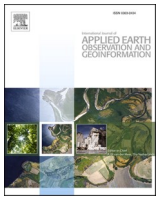
International Journal of Applied Earth Observations and Geoinformation 105 (2021) 102624

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Prediction of butt rot volume in Norway spruce forest stands using harvester, remotely sensed and environmental data

Janne Räty *, Johannes Breidenbach *, Marius Hauglin, Rasmus Astrup

Norwegian Institute of Bioeconomy Research (NIBIO), Høgskoleveien 8, 1433 Ås, Norway

ARTICLE INFO

Keywords:
Cut-to-length harvester data
Forest health
Heterobasidion spp.
Wood decay
Lidar

ABSTRACT

Butt rot (BR) damage of a tree results from a decay caused by a pathogenic fungus. BR damages associated with Norway spruce (*Picea abies* [L.] Karst.) account for considerable economic losses in timber production across the northern hemisphere. While information on BR damages is critical for optimal decision-making in forest management, maps of BR damages are typically lacking in forest information systems. Timber volume damaged by BR was predicted at the stand-level in Norway using harvester information of 186,026 stems (clear-cuts), remotely sensed, and environmental data (e.g. climate and terrain characteristics). This study utilized Random Forests models with two sets of predictor variables: (1) predictor variables available after harvest (theoretical case) and (2) predictor variables available prior to harvest (mapping case). Our findings showed that forest attributes characterizing the maturity of forest, such as remote sensing-based height, harvested timber volume and quadratic mean diameter at breast height, were among the most important predictor variables. Remotely sensed predictor variables obtained from airborne laser scanning data and Sentinel-2 imagery were more important than the environmental variables. The theoretical case with a leave-stand-out cross-validation resulted in an RMSE of 11.4 $m^3 \cdot ha^{-1}$ (pseudo-R^2: 0.66) whereas the mapping case resulted in a pseudo-R^2 of 0.60. When spatially distinct clusters of harvested forest stands were used as units in the cross-validation, the RMSE value and pseudo-R^2 associated with the mapping case were 15.6 $m^3 \cdot ha^{-1}$ and 0.37, respectively. The findings associated with the different cross-validation schemes indicated that the knowledge about the BR status of spatially close stands is of high importance for obtaining satisfactory error rates in the mapping of BR damages.

1. Introduction

Butt rot (BR) damages associated with coniferous forests account for considerable economic losses in the forestry sector of the northern hemisphere. BR damages are especially destructive in forests dominated by Norway spruce (*Picea abies* [L.] Karst). For example, it has been observed that 25 % of Norway spruce stems had BR damages in final fellings according to a Norwegian nation-wide stump survey (Huse et al., 1994). The most destructive fungus genus in coniferous forests is *Heterobasidion* spp. that comprises several species with varying host preferences, and *Heterobasidion parviporum* especially prefers Norway spruce as a host tree species. In total, *Heterobasidion spp*. causes an annual economic loss of approximately 800 million euros in Europe alone (Hodges, 1999).

BR infections can spread from an infected tree to a healthy tree via airborne fungus spores that occupy a new contamination surface, such as a fresh stump. The BR infection can also enter to a new host tree via below-ground root connections (Aosaar et al., 2020; Stenlid, 1987). The spread of BR infection is also dependent on the characteristics of forest stands. It has been found that the risk of infection is lower in stands with a mixture of Norway spruce and Scots pine (*Pinus sylvestris* [L.]) than in pure Norway spruce forest (Möykkynen and Pukkala, 2010). Tree attributes, such as diameter at breast height (DBH) and age, are also linked to the risk of the BR damages (Hylen and Granhus, 2018; Mattila and Nuutinen, 2007). Hylen and Granhus (2018) found that the risk of BR damage in Norway spruce increases with DBH up to a DBH of 30 cm. They also found that the risk of BR damage increases in terms of age, but the probability of damage is relatively stable for trees older than 80 years. BR damages have frequently been found on calcareous, limestone-rich and fertile soil types, and it has been suggested that a thick peat layer prevents the risk of the RB damages to some extent (Müller et al. 2018).

Cut-to-length harvesters collect tree-level data during harvest operations. Harvester datasets have been used with remotely sensed material

* Corresponding authors.
E-mail addresses: janne.raty@nibio.no (J. Räty), johannes.breidenbach@nibio.no (J. Breidenbach).

https://doi.org/10.1016/j.jag.2021.102624
Received 1 July 2021; Received in revised form 29 October 2021; Accepted 11 November 2021
Available online 18 November 2021

and other auxiliary datasets to produce or validate forest resource information. Previous studies have used harvester data, for example, for the modeling of diameter distributions (Maltamo et al., 2019; Söderberg et al., 2021), the prediction of sawlog volume (Peuhkurinen et al., 2008) or other forest attributes (Söderberg, 2015; Hauglin et al., 2018), and the validation of forest attribute maps (Vähä-Konka et al., 2020). A major restriction related to the operational application of harvester data is the lack of standardized technology for the positioning of the trees (Hauglin et al., 2017; Kemmerer and Labelle, 2020), which is essential if harvester data are a surrogate for conventional field measurements.

The quality characteristics associated with standing timber resources are laborious to assess in the field. It is possible to detect quality attributes such as branchiness or crookedness (Karjalainen et al., 2019), but the accurate detection of BR damages by a visual inspection of standing trees is practically impossible. Therefore, harvester data can be important sources of quality characteristics associated with timber resources. Harvester data provide information on timber assortments at the level of logs cut from individual trees. Typically, the commercial timber assortments are sawlog, pulpwood, energy wood. Timber damaged by BR is typically allocated to the pulpwood or energy wood assortments, which results in economic losses especially in mature forest stands. In order to separate healthy and infected logs, visual inspections of cross-cutting surfaces during harvest operation must be carried out.

Currently, forest owners cannot accurately evaluate economic loss caused by BR damages in Norway. Thus, our objective was to map timber volume with BR damages (henceforth BR volume) in spruce-dominated forests using harvester data, remotely sensed data, and environmental variables. The remotely sensed data comprised airborne laser scanning (ALS) data and Sentinel-2 satellite imagery whereas the environmental variables consisted of climate and terrain variables, and site-specific characteristics indicating, for example, growing conditions, and geographical position. To the best of our knowledge, the application of harvester data in the mapping of BR damages jointly with remotely sensed data, has not been studied so far.

2. Material and methods

2.1. Study area

The study area is located between the latitudes of 59° and 65°, and the longitudes of 8° and 12° in Norway. The large latitudinal range, changes in the distance to the coastline, and elevation shifts affect the growing conditions across the study area (Fig. 1). The high-altitude (above sea-level) mountain forests were not of interest, since the focus was on the operationally accessible forests that are under commercial timber production. The mean altitude associated with the forests of interest was 300 m whereas the maximum altitude was 900 m. Norway spruce and Scots pine are the most common tree species in the area of interest. Broadleaved species, mostly birch (*Betula spp.* [L.]), are typically growing as mixtures among the coniferous species.

2.2. Harvester data

The harvester data consisted of 323,724 trees and were collected in 2020 and 2021 using five different harvesters. The harvester data only included trees from clear-cut stands. Few retention trees were usually left in the clear-cut stands, but they were not registered. Further statistics associated with the harvester data are reported in Section 2.4.3.

The harvester data comprised information on each harvested tree. The trees were bucked during the harvest operation into the commercial timber product categories sawlog, pulpwood and energy wood. The harvester sensors recorded diameter measurements along stem, product lengths and product volumes (Nordström and Hemmingsson, 2018). Diameter at breast height (DBH) was estimated by the harvester's computing system based on an approximated stump height and the tapering of stem based on diameter measurements in 10 cm intervals along the stem.

