Int J Appl Earth Obs Geoinformation 83 (2019) 101942

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

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

Utility of hyperspectral compared to multispectral remote sensing data in estimating forest biomass and structure variables in Finnish boreal forest

Eelis Halme[a,*], Petri Pellikka[b,c], Matti Mõttus[a]

[a] VTT Technical Research Centre of Finland, P.O. Box 1000, FI-02044 Espoo, Finland
[b] Department of Geosciences and Geography, P.O. Box 64, FI-00014, University of Helsinki, Finland
[c] State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

ARTICLE INFO

Keywords:
Hyperspectral imaging
Sentinel-2
Forest structure variables
Machine learning
Support vector regression
Gaussian process regression

ABSTRACT

Three-quarters of Finland's land surface area is filled with forests, which compose a great part of the country's biomass, carbon pools and carbon sinks. In order to acquire up-to-date information on the forests, optical remote sensing techniques are commonly used. Moreover, in the future hyperspectral satellite missions will start providing data to support the needs of natural resource management practices, such as forestry. It is, however, unclear what would be the additional value from using hyperspectral data compared to multispectral in quantifying forest variables of Finnish boreal forest. In this study, we used the remote sensing data by hyperspectral AISA imager (128 bands, 400–1000 nm, resolution 0.7 m) and Sentinel-2 (10 bands, resolution 10 m) to assess the possible benefits of higher spectral resolution. As reference data, we used a new nationwide forest resource dataset (stand-level data), which has a high potential in further remote sensing applications. In addition, we used a set of independent in situ measurements (plot-level data) for validation. We applied two kernel-based machine learning regression algorithms (Gaussian process and support vector regression) to relate boreal forest variables with the remote sensing data. The variables of interest were mean height, basal area, leaf area index (LAI), stem biomass and main tree species. The regression algorithms were trained with stand-level data and estimations were evaluated with stand- and plot-level holdout sets. The estimation accuracies were examined with absolute and relative root-mean-square errors. Successful variable estimations showed that kernel-based regression algorithms are suitable tools for forest structure estimation. Based on the results, the additional value of hyperspectral remote sensing data in forest variable estimation in Finnish boreal forest is mainly related to variables with species-specific information, such as main tree species and LAI. The more interesting variables for forestry industry, such as mean height, basal area and stem biomass, can also be estimated accurately with more traditional multispectral remote sensing data.

1. Introduction

Forests are significant carbon sinks and play a great role in climate change mitigation globally. Research on vegetation parameter retrieval is of special relevance in order to extend our knowledge about the dynamics of forests and the environment in general at local and global scales (Verrelst et al., 2012a). Estimation of forest variables from remotely sensed spectral information has been of interest since the beginning of optical remote sensing. In the past, field surveys were essential to obtain information on the forests, whereas now they are often used as a supplement to remote sensing data. Satellite remote sensing has become a valuable and low-cost tool for the purposes of forest inventories and management planning. It is also the only tool capable of monitoring vast areas. For example, Finland has 22.8 million hectares of forest, which equals to three-quarters of the country's land surface area. The role of remote sensing in large area inventories, for instance in the National Forest Inventories (NFI) of Finland, is crucial. The forests of Finland serve as an important resource for the nation's nature conservation as well as for the forestry industry. Furthermore, they compose a great part of the country's carbon pools and sinks.

The fundamentals of any biophysical parameter retrieval includes relating spectral signatures to actual parameters on the ground. Quantification of biophysical parameters can be done essentially in two ways: utilizing statistical approaches or inverting radiative transfer models (Atzberger, 2004). The former is variable-driven method and requires direct experimental data on the retrieved variables, whereas the latter (also known as the physical method) is driven by radiometric data and rely on simulations of a radiative transfer model. The division

* Corresponding author.
E-mail addresses: eelis.halme@vtt.fi (E. Halme), petri.pellikka@helsinki.fi (P. Pellikka), matti.mottus@vtt.fi (M. Mõttus).

https://doi.org/10.1016/j.jag.2019.101942
Received 23 April 2019; Received in revised form 10 July 2019; Accepted 8 August 2019

of biophysical parameter retrieval methods into statistical and physical categories has been blurred over the years, as integrating the two has also yielded the hybrid methods.

Variable retrieval from inverting radiative transfer models can be computationally expensive, if the used model is advanced and complex. These models typically have many input variables and are more realistic than simpler models that are faster to invert. When using the advanced models, the inversion is by nature ill-posed (Atzberger, 2004). Consequently, the biophysical variable retrieval can be challenging. Furthermore, according to Verrelst et al. (2019) the utilization of radiative transfer model inversions is preferable when one is interested in the underlying radiative transfer processes, whereas in the presence of reliable reference data, statistical approaches suit better to quantification of a specific variable.

