub-episode was set so that the sum of the CLSE lengths corresponds to the length of the original sub-episode being sampled. Each resulting CLSE is considered as a representative sample of the machine time series for the current grid cell and for the current working mode. The route after the division into constant-length sub-episodes (CLSE) is shown in Fig. 2b. The chosen constant length affects greatly the number of total time series in a resulting fused dataset and has to be considered when building the dataset for analysis purposes. In general, the harvester sub-episodes have lower mean length than those of the forwarder, because the forwarder did not stop as often as the harvester. To clarify the terminology of the time series in this study, an overview is presented in Table 1.

Preprocessing is an important part in the automatic generation of the fused data sets. The dataset may lack some of the forest features for some grid cells, so removal of data points lacking some of the needed values was necessary. Forest parameters that are constant throughout the driven route are meaningless in the data analysis and were thus removed. Continuous forest data variables were normalized to zero mean and unit variance before analysis and nominal variables were coded as indicator variables for the machine learning algorithms. Furthermore, the dimensionality of the continuous forest variables was reduced with principal component analysis (PCA) to compact the representation. The variables describing the trees in the PCA transformation were *laser height, laser density*, *age*, *mean diameter, mean height*, *basal area* and *volume*. The laser height indicates a height were 85% of the laser observations are in a cumulative distribution for the grid cell, taking into account only observations that are above two meters. The percentage of the observations above the two meter level is described by the laser density. Other variables describe the average dimensions of the trees in the grid cell. In addition, ground height variation and change (above sea level) between the grid cells in the machine route were included in the PCA transformation when transforming the forest data for the field test area.

## Analysis of fused data

The fused dataset enables statistical analysis and machine learning modeling for seeking the relationships between the environment and the forest machine operation. Depending on the analysis method, the resulting fieldbus samples can be

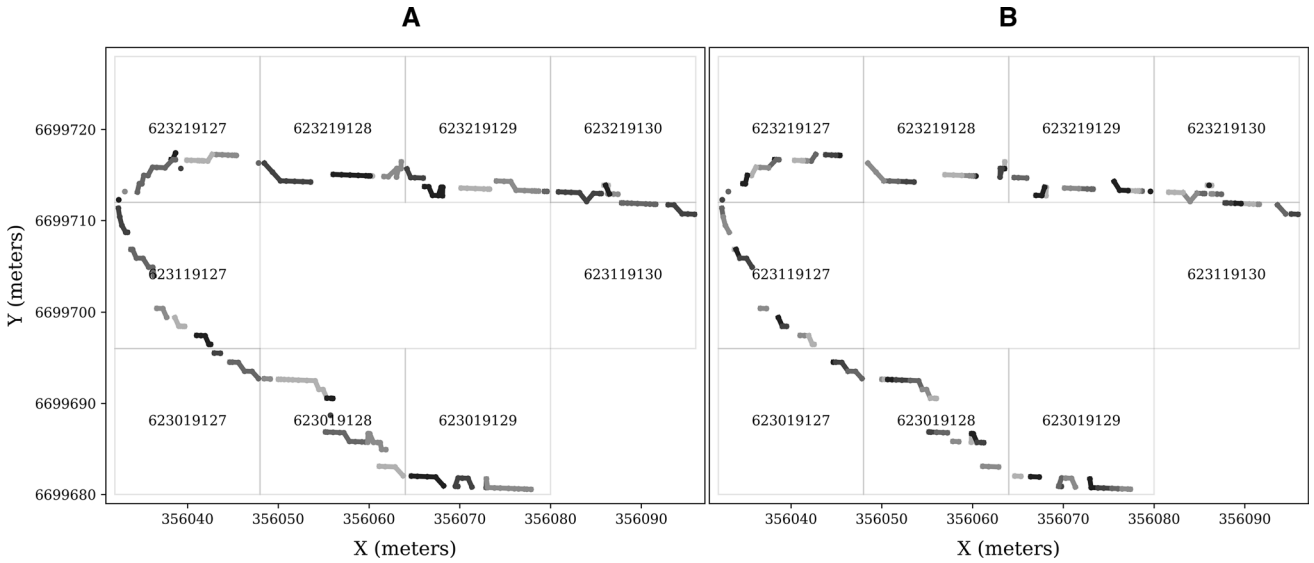


**Fig. 2** **a** Sub-episodes resulting from the driving motion of the forest machine. **b** Sub-episodes sampled with constant length (CLSE). Change in color indicates change in the sub-episode (**a**) or in the CLSE's (**b**)

**Table 1** The terminology of time series in this study

| Term | Explanation |
|---|---|
| Fieldbus time series | A recorded time series of the forest machine operation, consisting of several channels for different variables in the forest machine fieldbus |
| Positioned time series | A fieldbus time series that has the location of the forest machine, recorded by the GNSS receiver, for each time stamp |
| Grid-positioned time series | A fieldbus time series that has a grid cell identification number for each time stamp |
| Grid episode | Subsequences of the total fieldbus time series that are assigned into a specific grid cell. This can consist of multiple time series, if the machine visits the same grid cell repeatedly. See Fig. 1 |
| Sub-episode | A time series that is produced by dividing a grid episode according to the working mode (e.g., harvesting/moving). Thus, grid episodes usually consist of multiple sub-episodes. See Fig. 2a |
| Constant-length sub-episode (CLSE) | A section of the sub-episode that is sampled to have a predetermined constant length. Thus, the CLSE is commonly shorter than the sub-episode and a single sub-episode can be sampled to many CLSE's. See Fig. 2b. Note that the CLSE, like all other time series types described in this table, consists of several signal channels |

either transformed into a set of features or used as raw multivariate time series. Here, an arithmetic mean of each CLSE channel was adopted as the simplest feature for describing the individual CLSE channel sample. Such simplification of the time series into one simple feature enables straightforward correlation analysis and statistical significance tests for the fused data at the forest data grid level. In the fused dataset, the features of both the machine fieldbus time series and forest parameters (or the principal components of them) were located on the individual columns of the resulting dataset. Thus, a single row in the dataset represents a single fused sample dedicated to certain grid cell in the forest.

In this study, the Pearson correlation coefficient for each pair of the forest and machine variables quantified the linear dependencies in the entire fused dataset. High correlations within forest data or within machine signals allow reducing the set of variables in the dataset either directly or with methods, such as PCA. Then prediction models can be identified effectively between uncorrelated explanatory variables and uncorrelated target variables. Such correlations and models between continuous forest data variables and fieldbus time series signals can reveal important relationships in the machine–environment interaction.