Harvester operators manually determined the bucking of BR-damaged stems based on the severity of BR damage visible at the cross-cuttings. For this dataset specifically, the harvester operators manually recorded BR damages at each cross-cutting for each Norway spruce stem. The products damaged by BR were categorized into the products BR pulpwood, BR energy wood, and BR cut-off. The harvester data were stored in the Standard for Forest machine Data and communication (StandForD2010) format (Arlinger et al., 2012).

BR volume was calculated for each Norway spruce stem based on the damaged stem product or products. Damaged stems usually comprise both damaged and healthy timber products.

2.3. Remotely sensed and environmental data

ALS data covering the study area were collected in several flight campaigns between 2010 and 2018. The flight parameters were not identical among the data acquisitions and the resulting mean point densities varied between 2 and 5 points per square meter among the ALS campaigns. A digital terrain model (DTM, 1 m × 1 m) was created using the last returns of the ALS data (Kartverket, 2019). The DTM was subtracted from the orthometric height measurements of the ALS data to normalize them. The height-normalized ALS data were overlaid on the 16 m × 16 m grid cells and the following ALS features were calculated based on first-of-many and only echoes for the cells of the Norwegian forest resource map (SR16 map) (Hauglin et al., 2021): mean, variance, proportion of echoes above 2 m, and percentiles (25th, 50th, 75th, 90th, and 95th).

A mosaic of Sentinel-2 satellite imagery (Level 2A product) was obtained from the SentinelHub (Kirches, 2018). The mosaic was based on the images acquired in 2018. The medoid method (Flood, 2013) was used to obtain the most representative ground surface pixel value within each month, thus avoiding clouds and haze. After preliminary modeling attempts, we only used the B8 (near-infrared) band which has a spatial resolution of 10 m. Finally, the mosaic was resampled using bilinear interpolation to 16 m × 16 m grid cells.

The environmental variables were collected from several existing maps. Temperature sum (TS_{CLI}), precipitation (PS_{CLI}), altitude (AL_{CLI}), a terrain variable describing slope (SL_{TER}) and distance from the coastline (DC_{CLI}) were collected from the existing nation-level maps created for research purposes. The temperature sum was calculated as a sum of monthly mean temperatures (mean temperature threshold >5° Celsius) in the time period of 1989–2018. The precipitation was calculated as a sum of monthly precipitation of the months with mean temperature >5° Celsius in the same time period than the temperature sum.

The temperature and precipitation data were available for the study area in a 1 km × 1 km resolution and the temperature observations were adjusted using elevation when resampling to 16 m × 16 m grid cells (Skaugen et al., 2002). Altitude and slope data were calculated from a 10 m × 10 m resolution national digital terrain model provided by the Norwegian Mapping Authority. The terrain model was created using the ALS data collected in a national scanning campaign. Distance from the coastline was based on a 100 m resolution map produced by The Norwegian Water Resources and Energy Directorate. In this study, we resampled the raster data to match with the grid cells of the SR16 map (16 m × 16 m).

The soil characteristics (ST_{AR5}), and forest type (FT_{AR5}) classification were collected from the Norwegian national land resource map (AR5 map) (Ahlstrøm et al., 2019). The ST_{AR5} layer separates mineral soils from the organic soils, and the FT_{AR5} layer describes the composition of stands as coniferous, deciduous, or mixed. The AR5 map is originally in a vector format and was converted to raster files that align with the 16 m × 16 m grid cells. We also used the SR16 site index (BON_{SR16}) map which is based on a model fitted using site index values, recorded in the Norwegian national forest inventory field plots, as response and various environmental variables including depth to water, soil properties and

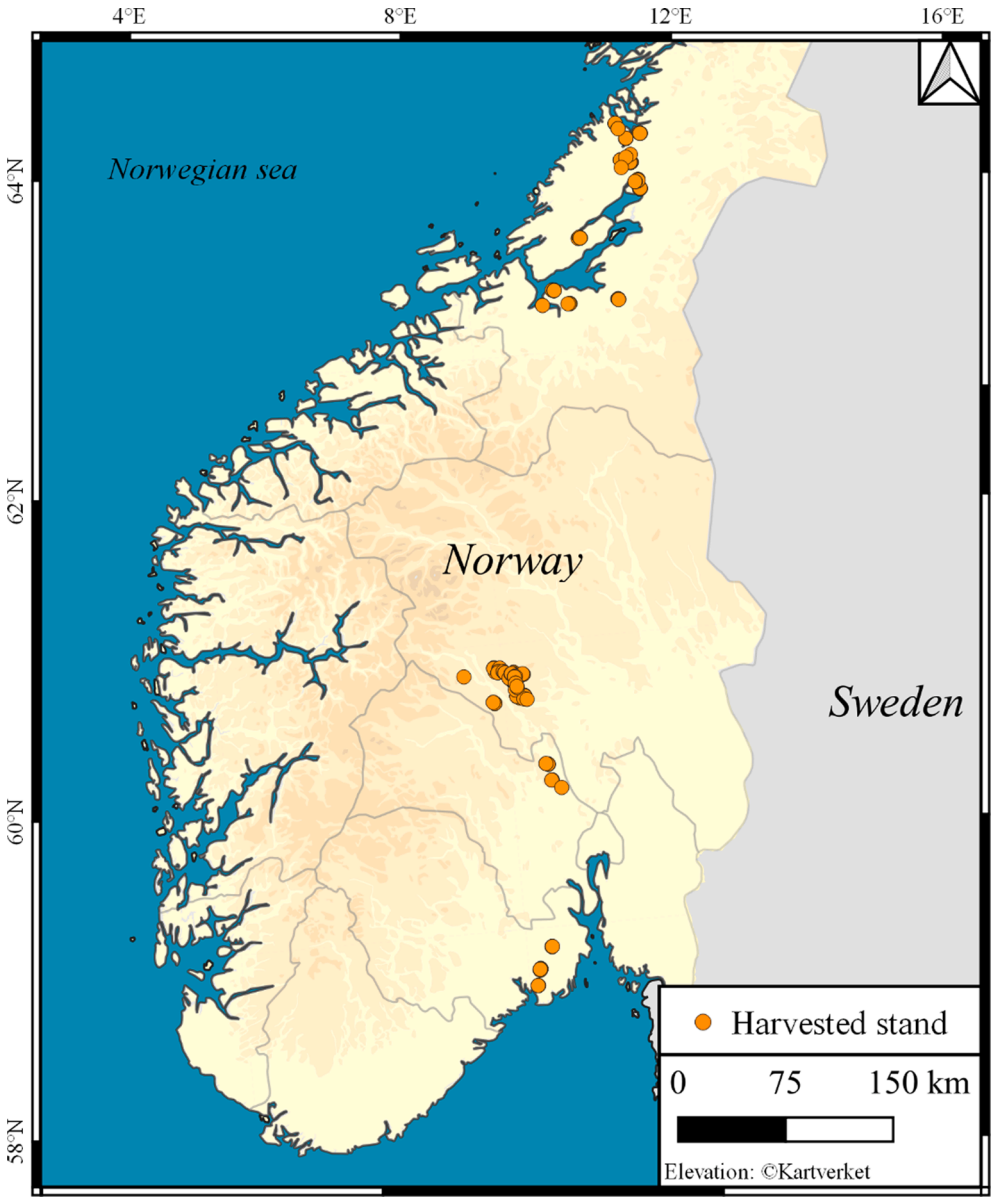


Fig. 1. Study area and the locations of the harvested forest stands.

altitude as predictor variables (Astrup et al., 2019).

We also collected characteristics from the soil map (SOIL) provided by Geological Survey of Norway (2020). The soil map describes the soil type more specifically than the ST_{AR5} map and has among others the categories moraine and old sea-floor soil types that indicate high soil pH. The soil map is originally in a vector format and was converted to raster files that align with the 16 m × 16 m grid cells.