The statistical approaches include parametric and nonparametric regression methods. The former includes, for instance, spectral indices that are essentially new variables generated by combinations of spectral bands. Vegetation index (VI) is one type of spectral index and extremely popular with biophysical variable estimation. With VI-based methods, estimations are commonly based on regression technique, where a parameterized expression is designed utilizing physical or statistical knowledge so that the expression would have a strong linear relation to some biophysical variable. The prime problem with the simple VI-based methods is that they become extremely dependent on the implicit assumptions being made. In geographic areas when the assumptions are not met, the estimation results could be poor (Tuia et al., 2011). In addition, on fine temporal and spectral scales, the VI-based regression models often fail to capture some physiological processes (Nichol et al., 2000). Moreover, according to Verrelst et al., 2012a these models underexploit the full potential of the contiguous spectral data of hyperspectral imaging, as it can be difficult to find the optimal band combination for biophysical variable retrieval.

In theory, nonparametric methods are more suitable for hyperspectral imaging data analysis than parametric ones. The nonparametric methods do not assume any fixed functional form, hence no a priori assumptions are made, and they only rely on the available data (Tuia et al., 2011). The nonparametric methods can be further divided into linear and nonlinear models, of which the latter can also be referred to as machine learning regression algorithms (Verrelst et al., 2015). These algorithms fit a flexible model directly from the data in order to learn the relation between the input and output (Verrelst et al., 2012b), which makes them perfect tools for hyperspectral imaging data analysis.

A diversity of machine learning regression algorithms has been developed over the years. Popular regression algorithms in different application domains include, for instance, random forest (RF), convolutional neural network (CNN) and Gaussian process regression (GPR). The foregoing, RF belongs to the family of decision trees that are commonly used more in classification than in regression tasks. CNN is a deep learning method, which belongs to the family of artificial neural networks. For the estimation of biophysical parameters, several architectures and models of neural networks have been considered; however, some significant drawbacks have been found (Gómez-Chova et al., 2011). GPR is part of the family of kernel-based regressions. Utilization of kernel methods is a promising alternative to neural networks. Kernel-based regression methods transfer the data into a high dimensional space using a kernel function in order to solve a nonlinear regression problem. With kernel-based methods, any linear method can be transformed to nonlinear one, while still operating with linear algebra (Gómez-Chova et al., 2011).

A recent study by Verrelst et al. (2012b) compared three kernel-based machine learning regression algorithms to neural networks in biophysical variable estimation. The comparison was made in terms of accuracy, goodness of fit, computational cost and robustness to low sample sized scenarios. The results showed that neural networks performed unstably, whereas Gaussian process regression (GPR) and support vector regression (SVR) provided the best accuracies, GPR being slightly more accurate. Several other studies have also noticed the good performance of GPR and SVR in biophysical variable estimation with hyper- and multispectral remote sensing data (e.g., Camps-Valls et al., 2006; Hultquist et al., 2014; Pasolli et al., 2008; Rabe et al., 2009; Tuia et al., 2011). Therefore, we selected these methods as the regression tools used in our study.

This study was carried out in the southern boreal forest zone in Finland. Previous studies of machine learning regression algorithms in vegetation biophysical parameter retrieval have mainly been carried out in different biomes than the boreal forest (e.g., Hultquist et al., 2014; Verrelst et al., 2012a, 2012b).

In the boreal forest of the same region, only Mutanen et al. (2016) have used GPR to estimate tree height. Other stand-level estimations in Finnish boreal forest have applied, for instance, the k-nearest neighbor (k-NN) method when estimating forest variables from optical space- or airborne imagery (Hyvönen, 2002; Tuominen et al., 2017). In addition, different segmentation (Anttila and Lehikoinen, 2002) and pattern recognition (Maltamo et al., 2003) methods have been used in stand-level estimations of forest variables in Finnish boreal forest. To our knowledge, machine learning regression algorithms have not been applied before to hyperspectral remote sensing data of this region.