Forest data includes many forest parameters that are categorical, rather than continuous. Soil type, main tree species, and trafficability class are examples of such variables. Obviously, their impact on continuous fieldbus time series cannot be evaluated through correlation measures; the question is that when CLSE channel features are grouped according to a categorical variable, does the statistics of the features differ significantly between the groups. The null hypothesis for statistical tests is that the categorical forest parameter, according to which the sub-episodes have been grouped, has no effect on the machine fieldbus time series and thus there are no significant differences between the feature statistics of the grouped sub-episodes. Most categorical forest parameters have more than two possible values; for example, soil type has six different classes for the grid cells within the field test set. For this reason, one-way analysis of variance (ANOVA) and equivalent nonparametric Kruskal–Wallis tests were utilized to evaluate if at least one of the resulting group means of features (CLSE channel means) differs statistically significantly from the other group means. The selected significance level was 0.05. As ANOVA assumes normally distributed data, the normality of each forest parameter groupings was tested (Jones et al. 2019), and the statistical test type was selected based on the results.

In addition to aforementioned statistical analysis, machine learning methods can be applied to the fused dataset. Machine learning can be divided into supervised and unsupervised methods, which differ in that the former requires the desired outcomes—labels or values—in the dataset, whereas the latter does not. Examples of methods from both categories in the analysis of the fused dataset are demonstrated in this study. Unsupervised machine learning covers clustering algorithms, which aim to find structure in data without known labels. In the current context, clustering can reveal structures or tendencies of the machine signals in different forest conditions, given only the fused data. In contrast, supervised machine learning algorithms can predict a value or a class if the target value is given in the training set but is not available currently. Here, the model can be trained for predicting a single fieldbus time series (CLSE channel) feature from the forest platform data, allowing predictions of the machine behavior for the grid cells never visited before. The main objective using any of the machine learning methods is to generalize rules from a fused training dataset for understanding the machine–environment relationship. All the machine learning analyses in this work were implemented with Python algorithms in the Scikit-learn library (Pedregosa et al. 2011).

Hierarchical clustering (Xu and Wunsch 2008; Seber 1984) of the fused dataset was employed as an example of the unsupervised learning methods. In hierarchical clustering, the data is clustered in different levels by merging or splitting clusters according to their distances. Hierarchical clustering is a descriptive method defining a process of collecting individual data points into clusters of increasing size. The bottom-up hierarchical clustering algorithm starts by assigning each data point to its own cluster and calculating a distance measure between each cluster. Then clusters that have the smallest intercluster distance are merged together. This creates clusters for a hierarchy level, and the algorithm continues iteratively by calculating the distance between the formed clusters and further merging the clusters. The merging is continued until all the data points are designated to a single cluster. In this study, the CLSE features were hierarchically clustered. A meaningful clustering of the fused data set is one in which all CLSE's of each grid cell are in the same cluster, thus also implying a clustering of the grid cells. Finding this kind of clustering of the grid cells based only on the machine time series data would indicate a meaningful relationship between the forest data and the machine data.

Hierarchical clustering is not used for associating a new data point to a cluster, whereas the well-known *K*-means—algorithm (Jain and Anil 2010) is suited for this purpose. In this work, Finnish forest grid cells were clustered into typical forest types with the *K*-means algorithm. *K*-means cluster centers allow association of any new data points to the clusters without the need of redoing the analysis with all the data. As the forest data platform enables access to forest data in the whole of Finland, the cluster centers can be identified for large forest areas representing all the types of forests in Finland. The grid cells of a field test area can then be allocated to the clusters according to their distances to the centers of the *K*-means clusters, thus enabling a more robust comparison between different field tests with varying forest types. However, the number of clusters in the *K*-means algorithm needs to be prescribed. It is not clear what is a good number of clusters with the forest data, and therefore, the clustering was experimented with number of clusters varying from 2 to 20. The number of forest grid cells in Finland is enormous, of the order of $10^9$. Thus ten large areas were queried, resulting to approximately 28 million grid cells (7000 $km^2$), distributed over 31 map sheets. The map sheets were further sampled to reduce the number of data lines, resulting a sample of 100,000 grid cells that was considered as a representative sample of the Finnish forests.

Supervised machine learning can be applied to solve classification and regression problems. In this study, regression algorithms were compared for predicting machine signal levels for a grid cell, when only forest data is given in advance. With such models, the aim is to predict forest machine performance before the operations. In the training phase, the target signal levels for the grid cells, essentially the mean values of the CLSE's channels, were given for the algorithm. As the length of the CLSE is a tuning parameter of the method, the method was experimented with different lengths to find the length that best correlates with the environment variables. This optimal CLSE length may vary depending on the settings and the features to be predicted. For instance, the signals resulting from the in-place work probably have different optimal CLSE length than the signals in driving motion. Many algorithms exist for learning the regression model from data, but here linear regression (LR) and random forest regression (RF) were utilized from the Scikit-learn library. The number of estimators in the random forest regressor was set to 100, but otherwise the algorithms were used with default parameters. The prediction accuracy was evaluated by calculating the coefficient of determination ($R^2$) for the regression results. The coefficients were calculated using *K*-fold cross-validation with five folds.

## Results

### Dataset statistics

Correlations were investigated between continuous forest parameters and machine time series features. Figure 3 presents a correlation matrix for CLSE channels in the case where the harvester was moving in the forest. The length of the CLSE's was 300 samples (6 s). The results are similar with the sub-episodes of the in-place working harvester and the driving motion of the forwarder.

Considering the data fusion approach, the most interesting correlations are between machine signals and the forest parameters. Many of the variables show no significant correlation, while some variable pairs between the groups have correlation with absolute value in the range 0.3–0.5. The figure also shows clearly some high correlations within the forest parameters and, respectively, within fieldbus time series features. This suggests that rather than identifying regression models between the original variables, those between some linear combinations of variables is preferable. For instance, tree basal area and tree volume have a close to perfect correlation, so there is no reason to include both as independent variables in the regression analyses.

Principle components are a natural way of describing highly correlated forest parameters. The four first components explain 95.4% of variation in the nine forest parameters (tree and soil) in Fig. 3. The percentage of variation explained and the loadings of the four components are given in Table 2.