2.4. Data preparation, modeling and validation

2.4.1. Study workflow

The methodology consisted of five steps: (1) post-processing of harvester data (2) the delineation of harvested stands, (3) the calculation of BR volume and predictor variables for the harvested stands, (4) the modeling of stand-level BR volume and (5) mapping and model validation using leave-stand-out and leave-cluster-out cross validation. The workflow is visualized in Fig. 2 and the steps are explained in detail in the next sections.

2.4.2. Post-processing of tree locations

The harvesters were equipped with a global navigation satellite system (GNSS) receiver that registered the machine's location during harvesting operation. Three harvesters were equipped with a positioning system that determined the XY location of a harvester head, which resulted in more accurate tree positions (52 % of the trees) than machine-based positioning. The XY locations of the harvested trees that were not positioned based on the harvester head followed stripe patterns in the harvested stands (Fig. 2, step 1). In order to better distribute the tree locations for the delineation of harvested stands, we added a uniformly distributed random value of ± 8 m to the XY coordinates of the machine to simulate the position of the harvester head. A preliminary analysis showed that the post-processing step of machine-based tree locations improved the delineation of stands from the harvester data.

2.4.3. Delineation of harvested stands

The SR16 map includes stand-like segments in forests, and we used the SR16 segments as the base information for the stand delineation (Fig. 2, step 2). The SR16 segmentation is based on canopy height, as well as site index, and tree species composition predicted using remotely

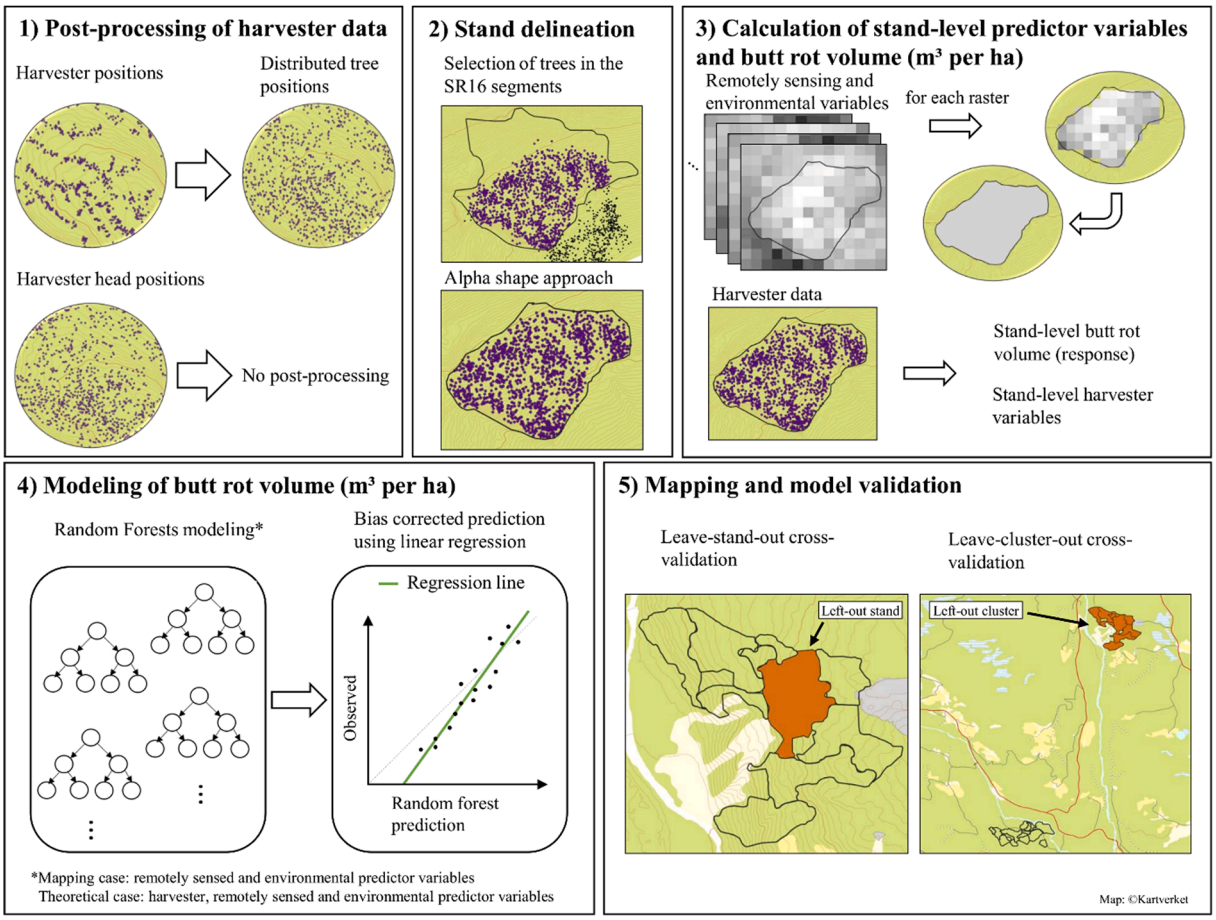


Fig. 2. Study workflow.

sensed and national forest inventory data (Astrup et al., 2019).

The harvested areas did not always match with the SR16 segments (Fig. 2, step 2) and therefore, the XY locations of the harvested trees were utilized to crop the SR16 segments. The harvested trees were attributed to the SR16 segments using their XY locations. The XY locations of trees were used to create two-dimensional alpha shapes (n = 667) using the *alphahull* package (Pateiro-Lopez and Rodriguez-Casal, 2019) in the R environment (R Core Team 2021). A buffer of two meters was added on the boundary of each alpha shape to account for the distance between crown edge and stem position. Finally, the SR16 segments were cropped by overlaying the alpha shapes to create segments of the harvested area. Henceforth, we refer to these cropped SR16 segments as *harvested stands*. The alpha parameter associated with the alpha shape approach was set at 25. Harvested stands with an area less than 0.3 ha, with less than 30 harvested trees, and mixed stands with a volume proportion of Norway spruce less than 50 % were removed. Altogether, 256 spruce-dominated harvested stands with a total area 326 ha comprising 186,026 harvested stems were available for modeling BR rot volume. Statistics associated with the harvested stands are in Table 1.

Table 1
Statistics associated with the harvested stands. Note that butt rot damages were only registered for Norway spruce stems.

	Mean	Standard deviation	Minimum	Maximum
Response variable: harvested volume damaged by butt rot ($m^3 \cdot ha^{-1}$)	23.9	19.7	0.0	134.7
Proportion of harvested volume damaged by butt rot (%)	11.4	7.2	0.0	37.4
Harvested volume ($m^3 \cdot ha^{-1}$)	216.3	114.7	40.9	703.0
Quadratic mean diameter (cm)	22.0	3.3	14.2	32.3
Number of harvested stems (stems $\cdot$ ha^{-1})	743	260	203	1709
Number of harvested stems per stand (stems $\cdot$ ha^{-1})	727	617	108	4471
Harvested volume of Norway spruce (%)	90.0	11.6	52.8	100.0
Stand area (ha)	1.0	0.8	0.3	5.1

2.4.4. Calculation of butt rot volume and predictor variables for harvested stands

Rasters of remotely sensed and environmental variables with 16 m × 16 m grid cells were linked to the harvested stands by aggregating the grid cells with centers inside the boundaries of a harvested stand (Fig. 2, step 3). Continuous variables were aggregated as means whereas the categorical variables were aggregated as mode values.

Predictor variables calculated from the ALS and Sentinel-2 data are referred to as *remotely sensed variables*. The variables calculated from the existing maps (e.g., SR16, AR5, SOIL) will be referred to as *environmental variables*. A summary of variables is presented in Table 2.