We used a comprehensive stand-level forest resource dataset, which has recently become publicly available from the Finnish Forest Centre, for training the kernel-based regression methods. The data have a national coverage and thus a huge potential in further remote sensing applications. The data rely heavily on forestry measurements, but have been scaled to the level of a forest stand. To understand the spatial scale at which the data can be used in forest mapping (i.e., the spatial details of forest information that can still be retrieved with the models trained by the data), we tested the transferability of the obtained regression models to independent plot-level measurements. We then used the stand-level data to assess the additional benefit of hyperspectral imaging data, in comparison with the more widely available multispectral data (e.g., Sentinel-2), in estimating forest variables of boreal forest. We hope that the study will contribute to a better understanding of the benefit of higher spectral resolution of future hyperspectral satellite missions and the sufficiency of currently existing forest data for the future needs of forestry industry. The specific objectives of this paper are:

1) to investigate the estimation accuracy of forest biomass and structure variables in Finnish boreal forest from stand-level data using kernel-based regression algorithms;
2) to study the suitability of the newly available Finnish Forest Centre stand-level data for training the kernel-based regression methods for forest variable retrieval;
3) to assess the additional value of hyperspectral remote sensing data, compared with multispectral optical satellite remote sensing data, in estimating forest biomass and structure variables of Finnish boreal forest.

2. Materials and methods

2.1. Study site

Our study site is located in the vicinity of Hyytiälä forestry field station (61°50′44″N, 24°17′10″E) in Juupajoki, Finland, in the southern boreal forest zone. The area around the station includes managed boreal forests, agricultural fields and wetlands. The main overstory species of the forests are Scots pine (*Pinus sylvestris*), Norway spruce (*Picea abies*) and Silver birch (*Betula pendula*), either as mixed or single-species stands. Different shrubs, lichens and mosses compose the forest floors.

2.2. Remote sensing data

We used airborne hyperspectral AISA (128 bands, 400–1000 nm, spectral resolution 4.7 nm, spatial resolution 0.7 m) and multispectral Sentinel-2 (10 bands in VNIR, spatial resolution 10 m) images. The AISA data was acquired in nine separate flight lines under clear sky on June 15, 2017. The solar zenith angle decreased from 42.3° to 40.3° during the data acquisition between 11:37 AM and 12:14 PM (Eastern European Summer Time, UTC + 3). The AISA Eagle II hyperspectral scanner (Specim, Oulu, Finland) had a 37.7° field of view. A more thorough description of the scanner is reported by Markiet et al. (2017), together with the data pre-processing information.

Sentinel-2 image (Level-2A product downloaded from Copernicus Open Access Hub) was acquired on June 30, 2017 at 10:00 AM, when the sun zenith angle was 39°. Sentinel-2 products have 13 spectral bands ranging from about 400 to 2200 nm. Since the wavelength range of the hyperspectral AISA image was approximately from 400 nm to 1000 nm, the Sentinel-2 bands B10, B11 and B12, with central wavelengths of 1373 nm, 1614 nm and 2202 nm, respectively, were not used. Furthermore, the desired pixel size for the Sentinel-2 image was selected to be 10 m. Bands that had a lower spatial resolution were resampled to 10 m using the nearest neighbor resampling method in SNAP software (version 6.0.4).

2.3. Reference data

2.3.1. Forest Centre data

Collecting forest resource information on privately owned Finnish forests is a statutory assignment for the Finnish Forest Centre. The forest resource data they provide is gathered by field measurements and remote sensing techniques, mainly laser scanning and aerial imaging. The data is simulated and updated to certain epochs with the help of growth models and all available reports on occurred forest operations, such as loggings or thinnings. The forest resource data was published as freely downloadable under Creative Commons Attribution 4.0 International (CC BY 4.0) license during the spring of 2018. The forest data is distributed in two different forms: as 16 m × 16 m grid cells and as stands. The grid cell is the inventory unit for which the forest data is computed, and the stand-level data is aggregated or averaged from the grid cells that intersect with the stand boundaries (Fig. 1).

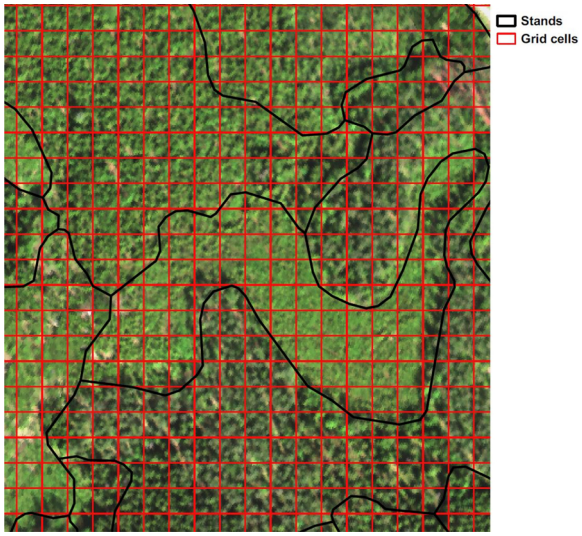

Fig. 1. Stand geometries and the grid of cells in Hyytiälä, southern Finland. The size of a cell is 16 m. The forest variables of the stands are aggregates or weighted averages of the grid cell data that fall inside the stands.