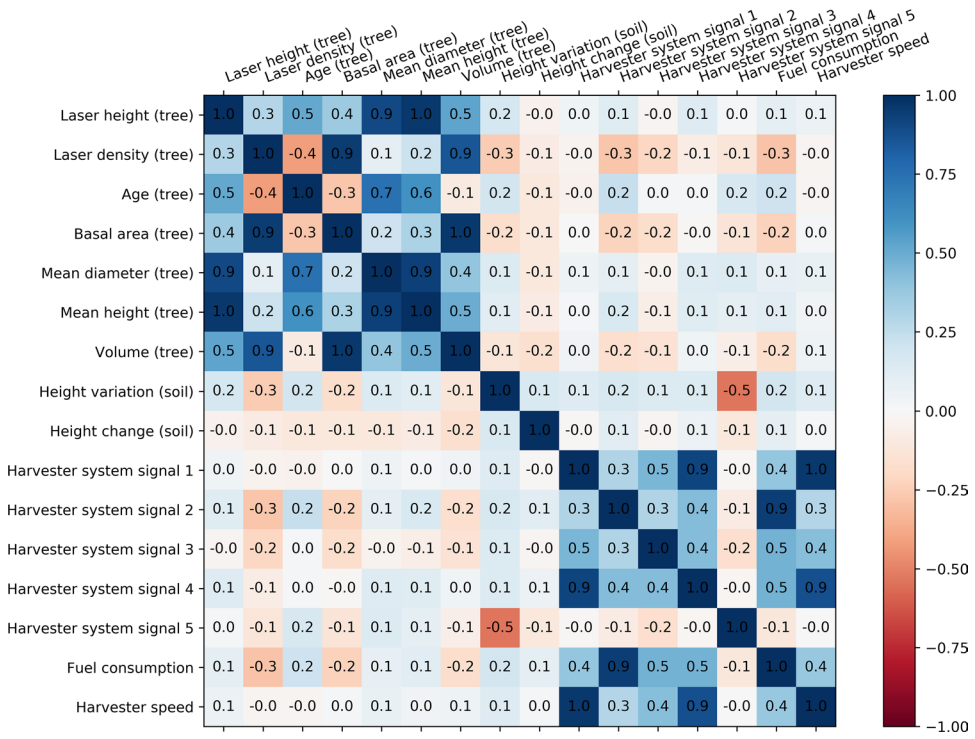


**Fig. 3** A correlation matrix between forest data and harvester CLSE mean values when moving

Figure 4 shows the correlation matrix after the PCA transformation of the forest parameters. The harvester system signal 5, which is related to the cooling of the harvester system, shows the best correlation with principal components three and four of the forest data. The second principal component, consisting of 31% of the total forest data variation, has correlation of 0.4 with the fuel consumption and the harvester system signal 2 (related to the diesel engine of the harvester).

The effect of a single categorical-type forest parameter to each fieldbus signal channel was evaluated by grouping the CLSE's with a single forest parameter at a time and conducting ANOVA or Kruskal–Wallis tests to determine the statistical significance of the resulting differences in the CLSE means. Figure 5 shows boxplots for the sub-episode means when grouped by the soil type. All the presented time series channels show statistically significant differences between some of the groups (significance level 0.05).

## Clustering

Hierarchical clustering methods were applied to the datasets generated from the fieldbus data. The clustering of the harvester data when the harvester was moving divided both the machine and the forest data in a sensible manner, meaning that most of the forest data samples from the same grid cell are in the same cluster. The tolerance for the correct sample division can be relaxed for a certain degree, allowing more clustering groups for the use of further analysis. The CLSE mean values inside the two clusters are distributed according to boxplots shown in Fig. 6.

After the data is divided into clusters, the clusters can be characterized, for example, by the mean value of each forest parameter. Table 3 shows the differences in forest parameters between the clustered groups. In the table, parameters that show statistically meaningful differences according to ANOVA are marked bold. The information in the table gives interpretation for the generated two groups. For example, in the case of moving harvester, the cluster 1 represents dense forest with low changes in ground height level and correspondingly cluster 2 represents grid cells with fewer trees and larger ground height variations. In addition to continuous forest parameters, categorical forest variables were also taken into account after the clustering of the machine signals. Statistical significance tests are not applicable to categorical variables, but the appearance frequency of certain categorical value inside a cluster can be calculated. For instance in the clustering of the harvester driving motion data, the dominant value for the forest parameter *trafficability* was type 1 ("accessible when frost-damaged") in the first cluster and type 2 ("accessible in normal summer conditions") in the second cluster.

In addition to clustering of local field test data, a representative sample of the forest data in Finland was clustered using the *K*-means algorithm. Before clustering, the tree data was transformed to three principal components using PCA. Table 4 shows the explained variance of each principal component and contributions of single forest

**Table 2** PCA results for forest data in field tests

| Principal component | Explained variance (%) | Forest variable contribution (loadings of standardized variables) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Laser height (tree) | Laser density (tree) | Age (tree) | Mean diameter (tree) | Mean height (tree) | Basal area (tree) | Volume (tree) | Ground height variation | Ground height change |
| PC1 | 43.4 | 0.46 | 0.29 | 0.18 | 0.42 | 0.45 | 0.35 | 0.41 | 0.01 | −0.09 |
| PC2 | 31.0 | 0.19 | −0.47 | 0.50 | 0.31 | 0.23 | −0.42 | −0.32 | 0.25 | 0.03 |
| PC3 | 12.3 | 0.10 | 0.08 | −0.19 | −0.05 | 0.01 | 0.07 | 0.04 | 0.56 | 0.79 |
| PC4 | 8.7 | −0.04 | 0.01 | −0.11 | −0.10 | −0.10 | 0.07 | 0.08 | 0.78 | −0.59 |

tree parameters. The three principal components explain together 96.3% of the total forest tree data variance.

Figure 7a shows the clustering result of the forest tree data sampled from 26 million grid cells of different forest areas in Finland with six clusters. The *K*-means clustering was experimented with several different cluster numbers, but the distances of the individual samples from their cluster centers decreased quite steadily while increasing the cluster number, so no single best cluster number was found. The forest tree data from the field tests in Vihti, shown in Fig. 7b, was partitioned according to the six centers.

All the clusters are represented in the field test data, but clusters 1, 5 and 6 have significantly more grid cells than others. Sub-episodes can be grouped according to the clusters, enabling comparison of signal features under general forest types in Finland. Figure 8 shows the differences in harvester fuel consumption and system signal 1 when driving forward in the field tests.

Taking into account the most represented clusters (1, 5 and 6) in the one-way ANOVA, significant differences exist between some of the clusters in the fuel consumption ($p < 0.0001$) and the harvester system signal 1 ($p = 0.006$) CLSE mean values.

## Prediction

For fieldbus signal feature prediction, linear and random forest regression models were trained with the forest data alone as the explaining variables and a mean value of a CLSE channel as the target variable. The prediction results for the moving harvester and forwarder data are collected to Table 5. The length of CLSE was set to 300 samples (6 s) after examining the results with lengths between 50 to 1000 samples as both the harvester and forwarder datasets showed best predictability in average at this window size. However, the prediction of system signal 5, for example, returned better results with the window size of 100 samples, so the optimal window size may be individual for each fieldbus signal.

For visualizing the best prediction performance of the random forest regression, Fig. 9 shows the sorted values for the true and predicted CLSE means of system signal 5 over an independent test set containing 20% of the total dataset values (CLSE window of 100 samples, $R^2 = 0.93$). The most important predictors for the system signal 5, reported by the Python random forest regressor, are the third principal component of the tree data (43.5%), the soil type of class 20 (21.8%) and the second principal component of the tree data (16.9%). In this respect, the system signal 5 is an exception, as all other signals have the first three principal components of the tree data as the most important predictors.