The stand-level response variable (BR volume) was calculated using the stem product information associated with the harvested trees. We also computed the following predictor variables from the harvester data: quadratic mean diameter of Norway spruce stems (QMD_{HRV}), harvested timber volume per hectare (V_{HRV}), the proportion of harvested Norway spruce volume, and the width of the DBH distribution (DR_{HRV}). The width of harvested DBH distribution was determined as the difference between the 90th and 10th percentiles of the DBH distribution. We henceforth refer to the predictor variables computed from the harvester

Table 2
Predictor variables calculated for the harvested stands.

Available after harvest		Available prior to harvest			
Harvester variables		Remotely sensed variables		Environmental variables	
Variable(s)	Description	Variable(s)	Description	Variable(s)	Description
V_{HRV}, N_{HRV}	Harvested timber volume (V_{HRV}, $m^3 \cdot ha^{-1}$) and the number of harvested stems (N_{HRV}, stems $\cdot$ ha^{-1})	$Hmean_{ALS}$, $Hvar_{ALS}$	Mean, and variance associated with the height measurements of ALS data	AL_{CLI}, SL_{TER}, TS_{CLI}, PS_{CLI}, DC_{CLI}	Altitude above sea level, slope, temperature sum (TS_{CLI}), precipitation (PC_{CLI}), and distance to coast (DC_{CLI})
QMD_{HRV}	Quadratic mean DBH of harvested Norway spruce stems (cm)	HP_{ALS}	Percentiles associated with the distributions of height measurements of ALS data. P = {25, 95 %}	BON_{SR16}	Site index extracted from the Norwegian forest resource map SR16
DR_{HRV}	Difference between the 90th and 10th percentiles of the DBH distribution. (cm)	$D2_{ALS}$	Proportion of ALS height measurements (first echoes) above a threshold of 2 m	FT_{AR5}, ST_{AR5}, SOIL	Forest type (FT_{AR5}) and soil type (ST_{AR5}) extracted from the Norwegian land resource map (AR5). The SOIL variable describes geological soil characteristics and was extracted from the map provided by the Geological Survey of Norway
SPP_{HRV}	Proportion of harvested timber volume of Norway spruce (%)	NIR_{S2}	Optical image variables extracted from the Sentinel-2 image mosaic. The following band was used: Near-infrared (NIR, band 8)	X, Y	X and Y coordinates of the centroids of the forest stands

Note: HRV – harvester variable, ALS – airborne laser scanning, S2 – Sentinel-2, CLI – climate variable, TER – terrain variable, SR16 – Norwegian forest resource map, AR5 – Norwegian national land resource map.

data as *harvester variables*.

The harvester, remotely sensed and environmental variables were further categorized into two categories which were used to train two separate models: i) All predictor variables. This is the *theoretical case* with the availability of observed information on forest attributes because the harvester variables are available only after harvest. ii) Predictor variables available prior to harvest (i.e. harvester variables excluded). This is the *mapping case* because these variables are available from the existing forest attribute maps before any harvest.

2.4.5. Modeling butt rot volume

We used the Random Forests (RF) regression method (Breiman, 2001) to model and subsequently map BR volumes at the stand-level. RF is a widely used non-parametric and non-linear approach which is based on classification and regression trees (CARTs). The RF method also enables the tracking of variable importances, which is an useful feature in the interpretation of models with numerous predictor variables. We used the RF implementation of the *randomForest* package (Liaw and Wiener, 2002) in the R environment. RF is controlled by three hyperparameters which determine the number of decision trees (*ntree*), the number of predictor variables selected in each node splitting (*mtry*) and the depth of a tree (*nodesize*). The hyperparameter values were fixed at their defaults of 500 and 5 for *ntree* and *nodesize*, respectively. The hyperparameter *mtry* is dynamically determined as the number of predictor variables divided by three brought up to a round integer. Preliminary analysis showed that changes in the hyperparameter settings only marginally affected the results.

It has been observed that RF regressions tend to overestimate small observations and underestimate large observations (Zhang and Lu, 2012). We reduced this prediction bias using a simple linear regression approach which relates observed values with RF predictions (Song, 2015). That means, our final predictions are based on an ordinary least squares regression model with the observed values as the response and the RF predictions as the only predictor variable (Fig. 2, step 4).

2.4.6. Validation and performance assessment

We applied two different validation strategies in the performance assessment of the RF models (Fig. 2, step 5). In order to study the importance of close-by training data for the predictive performance, we used k-means clustering to create geographically independent groups of harvested stands. The k-means clustering was carried out using the *stats* R-package (R Core Team, 2021). Only clusters with five or more harvested stands were allowed which resulted in 23 clusters. A distance to the center of the nearest cluster was on average 10 km, at minimum 0.5 km, and at maximum 57 km. There were on average 11 harvested stands per cluster. The resulting clusters were used to carry out a leave-cluster-out cross validation (ClusterCV). We also carried out a leave-stand-out cross validation (StandCV), which allows the inclusion of the geographically neighboring stands in the training data of the RF model.

In addition to the cross-validation strategies, we also evaluated the estimation of BR volume based on harvested stand-level timber volume and the mean proportion of harvested volume damaged by BR in the training data. This evaluation strategy is referred to as a null-model and shows the level of error that can be achieved when only the mean proportion of volume damaged by BR (per stand) in the study area is known.

We evaluated the predictive performance of the models using a pseudo-coefficient of determination (R^2):

$$Pseudo - R^2 = 1 - MSE \Big/ \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \quad (1)$$

where $\mathrm{MSE} = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2}{n}$ is the mean squared error, and y_i and $\widehat{y}_i$ are observed and predicted BR volumes in stand i, n refers to the number of harvested stands, and $\bar{y}$ is the mean of observed BR volume over all stands. For simplicity, we will refer to the pseudo-R^2 value as R^2.

The errors associated with predicted BR volume predictions were evaluated using the root-mean-square error (RMSE, Eq. (1)) and mean difference (MD, Eq. (2)). The relative error is the absolute error divided by the observed mean of the response multiplied by 100.

$$RMSE = \sqrt{MSE} \quad (2)$$

$$MD = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)}{n} \quad (3)$$

3. Results

3.1. Prediction of butt rot volume

Two cross-validation strategies, namely StandCV and ClusterCV, were employed in order to evaluate the predictive performance of the models. The latter was used to evaluate the importance of close-by reference observations on the models' predictive performances. In addition to StandCV and ClusterCV, the null-model was utilized to evaluate the achievable predictive performance given that only the

Table 3
Root-mean-square errors (RMSE), mean differences (MD) and pseudo-R^2 (R^2) values associated with predicted volume damaged by butt rot using different sets of predictor variables and three different strategies to evaluate the predictive performance. The harvested stands were used as modeling units. CV – cross-validation, ClusterCV – Leave-cluster-out CV, StandCV – Leave-stand-out CV

Evaluation strategy	Predictor variables	RMSE ($m^3 \cdot ha^{-1}$)	MD ($m^3 \cdot ha^{-1}$)	RMSE (%)	MD (%)	R^2
null-model	–	16.96	−0.77	70.86	−3.20	0.26
ClusterCV	All (theoretical case)	14.38	−0.64	60.11	−2.69	0.47
	Prior to harvest (mapping case)	15.64	0.07	65.36	0.30	0.37
StandCV	All (theoretical case)	11.42	−0.03	47.73	−0.12	0.66
	Prior to harvest (mapping case)	12.44	0.05	52.00	0.22	0.60

mean information on the proportion of BR damaged volume at the level of study area is available.

The null-model resulted in the largest error rates (and the smallest R^2 value) whereas the StandCV strategy resulted in smaller error rates than the ClusterCV strategy. With both cross-validation strategies, the exclusion of harvester variables (i.e., the mapping case compared with theoretical case) increased the error rates associated with the predicted BR volumes. The exclusion of the harvester variables in the mapping case increased the RMSE values by 8.7 % and 8.9 % compared with the theoretical case, for ClusterCV and StandCV, respectively. The magnitude of MD was moderate in all cases. Table 3 shows the RMSE and MD values associated with the null-model and cross-validated predictions of BR volume. The predicted versus observed values by predictor variable sets (mapping and theoretical case) and the cross-validation strategies are shown in Fig. 3.