Data in a single grid cell are obtained with models that have been developed for each forest variable using spatially accurate field measurements and remote sensing data. The reference plots measured in the field have typically a radius of nine meters, and approximately 600 to 800 plots are measured per each inventory area. The high accuracy is guaranteed with multiple measurements, always done by different persons. The variables measured in the field are mean height, basal area, stem count, mean diameter and age, with other forest variables derived using models. The reference plots were then used to retrieve forest variables from the laser scanning data's height distribution and aerial images typically using area-based nonparametric methods, such as k-Most Similar Neighbors and sparse Bayesian estimation. The aerial images were specifically used for tree species identification. The outline of the laser scanning based stand-level inventory is described by Maltamo et al. (2011).

In the database provided by the Finnish Forest Centre, stand-level data is divided into strata, homogeneous sets of trees within one stand, including only one species. A stand might include several strata of the same species, if there exists, for example, young and mature spruce trees within one stand. For some variables of interest, such as mean height, basal area and stem biomass, aggregated stand-level values were directly available. Other variables were calculated for each stratum and then added to obtain stand-level values (Table 1). We calculated the leaf area index (LAI) from foliage mass (leaf biomass) provided by the Forest Centre using species-specific values of specific leaf area (SLA), reported in Majasalmi et al. (2013). The calculated (allometric) LAI conforms to the common definition of leaf area index as one-sided green leaf area per unit area of ground (Chen and Black, 1992). We next applied a shoot-level clumping correction to the allometric LAI as described by Majasalmi et al. (2013) to obtain the effective LAI values for all stands and assumed it to be the biophysical variable driving forest reflectance instead of the true, allometric LAI.

We used two methods to compute the main tree species of a stand: based on basal area and LAI (Table 1). For the former, the species that contributes the most to the total stand basal area is assigned as the main tree species of the stand; for the latter, contributions of tree species were quantified by their effective LAI. The risk of misidentifying the dominant species decreases with increasing dominance of a species (Varvia, 2018); it is the largest for stands made up of equal amounts of different overstory species. In the extreme case, no dominant tree species exist. Therefore, we documented two different main tree species estimations: one using all stands and another using only the stands where a species had a dominance of over 75%.

Table 1
Forest variables of interest. Derived variables were based on stratum information.

Forest variable of interest	Unit
Directly read variables	
mean height	m
basal area	m^2/ha
stem biomass	t/ha
Derived variables	
LAI	unitless
main tree species (LAI-weighted)	unitless
main tree species (basal area-weighted)	unitless

2.3.2. Independent in situ measurements

An independent set of field measurements collected between June 24 and August 29, 2013 around the Hyytiälä forestry field station was used for the final evaluation of the forest variable estimations. The regularly located plot-level measurements included data on basal area, mean height, LAI and main tree species. The main tree species classification of a plot was based on the basal area of the dominating species. Full description of the measurements is given by Majasalmi et al. (2015).