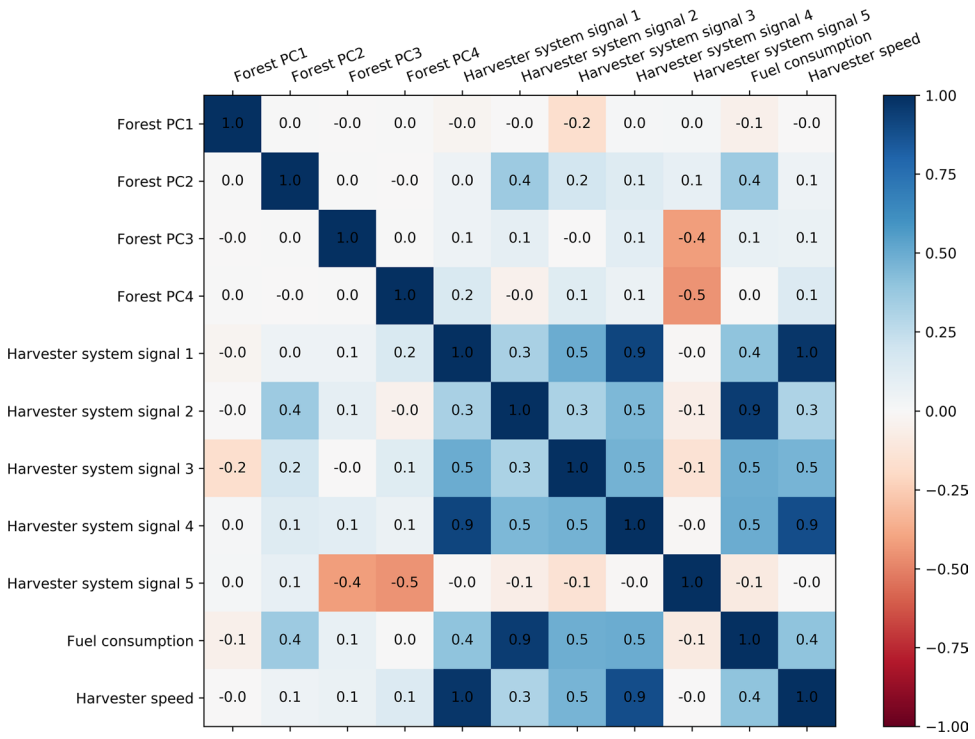


**Fig. 4** Correlations between principal components of the forest parameters and machine signals

## Discussion

The greatest benefit in the use of the forest data platform for predicting the forest machine behavior is that the data is continuously available for the entire forest area in Finland; each site visited can be automatically analyzed in terms of the environment parameters and recorded fieldbus time series. Another advantage is that a model, trained with the forest data and fieldbus measurements from field tests, can predict working conditions for the grid cells in the sites to be visited in future. For example, harvester fuel consumption for the planned route could be estimated by downloading the forest data and predicting the fuel consumption level with the pre-trained model for each grid cell along the route. However, the problem in the predictive setting is that the model requires a comprehensive set of training examples from different forest conditions before reliable predictions can be produced. Here the limiting factor is the availability of the fieldbus data, as the data typically needs to be separately collected in field tests or analyzed directly on the on-board computer of the forest machine. Hence, the possible advantages gained from the comprehensive training data give motivation for continuous collection of fieldbus data.

The relative positioning of the datasets for fusion was based on the GNSS location recorded by the forest machine, by assigning each fieldbus data point to a platform grid cell. The fused dataset contains noise resulting from the forest data, the machine data and the data fusion process. One relevant error source is the location recorded by the GNSS receiver, in particular when the forest canopy is blocking the satellite signal (Kaartinen et al. 2015; Blum et al. 2016). The resulting machine location is somewhat erratic, as can be seen in Fig. 1. The noise in the forest machine path can cause the time series to be positioned off from the correct grid cell in the fusion process, if the location erroneously jumps near the grid cell border. Furthermore, the GNSS receiver records the location from the roof of the machine operator cabinet, so the 8-m-long machine can be mostly on the neighboring cell to that of the cabinet, again resulting in a wrong cell label for the machine. However, typically the forest parameters vary smoothly between neighboring cells, so the error in the machine location does not necessarily lead to large errors in the analysis. Some modern harvesters are capable of determining harvested tree locations using the boom location data, which would help in positioning harvested trees to correct grid cells while working.

The structure of the fused data set was examined with correlation analysis and unsupervised learning techniques, such as PCA, and hierarchical clustering. The dimensionality of the forest data was reduced with PCA, showing that three first principal components carry more than 86% of the variability. The main principal components, based on their loadings of forest tree variables (Table 2), have natural interpretations. The first component differentiates the grid cells based on their total tree mass, as all the loadings have the same sign and increase in each variable implies more tree mass. The second principal component increases with age and mean diameter and decreases when tree density