An example of BR mapping using the RF model and the SR16 segments is shown in Fig. 4. The models presented in this study are not applicable for young forest stands. The "not applicable" stands shown in Fig. 4 were filtered out by comparing the attributes associated with our training data and the attributes provided in the SR16 map (Not applicable: 95th percentile of ALS height distribution <12 m and spruce volume proportion <50%).

3.2. Importance of predictor variables

A list of the predictor variables and their importance in the prediction of BR volume is in Fig. 5. The 95th percentile of the ALS height distribution ($H95_{ALS}$) was the most important variable. In the RF model with all predictor variables, harvested volume (V_{HRV}) was the second most important among all variables and the most important harvester variable. Harvester variables associated with the DBH distribution,

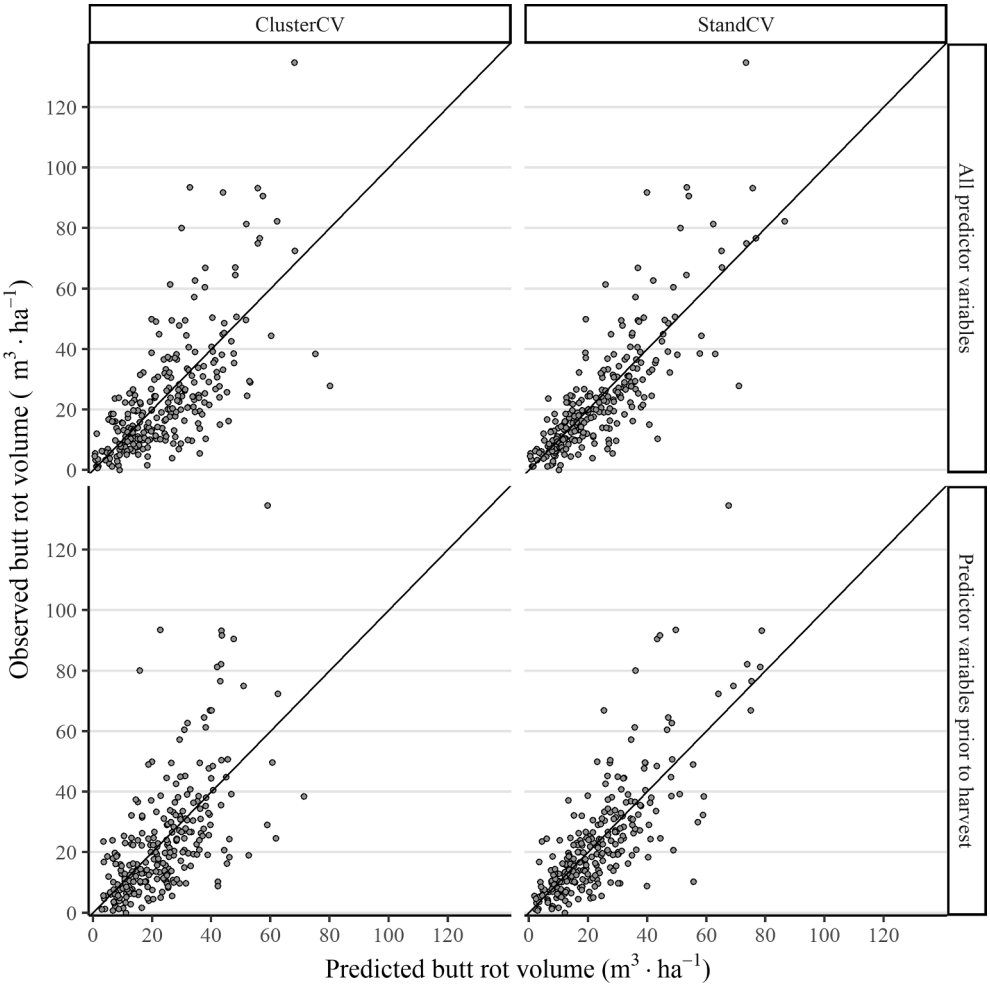


Fig. 3. Observed versus predicted butt rot volume using leave-cluster-out (ClusterCV) and leave-stand-out cross-validation (StandCV) strategies. The top row: all available predictor variables (theoretical case); bottom row: predictor variables available prior to harvest (mapping case).

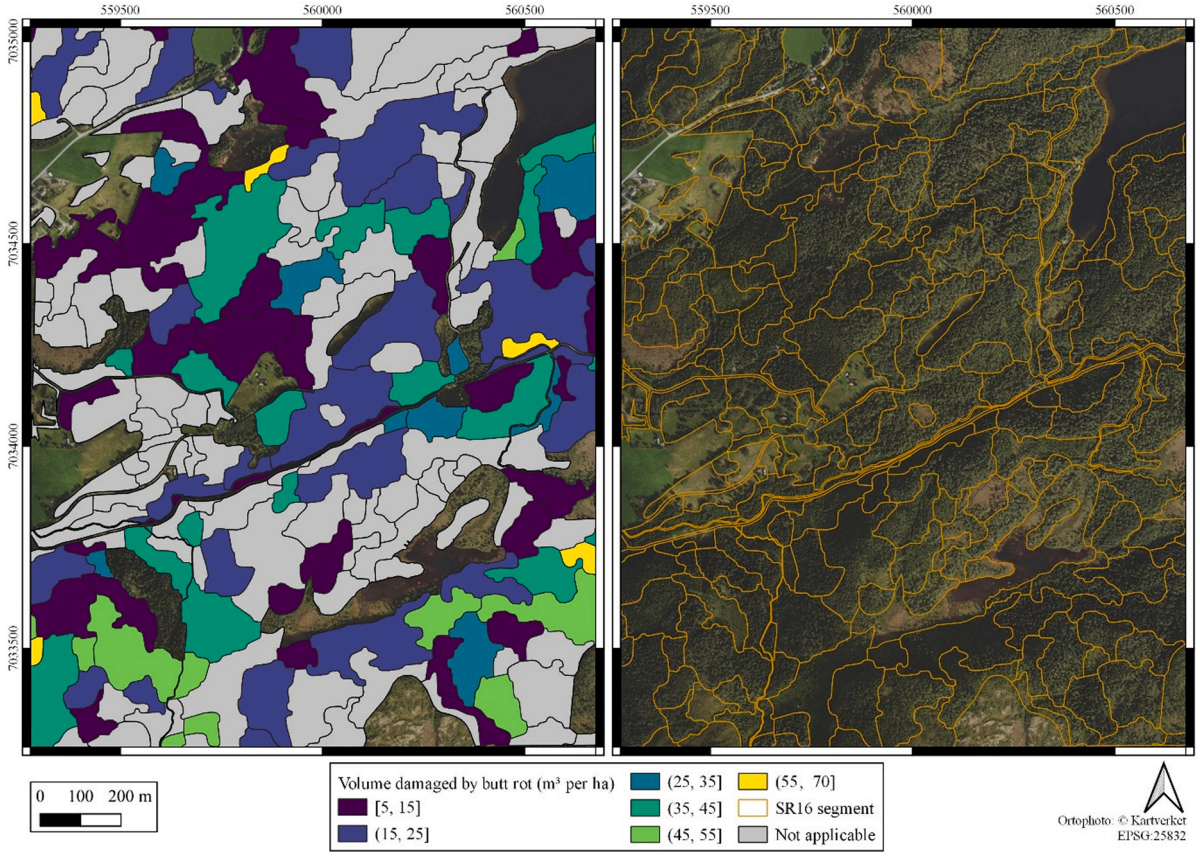


Fig. 4. Mapping of predicted timber volume damaged by butt rot for stand segments of the Norwegian forest resource map SR16. "Not applicable" refers to non-mature forests or forests not dominated by spruce.

especially quadratic mean DBH (QMD_{HRV}), were also observed to be important in the RF model. Harvested volume (V_{HRV}) and quadratic mean diameter (QMD_{HRV}) and the width of DBH distribution (DR_{HRV}) had Spearman correlations of larger than 0.5 with BR volume (Fig. 6). Other variables with Spearman correlations of larger than 0.5 with the response were $H95_{ALS}$ and the variance of ALS heights ($Hvar_{ALS}$).