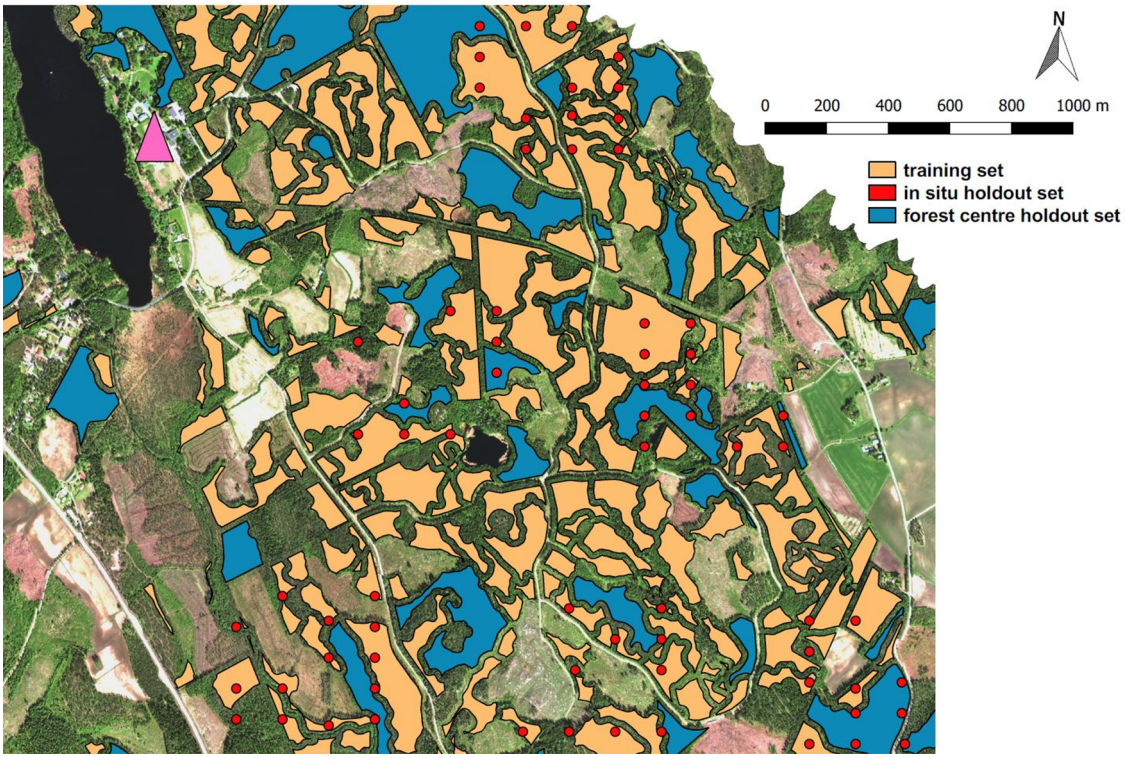

Fig. 2. A subset of the stands and in situ plots used in this study. Hyytiälä forestry field station is located in the upper left corner by the Lake Kuivajärvi, marked with a triangle. The AISA image was used as background raster visible between the buffered stands.

We used only the plots that had center points within a stand provided by the Finnish Forest Centre. Furthermore, we extracted a circular area of 15-meter radius around the center point of each plot in order to make them slightly more similar with natural stands. Eventually, 120 plot geometries were used, for which hyperspectral AISA data was available. In order to evaluate the main tree species estimations with the independent measurements, we utilized only the plots that intersected with the stands for which we had collected the stand-level data. This decreased the number of used in situ plots by fourteen.

The effective LAI was measured using the LAI-2000 Plant Canopy Analyzer (LI-COR, Lincoln, Nebraska, USA). This optical instrument gives an estimate of the canopy silhouette area index rather than the true LAI (Stenberg et al., 2014). The silhouette area index, also known as effective or optical LAI, should be divided by hemispherical clumping index (Γ) to obtain the true LAI corresponding to the allometric value calculated from foliage mass. However, quantitative estimates were only available for shoot-level clumping, with the clumping caused by other structural levels (most importantly, the crown) left unaccounted for. Therefore, the calculated stand-level effective LAI is not completely equivalent to the effective LAI of the in situ measurements. Although, we expect a high correlation between the two.

2.4. Machine learning regression algorithms

Based on previous studies (e.g., Camps-Valls et al., 2006; Hultquist et al., 2014; Verrelst et al., 2012b), we chose two kernel-based algorithms for the study: Support vector regression (SVR) and Gaussian process regression (GPR). The former, is based on ε-insensitive error function ($\varepsilon > 0$), where the error function gives zero error, if the difference between prediction and actual value is smaller than ε. A more extensive and precise description of SVR can be found from the works of Smola and Schölkopf (2004) and Bishop (2006). GPR is a machine learning approach based on Gaussian process (GP) theory. The approach is described more comprehensively in the works of Rasmussen and Williams (2006) and Pasolli et al. (2010). The chosen algorithms were implemented with scikit-learn that is a widely used machine learning library for Python programming language (Pedregosa et al., 2011). Kernel methods require tuning of hyperparameters. With the GPR implementation, hyperparameter tuning concentrated more on finding a suitable kernel. Based on preliminary trials, a radial basis function (RBF) kernel was chosen with added white noise kernel. For SVR, we chose also RBF.