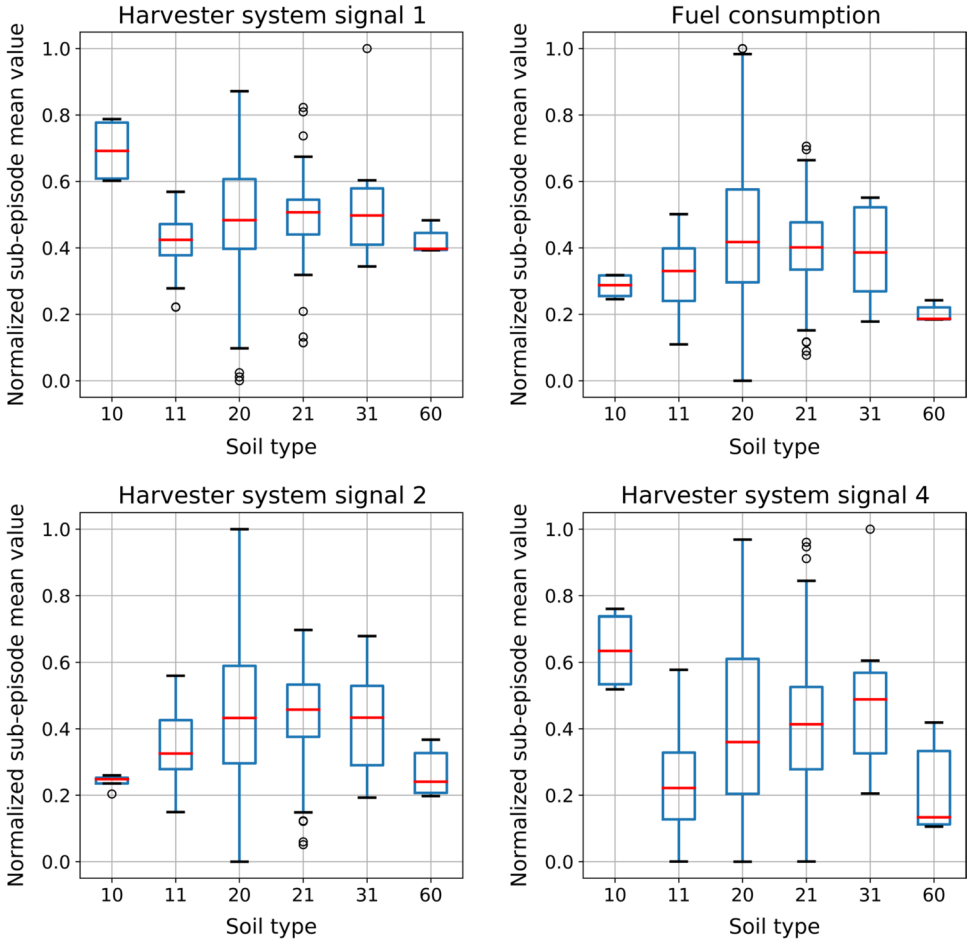


**Fig. 5** Differences in CLSE mean values grouped by the soil type

| Soil type code | Explanation | Sub-episode occurrence |
|---|---|---|
| 10 | Coarse graded moorland | 4 |
| 11 | Coarse graded moraine | 60 |
| 20 | Fine graded moorland | 241 |
| 21 | Fine graded moraine | 115 |
| 31 | Stony coarse graded moraine | 20 |
| 60 | Peat soil | 6 |

and basal area increase. This suggests that the component differentiates dense forests with young and thin trees from sparse forests with old, thick and high trees. The principal components three and four are strongly related to the variation of the ground height: Both components increase with the height variation inside the grid cell, but the third component is related also to increase in the ground height between the cells, suggesting uphill in the movement direction. The hierarchical clustering of the CLSE's mean values revealed that the clusters found in machine signals can be associated to certain grid cells, thus to the attributes of the forest. The hierarchical clustering of sub-episodes reveals how machine behaves in different forest environments within the field test. The number of clusters was selected by requiring that samples from any given grid cell should be assigned to the same cluster. This resulted into division of the sub-episodes to two clusters, but the results could be applicable only to the current field test conditions.

The problem in the generalization of single field test data can be at least partially solved by clustering larger set of the forest data, a representative sample of the Finnish forests, with *K*-means algorithm. Figure 7a shows that the forest data, when projected to the plane of two principal components, forms a single large connected area with few dense regions. Thus the *K*-means algorithm rather divides the area to a given number of sections than finds distinct clusters. Correspondingly to the PCA conducted for the forest data of the field test area (Table 2), the principal components for the forest data of whole Finland (Table 4) have similar interpretations in respect to the tree data. The ground height

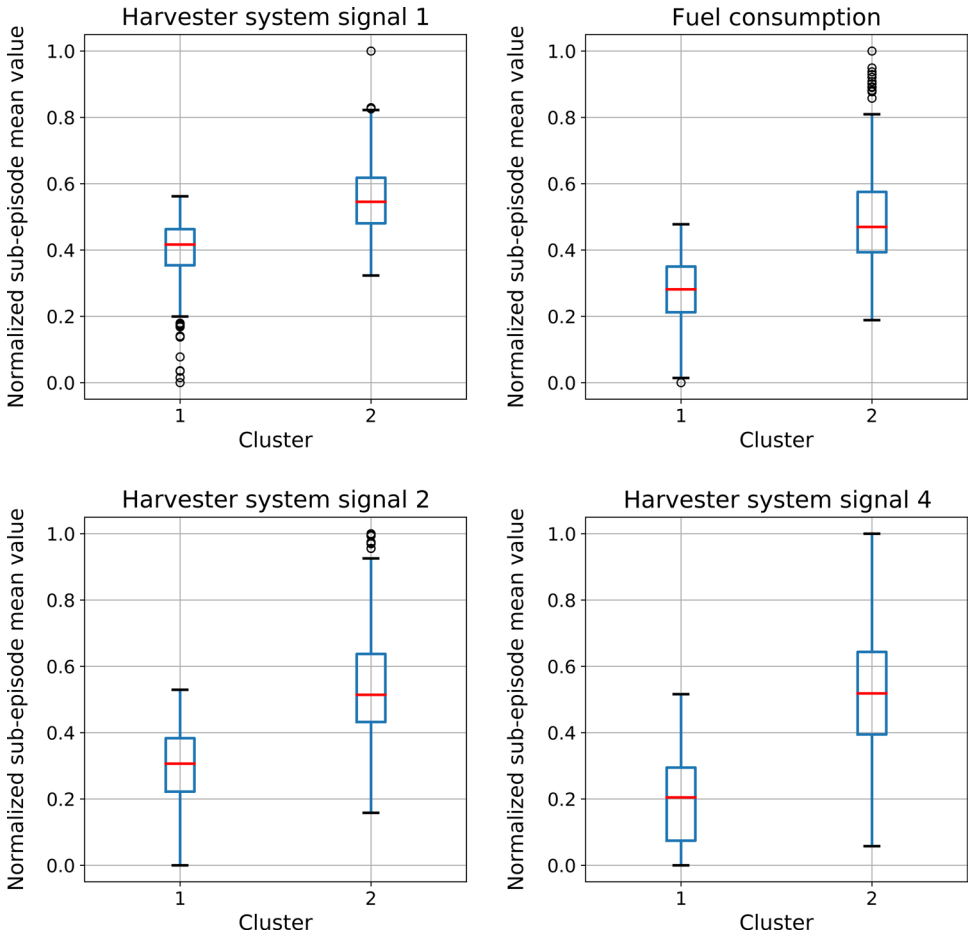


**Fig. 6** Differences in CLSE mean values in four channels (harvester in motion) inside hierarchical clusters

**Table 3** Mean values and standard deviations for forest parameters in clustered groups

| Data type | Cluster no. | Laser density—tree (%) | Age—tree (years) | Basal area –tree ($m^2$/ha) | Stem count—tree (pcs/ha) | Mean diameter—tree (cm) | Volume—tree ($m^3$/ha) | Ground height STD (m) |
|---|---|---|---|---|---|---|---|---|
| Harvester moving | 1 | **64.3 (30.8)** | **31.6 (11.2)** | 15.0 (9.7) | **1658.8 (888.2)** | **13.3 (2.6)** | 97.7 (72.9) | **3.7 (3.1)** |
| | 2 | **53.0 (31.4)** | **35.9 (11.1)** | 13.1 (9.8) | **1292.0 (745.8)** | **14.6 (3.9)** | 93.5 (75.1) | **5.0 (2.1)** |
| Harvester working | 1 | 83.2 (18.1) | 29.1 (7.0) | 21.8 (7.6) | **2173.1 (746.4)** | **13.5 (2.7)** | 147.3 (63.6) | 4.0 (3.0) |
| | 2 | 82.9 (17.4) | 29.5 (7.2) | 21.8 (7.2) | **2100.6 (767.3)** | **13.8 (2.9)** | 149.2 (60.6) | 4.0 (3.0) |
| Forwarder moving | 1 | **75.5 (23.9)** | 29.6 (7.5) | **18.4 (8.4)** | **1848.0 (788.8)** | **13.5 (2.8)** | **122.5 (64.2)** | **3.5 (3.5)** |
| | 2 | **59.8 (28.3)** | 31.8 (13.0) | **14.3 (8.7)** | **1551.5 (741.7)** | **14.0 (4.2)** | **95.8 (70.4)** | **4.8 (2.3)** |