Remotely sensed variables were generally more important predictor variables than the environmental variables. Figs. 5 and 6 show that the remotely sensed variables, especially ALS variables, are related to the response variable, and the variable importance values associated with remotely sensed variables were comparable with harvester variables. The most important environmental variable was the Y coordinate associated with the harvested stand. In terms of the Spearman correlation, slope (SL_{TER}) had the largest correlation among the environmental variables, but its importance values associated with both RF models were small. The X and Y coordinates also had a weak correlation with the response indicating no strong spatial trend in BR abundance. The predictor variables were positively correlated with the response variable except the near-infrared band of Sentinel-2 (NIR_{S2}), distance to coast (DC_{ENV}), altitude (AL_{ENV}) and the Y coordinate of harvested stand (Y).

4. Discussion

We scrutinized the modeling and subsequent mapping of stand-level spruce BR volume using harvester data in Norwegian spruce-dominated forests. The recording of BR damages at the level of individual stems requires additional effort from the harvester operator since BR damages must be visually observed at the crosscuttings of stems and manually recorded. BR damages are therefore not routinely recorded during harvest operations in Norway so far. This study indicates that the harvester data are a potentially valuable source for the mapping of BR damages in mature spruce-dominated forests.

The harvester variables, namely volume and quadratic mean DBH, and remotely sensed variables extracted from ALS data ($H95_{ALS}$ and $Hvar_{ALS}$) were among the most important predictor variables. The abovementioned predictor variables are associated with the maturity of forest which is known to positively correlate with the risk of BR damages (Hylen and Granhus, 2018; Müller et al., 2018). The large importance of harvester variables indicates that there is potential to decrease error rates associated with BR volume predictions by improving forest attribute maps, such as timber volume (Rahlf et al., 2021) and DBH distributions (Räty et al., 2021).

We focused on Norway spruce-dominated forests which means that the harvested stands were not always pure Norway spruce stands. Mixed stands are linked to the slower spread of BR damages compared with monocultures (Möykkynen and Pukkala, 2010). It is also evident that the likelihood of observing a large absolute BR volume is higher when the spruce volume proportion is large. Therefore, it is important to employ predictor variables that provide information on the tree species in the model. The tree species can be mapped using remotely sensed variables, such as the optical bands of Sentinel-2, or indirectly, for example, with predictor variables characterizing growing conditions in forests (Breidenbach et al., 2020). Several predictor variables may indirectly provide information on tree species compositions, which may be the reason why the harvester data-based spruce volume proportion (SPP_{HRV}) was not among the most important predictor variables in this study. It should also be noted that the harvested stands used in this study were strongly dominated by Norway spruce (Table 1).

Hylen and Granhus (2018) found that BR damages were linked to temperature sum and altitude. In our study, the environmental variables were generally not as important predictor variables as harvester or remotely sensed variables. It is critical to note that our study had smaller

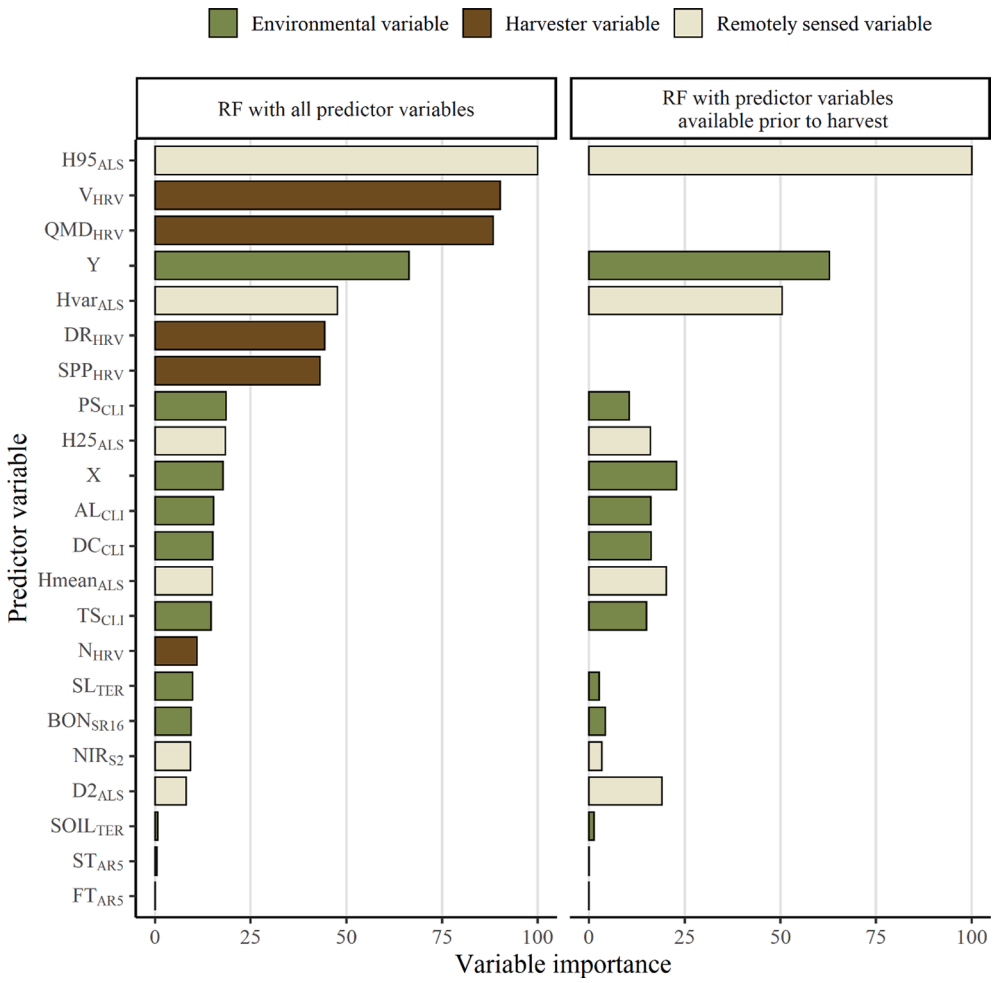


Fig. 5. Variable importance values associated with predictor variables of the Random Forests (RF) models. The variable importance values were scaled to 0–100 range. See Table 2 for the description of the predictor variables.

geographical coverage and the forests were structurally more homogenous compared with the study of Hylen and Granhus (2018). These differences may lead to the underestimation of the predictive power associated with the environmental variables in this study. We also used the geographical coordinates of the harvested stands in the RF models, and they achieved relatively large variable importance. However, it was found that the coordinates did not correlate with the response variable. Care must be taken when interpreting the variable importance values associated with the X and Y coordinates of the harvested stands. Their importance values do not directly indicate differences in terms of south-north or west-east directions, since the geographical coverage of our data was not comprehensive. The large importance values associated with the geographical coordinates rather resulted from the spatial autocorrelation associated with the BR observations among the harvested stands at the level of sub-regions.

StandCV resulted in smaller error rates (and larger R^2 values) than the null-model and ClusterCV, which confirms our hypothesis regarding the spatial autocorrelation of the BR damages. This can be explained by the fact that a target forest and its geographically nearest harvested stands are likely similar in growing conditions and silvicultural history. Thus, the results suggest that geographically comprehensive harvester data are required in order to further reduce prediction errors of BR volume. Especially, care must be taken when creating BR volume maps for new forested areas without harvested reference stands nearby.