2.5. Methods

2.5.1. Cross-validation

In general, when working with machine learning, the available data should be divided into three sets: training, validation and testing. Training set is used for model training, validation set is used for evaluation, and when a strong model has been found it is tested using the test set. However, this requires a large amount of data. In this study, the hyperspectral AISA data were spatially limited. A solution for this was k-fold cross-validation, which is a statistical method to estimate the skill of machine learning models. When using cross-validation, it is enough to divide the available data in two: training and holdout sets. In this study, we used 5-fold cross-validation and two different holdout sets: one based on the Forest Centre data and another based on the independent in situ measurements. The training set was created from the Forest Centre data.

2.5.2. Training and holdout sets

Data for the training set and to the Forest Centre holdout set were created from the stand-level data. For each stand, the forest variables of interest were the target values and the zonal mean reflectances were the feature values. For the in situ holdout set, the zonal mean reflectances were computed from the plot geometries. As we used two different remote sensing images, we had two different datasets with different zonal mean reflectances, both covering the same 745 stands. In order to have the Forest Centre holdout set the same size as the in situ holdout

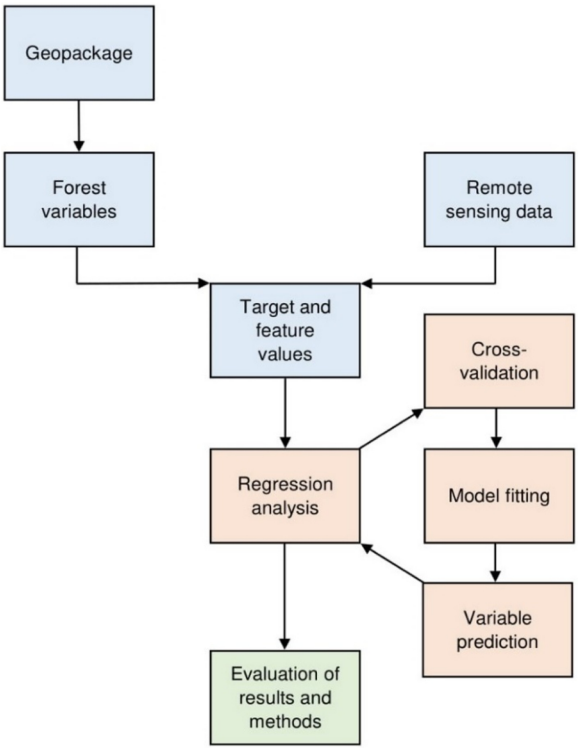

Fig. 3. The analysis process: geopackage contains the data from the Finnish Forest Centre, remote sensing data includes AISA and Sentinel-2 images, target values are forest variable values, and feature values are the zonal mean reflectances for each stand.

set, 120 stands were picked from the AISA and Sentinel-2 datasets to the Forest Centre holdout sets. The remaining 625 stands were left for training. The stands in the holdout set were evenly distributed across the study area and stands of different sizes were included (Fig. 2).

Only stands that had an area of at least 0.5 ha were used in the study. In addition, the stands of the AISA and Sentinel-2 datasets were downsized with 10-meter buffer before calculating the zonal mean reflectances so that possible spectral mixing was avoided. Furthermore, a threshold value of 0.86 was set for LAI to exclude stands that were logged between the acquisition of the remote sensing data (June 2017) and the epoch of the used reference data (beginning of the year 2018). In addition, a threshold value of 0.61 was set for NDVI (calculated from hyperspectral data) to exclude stands with low vegetation corresponding to non-forest (i.e., recently logged stands with little above-ground biomass). Validation of the chosen thresholds was based on visual analysis with the hyperspectral AISA image.

2.5.3. Estimations of forest variables

The estimations were performed separately for each forest variable (Fig. 3): the training set was given to the machine learning regression algorithm, which trained a predictive model. Then the features (zonal mean reflectances) of a holdout set (Forest Centre or in situ) were given to the trained model, which produced new estimations. The main tree species estimations, however, differed from the other variables, as it was a categorical variable. To make it compatible with regression models, we combined it with species-specific information on basal area and LAI by creating variables quantifying the basal areas and LAI values for pine, spruce and broadleaved trees within each stand. These species-specific values were given as target vectors to the regression algorithms. The predicted dominant tree species was calculated from the estimated species-specific values.

2.5.4. Accuracy assessment

A holdout set was utilized to evaluate the accuracy of the predictions (Fig. 3) with the use of root-mean-square error (RMSE). In order to compare the prediction performance between the forest variables of interest, RMSE was normalized to the mean of the target vector of the holdout set to obtain the relative root-mean-square error (rRMSE) that was expressed as percentages. In addition, we computed relative bias as the difference between the mean of the holdout set's target vector and the mean of the estimated values divided by the mean of the holdout set's target vector. The accuracy of successfully determining the correct main tree species was presented as overall accuracy, calculated by dividing the number of correct predictions by the total number of data points. We also calculated the coefficient of determination (R^2) as the square of the Pearson correlation coefficient between the holdout set and estimated values.