**Table 4** The contributions of single variables to principal components in forest tree data in Finland

| Principal component | Explained variance (%) | Forest tree variable contribution (loadings of standardized variables) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Laser height | Laser density | Age | Mean diameter | Mean height | Basal area | Volume |
| PC1 | 76.0 | 0.41 | 0.27 | 0.35 | 0.4 | 0.41 | 0.38 | 0.4 |
| PC2 | 16.8 | 0.14 | −0.64 | 0.44 | 0.32 | 0.21 | −0.41 | −0.24 |
| PC3 | 3.5 | 0.02 | −0.70 | −0.28 | −0.13 | −0.03 | 0.35 | 0.54 |

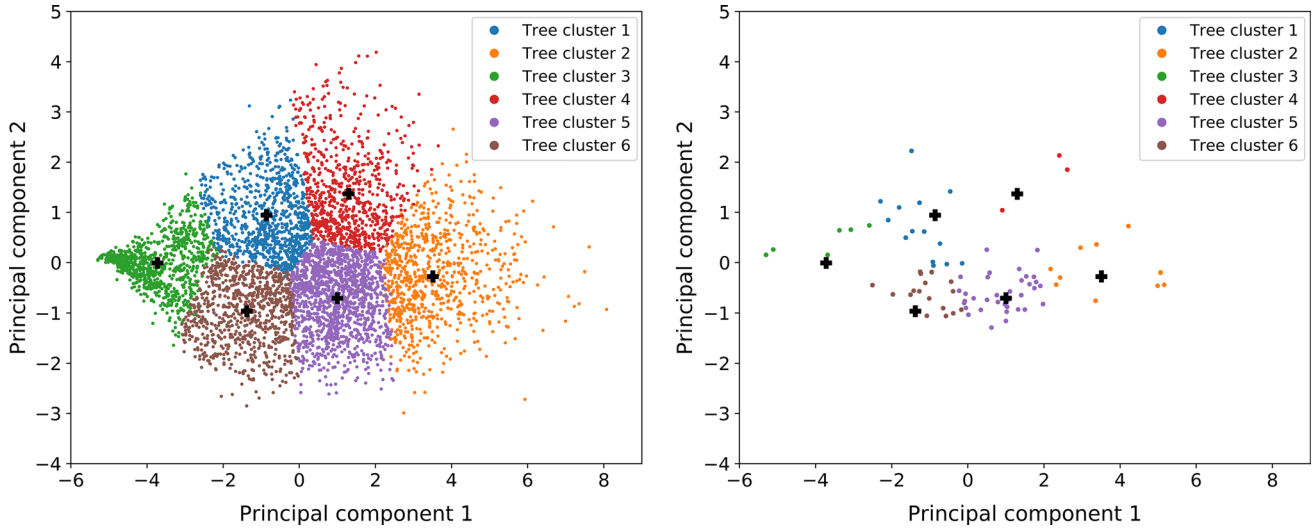


**Fig. 7** Clustered forest tree data points according to the first two principal components in Finland (**a**) and in Vihti field tests (**b**). Only 5% of the total data points are plotted in (**a**) for better visualization of the clusters. Corresponding cluster centers are marked with black crosses

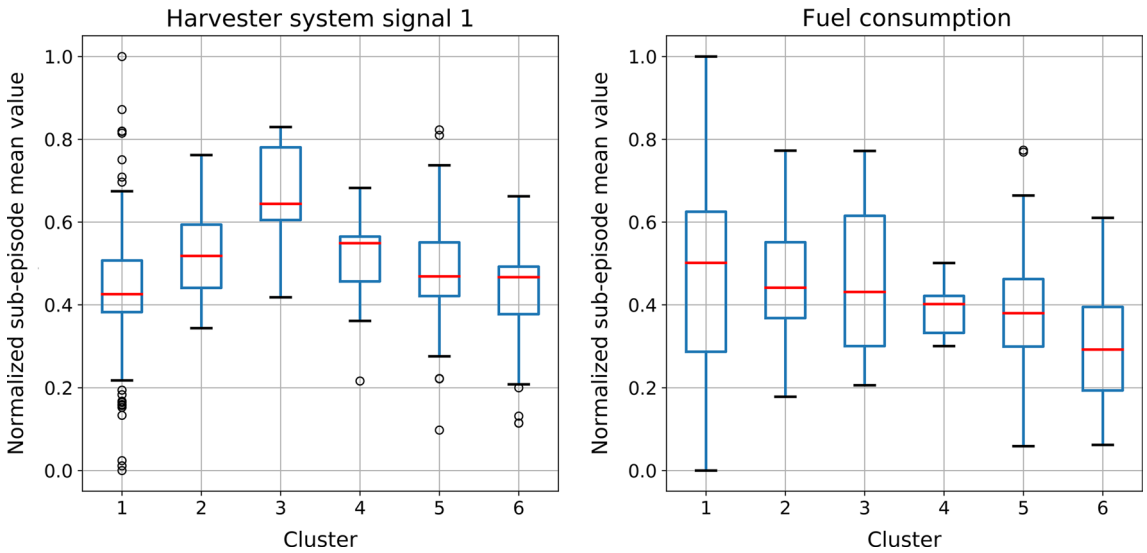


**Fig. 8** Harvester fuel consumption and system signal 1 mean distributions when clustered with six clusters formed from the Finnish forests

**Table 5** Fieldbus signal prediction results for driving motion using linear regression (LR) and random forest regressor (RF) using sample window of 300 samples

| Fieldbus signal channel | $R^2$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Harvester | | Forwarder—total | | Forwarder—1st run | | Forwarder—2nd run | |
| | LR | RF | LR | RF | LR | RF | LR | RF |
| Fuel consumption | 0.10 | 0.49 | 0.16 | 0.46 | 0.34 | **0.71** | **0.58** | **0.86** |
| Speed | 0.07 | 0.35 | 0.15 | **0.66** | 0.16 | **0.70** | 0.17 | **0.74** |
| System signal 1 | 0.09 | 0.43 | 0.16 | **0.67** | 0.18 | **0.74** | 0.17 | **0.73** |
| System signal 2 | 0.14 | **0.61** | 0.15 | 0.45 | 0.31 | **0.68** | **0.58** | **0.86** |
| System signal 3 | 0.19 | **0.53** | 0.08 | 0.49 | 0.15 | **0.69** | 0.28 | **0.72** |
| System signal 4 | 0.15 | 0.49 | 0.26 | **0.74** | 0.24 | **0.82** | 0.34 | **0.81** |
| System signal 5 | 0.45 | **0.83** | **0.53** | **0.64** | 0.62 | **0.88** | **0.69** | **0.95** |