The harvester data were not a probability sample over the study area and have a selection bias towards clear-cut stands. This means that the harvester data are usually limited to mature forest stands, which affects the applicability of the models fit based on harvester data. Therefore, the use of harvester data is usually studied in the context of timber procurement which is mostly associated with mature forests (Hauglin et al., 2018; Karjalainen et al., 2020; Peuhkurinen et al., 2008; Söderberg et al., 2021). There are also a few general challenges recognized in the application of harvester datasets. For example, the total timber volume is underestimated by harvesters (Kemmerer and Labelle, 2020), which is, however, not a problem in this study since the RB damage never reaches the top of a tree. Furthermore, retention trees are not typically recorded by the harvester, which may potentially decrease observed BR volumes at the stand-level. In addition, differences among the bucking schemes of the harvester machines (e.g. minimum allowed length of pulpwood log) may affect the accumulation of timber assortment volumes at the stand-level among the operation areas. We did not have access to the bucking schemes used by the harvesters.

Due to the technological differences in the GNSS systems of the harvesters, the positioning errors varied among the harvesters. It is realistic to assume that the average positioning error of the harvester head-positioned trees in our dataset likely ranges between 5 m and 10 m. For the other trees without harvester head positions, the average positioning error is likely larger than 15 m. An average positioning accuracy of 1 m can be achieved with an integrated positioning system which utilizes GNSS receivers and other sensors mounted in the harvester (Hauglin et al., 2017; Noordermeer et al., 2021). The positioning errors negatively affect the model errors of forest attributes and the effect increases with the decreasing size of the modeling units (Saukkola et al., 2019). We minimized the effect caused by positioning errors by using

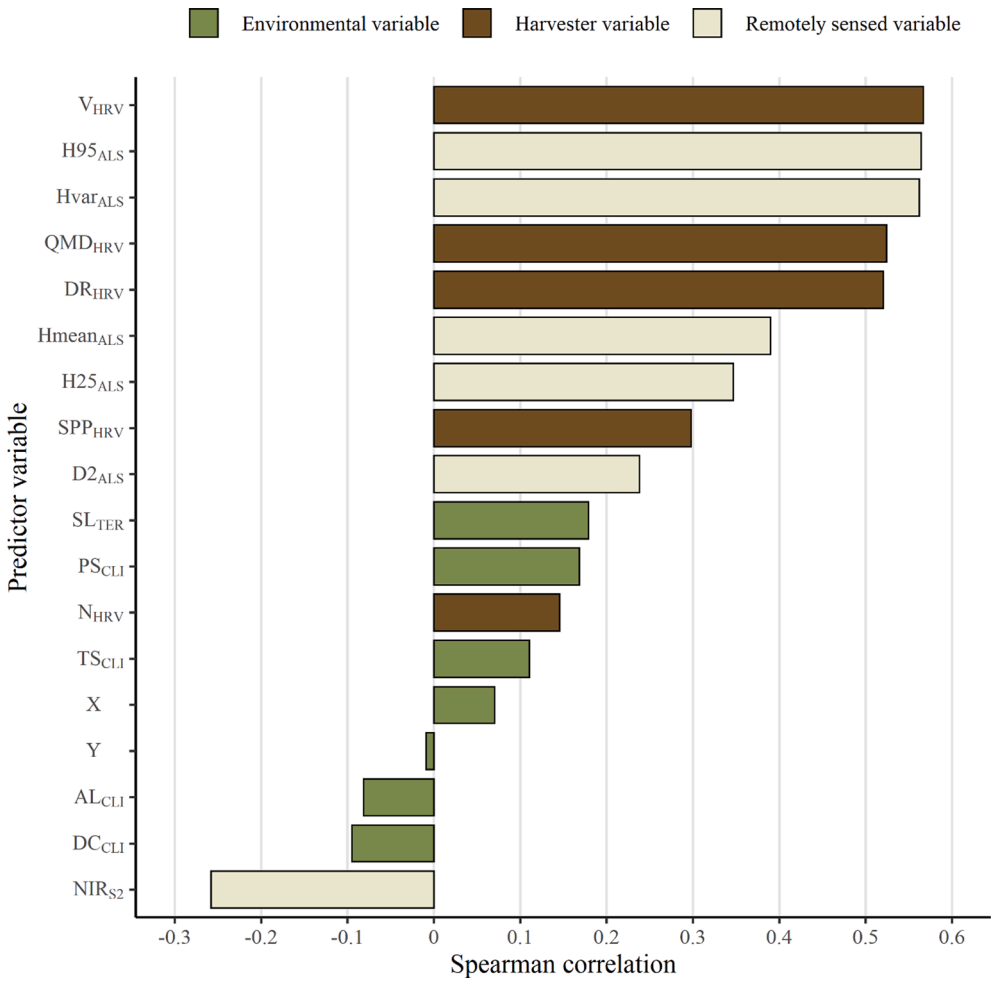


Fig. 6. Spearman correlations between the response variable (volume damaged by butt rot $m^3 \cdot ha^{-1}$) and each numeric predictor variable used in the Random Forests models. See Table 2 for the description of the predictor variables.

harvested stands as the modeling units and excluding harvested stands with few recorded trees or small area.

The findings of this study suggest that harvester data are a potential source for the mapping of BR volumes in mature spruce-dominated forests. We showed that a geographically comprehensive reference database is needed to minimize the error rates associated with the mapping of BR damages. Future work should consider different methodological solutions for the utilization of the continuous dataflow of harvester data in the mapping of BR damages.

5. Conclusions

We draw the following conclusions from this study: (1) Volume damaged by butt rot can be mapped using observations from cut-to-length harvester data combined with remotely sensed and environmental predictor variables. (2) Geographically comprehensive training data from an area of interest are required to map butt rot damages with satisfactory accuracy. (3) Predictor variables that characterize the maturity of a forest stand, such as remote sensing-based height characteristics, were the most important predictor variables in the modeling of butt rot volume. (4) The use of forest attributes obtained from harvester data as predictor variables, in addition to the remotely sensed and environmental variables, decreased error rates, which suggests that improved forest attribute maps may improve butt rot volume maps.

CRediT authorship contribution statement

Janne Räty: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Johannes Breidenbach:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Marius Hauglin:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Rasmus Astrup:** Conceptualization, Resources, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We express our gratitude to Simon Berg for the processing of harvester data. We would also like to thank Johannes Rahlf and Johannes Schumacher for their efforts in various data processing steps. We appreciate Ari Hietala's comments on the manuscript.

Funding

This study was supported by the Norwegian research council through the PRECISION project (NFR# 11067).

References

Ahlstrøm, A., Bjørkelo, K., Fadnes, K., 2019. AR5 klassifikasjonssystem, NIBIO bok. ed, NIBIO bok.

Aosaar, J., Drenkhan, T., Adamson, K., Aun, K., Becker, H., Buht, M., Drenkhan, R., Fjodorov, M., Jürimaa, K., Morozov, G., Pihlak, L., Piiskop, K., Riit, T., Varik, M., Väär, R., Uri, M., Uri, V., 2020. The effect of stump harvesting on tree growth and the infection of root rot in young Norway spruce stands in hemiboreal Estonia. For. Ecol. Manage. 475, 118425. https://doi.org/10.1016/j.foreco.2020.118425.

Arlinger, J., Nordström, M., Möller, J., 2012. StandForD 2010: Modern kommunikation med skogsmaskiner.

Astrup, R., Rahlf, J., Bjørkelo, K., Debella-Gilo, M., Gjertsen, A.-K., Breidenbach, J., 2019. Forest information at multiple scales: development, evaluation and application of the Norwegian forest resources map SR16. Scand. J. For. Res. 34 (6), 484–496. https://doi.org/10.1080/02827581.2019.1588989.