3. Results

3.1. Evaluating with Forest Centre holdout set

The estimation accuracy differences between different combinations of algorithms and datasets were small for all variables. In general, GPR provided slightly better or equally good estimation accuracies for each forest variable compared to SVR (Table 2). The best estimation accuracy was obtained for mean height (rRMSE = 15%), with basal area being the second best (rRMSE = 17%). Leaf area index (LAI) and stem biomass had poorer estimation accuracies (rRMSE 20% and 28%, respectively).

The combination of GPR and AISA dataset had the smallest rRMSE for estimating mean height and LAI (Table 2). The AISA dataset provided also the smallest rRMSE for basal area together with SVR. However, the difference in RMSE was very small (0.01 m^2/ha) to GPR and AISA combination. GPR together with Sentinel-2 dataset provided the smallest rRMSE for stem biomass.

Estimation maps of stem biomass, basal area and LAI using SVR with Sentinel-2 dataset are given in Fig. 4.

Table 2
Relative and absolute root-mean-square errors of estimated forest variables for support vector (SVR) and Gaussian process regression (GPR) algorithms. AISA and Sentinel-2 correspond to the remote sensing images that were used to compute the stand zonal mean reflectances to the used datasets. Also, relative bias and coefficient of determination (R^2) are given for each estimation.

	AISA		Sentinel-2	
Forest variable	**SVR**	**GPR**	**SVR**	**GPR**
rRMSE				
mean height	17%	15%	17%	17%
basal area	17%	17%	18%	18%
LAI	22%	20%	27%	25%
stem biomass	29%	28%	29%	28%
RMSE				
mean height (m)	2.88	2.60	2.95	2.98
basal area (m^2/ha)	3.84	3.85	4.01	3.93
LAI	0.72	0.66	0.86	0.80
stem biomass (t/ha)	22.50	21.78	22.33	21.46
Relative bias				
mean height	–3%	–2%	–3%	–3%
basal area	−2%	0%	–2%	0%
LAI	–2%	0%	–3%	0%
stem biomass	–6%	–3%	–5%	–2%
R^2				
mean height	0.59	0.66	0.57	0.56
basal area	0.68	0.67	0.65	0.66
LAI	0.81	0.83	0.72	0.75
stem biomass	0.61	0.62	0.61	0.63