$R^2$ values higher than 0.5 are marked in bold

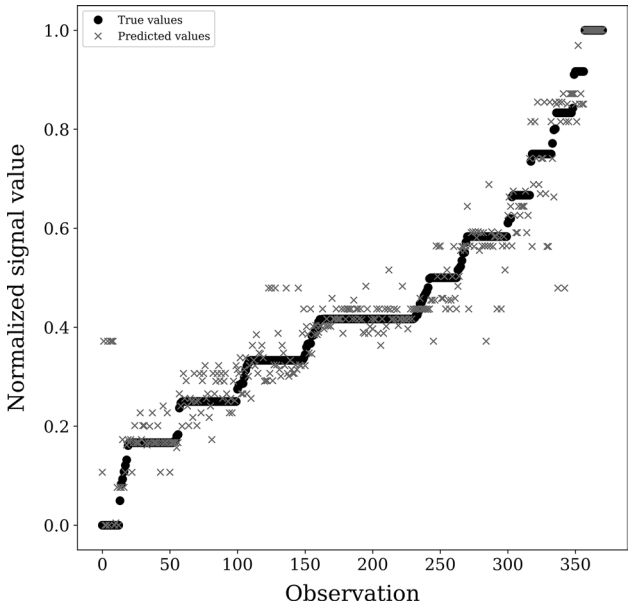


**Fig. 9** Predicted (normalized) mean values of system signal 5 (harvester, CLSE of 100 samples) using only forest data

variation between the grid cells is not applicable here as there is no predefined route in the dataset. In Fig. 7a, the first principal component divides data in four regions and the second component further halves the middle regions. The resulting cluster centers, representing the general forest types in Finland, serve as the basis for robust comparisons of the data from field tests. As show in Fig. 7b, the clusters found are not equally represented at a test location, but a few clusters dominate the dataset. Knowing the general forest types of the single field test helps in interpreting the fusion results from a more general viewpoint, meaning that the next field test results can be compared to current results, if the same forest types are present. For example, a minimum number of observations of the forest type could be necessitated before the data fusion analysis results would be considered comparable in the more general level.

In this study, supervised machine learning was demonstrated by predicting a mean value of a fieldbus time series channel for each grid cell in the dataset using only forest data. The instantaneous values of fieldbus time series depend on numerous factors, but the working environment could hypothetically cause changes in the average time series levels available in the fused dataset. Predicting the mean values beforehand for the grid cells in the target forest site could offer valuable information for route planning, for instance. The prediction results for the traveling forest machines, collected to Table 5, suggest that some variables, in particular *'System signal 5',* were indeed predictable using the forest data alone. This signal is related to the cooling of the forest machine, contrary to other system signals directly related to the transmission and the diesel engine. In this study, a single soil type class (*fine graded moorland*) was found to be important when predicting this signal, so the soil type appears to have an effect on the cooling of the harvester. One useful result for future studies was the observation of the best CLSE length to be 6 s in the fieldbus time series, suggesting that the influence of the forest parameters is on the average best observable in the fieldbus data with this window size. As the random forest regression predicted all the time series levels better than the linear regression, nonlinear relationships between the forest and machine data are of importance. In general, the predictability was higher in the datasets where the route was travelled only once and in the same direction (harvester and individual forwarder runs), suggesting that the direction of which the grid cell is travelled is an essential piece of information. A natural explanation for this is the topography of the grid cell at tilted terrain. Furthermore, the number of runs already driven along a route appears to have an effect and needs to be considered in particular when predicting forwarder performance.

In the current paper, the forest data for the clustering consisted only of the tree attributes without information about the tree species or details of the soil characteristics, such as the soil type, fertility and the load-carrying capacity. The clustering of the general forest types should consider these as well, but this requires dimension reduction methods for joint continuous and categorical data. Furthermore, the fieldbus time series could be more extensive, covering, for instance, machine inclinations and signals related to the tree cutting process of the harvester. The best prediction results were obtained, when the forest machine was constantly traveling forward. The reason for this could be the selection of the fieldbus time series channels, which were more related to the mobility of the harvester than the tree cutting process.

The forest data platform supports new data sources produced by the clients. Here, the forest machine data and the results gained from the data fusion can be thought as a new data source for the platform. Saving the essential analysis results, derived from the machine fieldbus signals, would emphasize more general analysis in the future, after the platform has data from several different forest locations in Finland. Currently, the platform supports addition of the data through CSV files, where a single line represents data for a particular grid cell. One possible data source, utilizing the results found in this paper, would include saving the overall mean of the CLSE's for each grid cell together with the main forest type found in the *K*-means clustering (Fig. 7a). For better usability in future, factors affecting the machine performance, such as the machine type or the operator ID, should be included. In addition, for comparing results inside a single forest operation, the time that the forest machine spends in the grid cell and the main driving direction could be added to the grid cell data.

## Conclusion

The fusion of forest machine fieldbus data and forest inventory data through the geographical location enables machine learning methods for discovering relationships between the forest machine and the environment. In this study, forest data were found to be a good predictor when the machine time series was considered in the grid cell level. In particular, regressions for the data in individual forwarder runs showed $R^2$ values higher than 0.80 with many of the fieldbus signals, and similar results were found with some fieldbus signals from the moving harvester. The study found the grid cells also useful in the clustering of machine signals, as they provided a link between machine data clusters and forest data. The key aspect was the use of the forest data platform, which enables the analysis of the machine and forest inventory data everywhere in Finland. This can accelerate the development of the forest machine and operations, as the effect of the environment can be analyzed without separate field tests for measuring the environmental conditions. The platform can also function as a data warehouse for the shown analysis results. Applications that currently utilize the fieldbus time series directly, like condition monitoring or operator guidance, could benefit from the signals that can be standardized in terms of the environmental effect. The essential constraints in the analyses are the fieldbus time series collection for the data fusion and the generalization of the results with variations in the machine types and operators. The individual results shown in this paper were constructed using fieldbus data from a single field test; although some interesting results were found, more data needs to be gathered from different locations before the results can be generalized.

**Acknowledgements** Luke (Natural Resources Institute Finland) is greatly thanked for arranging the field tests in Vihti and kindly providing the CAN-bus data utilized in this study. Metsäteho Oy and its partners in the Forest Big Data platform project (CGI and Tampere University) are thanked for providing access to the platform and thus enabling automatic retrieval of forest data.