Breidenbach, J., Waser, L.T., Debella-Gilo, M., Schumacher, J., Rahlf, J., Hauglin, M., Puliti, S., Astrup, R., 2020. National mapping and estimation of forest area by dominant tree species using Sentinel-2 data. Can. J. For. Res. 51 (3), 365–379. https://doi.org/10.1139/cjfr-2020-0170.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Flood, N., 2013. Seasonal composite landsat TM/ETM+ images using the Medoid (a Multi-Dimensional Median). Remote Sens. 5, 6481–6500. https://doi.org/10.3390/rs5126481.

Geological Survey of Norway, 2020. Løsmasse og marin grense.

Hauglin, M., Hansen, E., Sørngård, E., Næsset, E., Gobakken, T., 2018. Utilizing accurately positioned harvester data: modelling forest volume with airborne laser scanning. Can. J. For. Res. 48 (8), 913–922. https://doi.org/10.1139/cjfr-2017-0467.

Hauglin, M., Hansen, E.H., Næsset, E., Busterud, B.E., Gjevestad, J.G.O., Gobakken, T., 2017. Accurate single-tree positions from a harvester: a test of two global satellite-based positioning systems. Scand. J. For. Res. 32 (8), 774–781. https://doi.org/10.1080/02827581.2017.1296967.

Hauglin, M., Rahlf, J., Schumacher, J., Astrup, R., Breidenbach, J., 2021. Large scale mapping of forest attributes using heterogeneous sets of airborne laser scanning and national forest inventory data. For. Ecosyst. 8, 65. https://doi.org/10.1186/s40663-021-00338-4.

Hodges, C.S., 1999. Heterobasidion annosum. Biology, Ecology, Impact and Control. Plant Pathol. 48, 564–565. https://doi.org/10.1046/j.1365-3059.1999.0366b.x.

Huse, K., Solheim, H., Venn, K., 1994. Råte i gran registrert på stubber etter hogst vinteren 1992 [Stump inventory of root and butt rots in Norway spruce cut in 1992]. Rapp Skogforsk 23, 1–26.

Hylen, G., Granhus, A., 2018. A probability model for root and butt rot in Picea abies derived from Norwegian national forest inventory data. Scand. J. For. Res. 33 (7), 657–667. https://doi.org/10.1080/02827581.2018.1487074.

Karjalainen, T., Mehtätalo, L., Packalen, P., Gobakken, T., Næsset, E., Maltamo, M., 2020. Field calibration of merchantable and sawlog volumes in forest inventories based on airborne laser scanning. Can. J. For. Res. 50 (12), 1352–1364. https://doi.org/10.1139/cjfr-2020-0033.

Karjalainen, T., Packalen, P., Räty, J., Maltamo, M., 2019. Predicting factual sawlog volumes in Scots pine dominated forests using airborne laser scanning data. Silva Fennica 53. https://doi.org/10.14214/sf.10183.

Kartverket, 2019. Høydedata og terrengmodeller for landområdene.

Kemmerer, J., Labelle, E.R., 2020. Using harvester data from on-board computers: a review of key findings, opportunities and challenges. Eur. J. Forest Res. 140 (1), 1–17. https://doi.org/10.1007/s10342-020-01313-4.

Kirches, G., 2018. Sentinel 2 Global Mosaics: Copernicus Sentinel-2 Global Mosaic (S2GM) within the Global Land Component of the Copernicus Land Service, Algorithm Theoretical Basis Document JRC: European Comission. Data available from https://www.sentinel-hub.com/.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Maltamo, M., Hauglin, M., Naesset, E., Gobakken, T., 2019. Estimating stand level stem diameter distribution utilizing harvester data and airborne laser scanning. Silva Fenn. 53 https://doi.org/10.14214/sf.10075.

Mattila, U., Nuutinen, T., 2007. Assessing the incidence of butt rot in Norway spruce in southern Finland. Silva Fenn. 41 https://doi.org/10.14214/sf.473.

Möykkynen, T., Pukkala, T., 2010. Optimizing the management of Norway spruce and Scots pine mixtures on a site infected by Heterobasidion coll. Scand. J. For. Res. 25 (2), 127–137. https://doi.org/10.1080/02827581003667322.

Müller, M.M., Kaitera, J., Henttonen, H.M., 2018. Butt rot incidence in the northernmost distribution area of Heterobasidion in Finland. For. Ecol. Manage. 425, 154–163. https://doi.org/10.1016/j.foreco.2018.05.036.

Noordermeer, L., Sørngård, E., Astrup, R., Næsset, E., Gobakken, T., 2021. Coupling a differential global navigation satellite system to a cut-to-length harvester operating system enables precise positioning of harvested trees. Int. J. Forest Eng. 32 (2), 119–127. https://doi.org/10.1080/14942119.2021.1899686.

Nordström, M., Hemmingsson, J., 2018. Measure up! A Skogforsk Guide to Harvester Measurement. Skogforsk, Uppsala.

Pateiro-Lopez, B., Rodriguez-Casal, A., 2019. alphahull: Generalization of the Convex Hull of a Sample of Points in the Plane.

Peuhkurinen, J., Maltamo, M., Malinen, J., 2008. Estimating species-specific diameter distributions and saw log recoveries of boreal forests from airborne laser scanning data and aerial photographs: a distribution-based approach. Silva Fennica 42. https://doi.org/10.14214/sf.237.

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahlf, J., Hauglin, M., Astrup, R., Breidenbach, J., 2021. Timber volume estimation based on airborne laser scanning — comparing the use of national forest inventory and forest management inventory data. Ann. Forest Sci. 78, 49. https://doi.org/10.1007/s13595-021-01061-4.

Räty, J., Astrup, R., Breidenbach, J., 2021. Prediction and model-assisted estimation of diameter distributions using Norwegian national forest inventory and airborne laser scanning data. Can. J. For. Res. 51 (10), 1521–1533. https://doi.org/10.1139/cjfr-2020-0440.

Saukkola, A., Melkas, T., Riekki, K., Sirparanta, S., Peuhkurinen, J., Holopainen, M., Hyyppä, J., Vastaranta, M., 2019. Predicting Forest Inventory Attributes Using Airborne Laser Scanning, Aerial Imagery, and Harvester Data. Remote Sens. 11, 797. https://doi.org/10.3390/rs11070797.

Skaugen, T.E., Hanssen-Bauer, I., Førland, E.J., 2002. Adjustment of dynamically downscaled temperature and precipitation data in Norway.

Söderberg, J., 2015. A method for using harvester data in airborne laser prediction of forest variables in mature coniferous stands. Master thesis, Sveriges lantbruksuniversitet, Master thesis 1–37.

Söderberg, J., Wallerman, J., Almäng, A., Möller, J.J., Willén, E., 2021. Operational prediction of forest attributes using standardised harvester data and airborne laser scanning data in Sweden. Scand. J. For. Res. 36 (4), 306–314. https://doi.org/10.1080/02827581.2021.1919751.

Song, J., 2015. Bias corrections for Random Forest in regression using residual rotation. J. Korean Stat. Soc. 44 (2), 321–326. https://doi.org/10.1016/j.jkss.2015.01.003.

Stenlid, J., 1987. Controlling and predicting the spread of heterobasidion annosum from infected stumps and trees of picea abies. Scand. J. For. Res. 2 (1-4), 187–198. https://doi.org/10.1080/02827588709382457.

Vähä-Konka, V., Maltamo, M., Pukkala, T., Kärhä, K., 2020. Evaluating the accuracy of ALS-based removal estimates against actual logging data. Ann. Forest Sci. 77, 84. https://doi.org/10.1007/s13595-020-00985-7.

Zhang, G., Lu, Y., 2012. Bias-corrected random forests in regression. J. Appl. Statist. 39 (1), 151–160. https://doi.org/10.1080/02664763.2011.578621.