## References

Ala-Ilomäki J, Lamminen S, Sirén M, Väätäinen K, Asikainen A (2012) Using harvester CAN-bus data for mobility mapping. In: Special issue. Abstracts for international conferences organized by LSFRI Silava in cooperation with SNS and IUFRO, vol 25, No. 58, pp 85–87

Blum R, Bischof R, Sauter UH, Foeller J (2016) Tests of reception of the combination of GPS and GLONASS signals under and above forest Canopy in the Black Forest, Germany, Using Choke Ring Antennas. Int J For Eng 27(1):2–14. https://doi.org/10.1080/14942119.2015.1122945

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. https://doi.org/10.1080/14942119.2014.974309

Fardusi MJ, Chianucci F, Barbati A (2017) Concept to practices of geospatial information tools to assist forest management & planning under precision forestry framework: a review. Ann Silvic Res 41(1):3–14. https://doi.org/10.12899/asr-1354

Finnish Forest Center (2019) EServices for forest owners and service providers. https://www.metsaan.fi/. Accessed 1 Apr

Finnish Meteorological Institute (2019) Weather observations database. https://ilmatieteenlaitos.fi/havaintojen-lataus#!/. Accessed 1 Apr

Gülci N, Akay AE, Erda O, Wing MG, Sessions J (2015) Planning optimum logging operations through precision forestry approaches. Eur J For Eng 1:56–60

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. https://doi.org/10.1080/14942119.2016.1170495

Hämäläinen J (2016) Kohti Puuhuollon Digitalisaatiota—forest big data—Project. Metsätehon Tuloskalvosarja 11: http://www.metsateho.fi/kohti-puuhuollon-digitalisaatiota/. Accessed 1 Apr

Han H-S, Page-dumroese D, Han S-K, Tirocke J, Page-dumroese D (2006) Effects of slash, machine passes, and soil moisture on penetration resistance in a cut-to-length harvesting. Int J For Eng 17(2):11–24. https://doi.org/10.1080/14942119.2006.10702532

Holopainen M, Vastaranta M, Hyyppä J (2014) Outlook for the next generation's precision forestry in Finland. Forests 5(7):1682–1694. https://doi.org/10.3390/f5071682

Jain AK (2010) Data clustering: 50 years beyond *K* means. Pattern Recognit Lett 31(8):651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jones E, Oliphant T, Peterson P (2019) SciPy: open source scientific tools for python—normaltest. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html. Accessed Apr 1

Kaartinen H, Hyyppä J, Vastaranta M, Kukko A, Jaakkola A, Xiaowei Yu, Pyörälä J et al (2015) Accuracy of kinematic positioning using global satellite navigation systems under forest Canopies. Forests 6(9):3218–3236. https://doi.org/10.3390/f6093218

Lindroos O, Ringdahl O, La Hera P, Hohnloser P, Hellström T (2015) Estimating the position of the harvester head—a key step towards the precision forestry of the future? Croat J For Eng 36(2):147–164

Lu K, Bi H, Watt D, Strandgard M, Li Y (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(1):13–33. https://doi.org/10.1007/s11676-017-0517-1

Luke Natural Resource Institute of Finland (2019) National forest inventory (NFI). http://www.metla.fi/ohjelma/vmi/info-en.htm. Accessed 1 Apr

Mason, EG, Morgenroth JA, Bown HE (2016) Precision forestry research project—final report. University of Canterbury Research Repository. https://ir.canterbury.ac.nz/handle/10092/13431. Accessed 1 Apr

Melander L, Ritala R (2018) Time-of-flight imaging for assessing soil deformations and improving forestry

vehicle tracking accuracy. Int J For Eng 29(2):63–73. https://doi.org/10.1080/14942119.2018.1421341
National Land Survey of Finland (2019) Topographic Database. https://www.maanmittauslaitos.fi/en/maps-and-spatial-data/expert-users/product-descriptions/topographic-database. Accessed Apr 1
Obi OF, Visser R (2017) Influence of the operating environment on the technical efficiency of forest harvesting operations. Int J For Eng 28(3):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(4):3–7
Pedregosa F, Weiss R, Brucher M (2011) Scikit-learn machine learning in python. J Mach Learn Res 12:2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2
Prinz R, Spinelli R, Magagnotti N, Routa J, Asikainen A (2018) Modifying the settings of CTL Timber harvesting machines to reduce fuel consumption and $CO_2$ emissions. 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. Metsäteho Rep, No. 240. http://www.metsateho.fi/data-platform-promoting-forest-data-utilization/. Accessed 1 Apr
Salmivaara, A, Launiainen S, Ala-Ilomäki J, Kulju S, Laurén A, Sirén M, Tuominen S, et al (2017) Dynamic forest trafficability prediction by fusion of open data, hydrologic forecasts and harvester-measured data. Young Researchers Challenge 2017. Stockholm: Poster, 1. http://urn.fi/URN:NBN:fi-fe2017102450272. Accessed 1 Apr
Salmivaara A, Miettinen M, Finér L, Launiainen S, Korpunen H, Tuominen S, Heikkonen J et al (2018) Wheel rut measurements by forest machine-mounted lidar sensors—accuracy and potential for operational applications? Int J For Eng 29(1):1–12. https://doi.org/10.1080/14942119.2018.1419677
Seber GAF (1984) Hierarchical clustering: agglomerative techniques. Multivariate observations. Wiley, New York, pp 359–375
Sirén M, Salmivaara A, Ala-Ilomäki J, Launiainen S, Lindeman H, Uusitalo J, Sutinen R, Hänninen P (2019) Predicting forwarder rut formation on fine-grained mineral soils. Scand J For Res 7581:1–10. https://doi.org/10.1080/02827581.2018.1562567
Strandgard M, Walsh D, Acuna M, Strandgard M, Walsh D, Estimating MA, Strandgard M, Walsh D, Acuna M (2013) Estimating harvester productivity in pinus radiata plantations using StanForD stem files. Scand J For Res 28(1):73–80. https://doi.org/10.1080/02827581.2012.706633
Suvinen A, Saarilahti M (2006) Measuring the mobility parameters of forwarders using gps and CAN bus techniques. J Terrramech 43(2):237–252. https://doi.org/10.1016/j.jterra.2005.12.005
Venäläinen P, Räsänen T, Hämäläinen J (2015) Potential business models for forest big data—data to intelligence (D2I) project report, Task 3.2. Metsäteho Report, No. 235. http://www.metsateho.fi/potential-business-models-for-forest-big-data-data-to-intelligence-d2i-project-report-task-3-2/. Accessed 1 Apr
Xu R, Wunsch D (2008) Hierarchical clustering. Piscataway, Wiley, pp 31–62