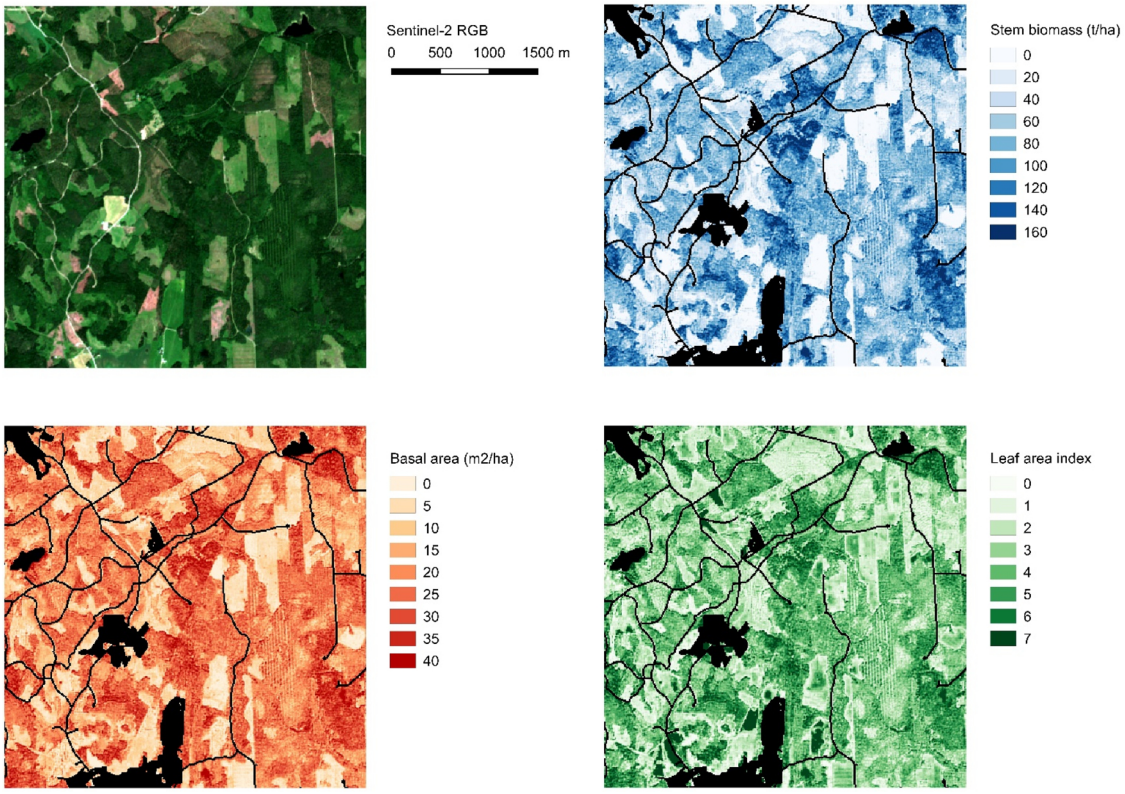

Fig. 4. Estimation maps of stem biomass, basal area and LAI. The estimations were performed for Sentinel-2 pixels (10 m resolution and 10 bands) with a support vector regression model that was trained using Forest Centre stand-level data. Non-forest areas were masked out (black areas).

Table 3
Overall accuracy of the main tree species estimations for scenarios where stand's tree species dominance of over 75% is ignored and taken into account. The abbreviations "BA" and "LAI" in parentheses correspond to basal area and leaf area index, respectively.

	AISA		Sentinel-2	
Forest variable	**SVR (%)**	**GPR (%)**	**SVR (%)**	**GPR (%)**
Species dominance ignored				
main tree species (BA)	90	93	88	87
main tree species (LAI)	88	87	83	84
Over 75 % dominance present				
main tree species (BA)	94	98	88	90
main tree species (LAI)	100	98	93	95

The main tree species estimation was better for the stands that had a clear dominant tree species compared to the results for all stands (Table 3). Nearly all accuracies were close to or over 90%. Excellent results were obtained when using the hyperspectral dataset and only the stands where over 75% dominance was present: the accuracy reached 100% for AISA data and LAI-based dominance.

Exclusions of stands with a species dominance below 75% decreased the sample size from 120 stands by more than 50%; for estimations based on basal area by 57% and based on LAI by 64%. The estimations based on basal area provided the highest overall accuracy when the species dominance was not taken into account and all holdout set stands were used for evaluation. Using only the stands where over 75% species dominance was present increased the overall accuracies. In contrast to former, the estimations based on LAI provided the best accuracies. The highest accuracies were obtained using the hyperspectral AISA dataset.

Table 4
Relative and absolute root-mean-square errors of estimated forest variables for support vector (SVR) and Gaussian process regression (GPR) algorithms. AISA and Sentinel-2 correspond to the remote sensing images that were used to compute the stand and plot zonal mean reflectances to the used datasets. Also, relative bias and coefficient of determination (R^2) are given for each estimation.

	AISA		Sentinel-2	
Forest variable	**SVR**	**GPR**	**SVR**	**GPR**
rRMSE				
mean height	39%	39%	38%	37%
basal area	45%	45%	47%	48%
LAI	68%	65%	68%	71%
RMSE				
mean height (m)	6.43	6.39	6.20	6.14
basal area (m^2/ha)	8.79	8.66	9.06	9.28
LAI	1.72	1.65	1.72	1.80
Relative bias				
mean height	7%	9%	10%	11%
basal area	25%	25%	28%	29%
LAI	46%	45%	50%	54%
R^2				
mean height	0.11	0.13	0.19	0.21
basal area	0.12	0.12	0.10	0.10
LAI	0.26	0.33	0.23	0.24

3.2. Evaluating with in situ holdout set

The smallest rRMSE values were obtained for basal area and LAI using the GPR and AISA combination (Table 4). The most accurate mean height estimation was obtained using the same algorithm, but the Sentinel-2 dataset. Mean height was estimated most accurately (rRMSE = 37%), whereas basal area had the second lowest rRMSE (45%). Estimations of LAI had the highest rRMSE (65%). Overall, the estimations evaluated with the in situ holdout set were less accurate than

Table 5
Overall accuracy of main tree species estimations, when using the independent in situ measurements as the holdout set. The abbreviation "BA" refers to basal area.

	AISA		Sentinel-2	
Forest variable	**SVR (%)**	**GPR (%)**	**SVR (%)**	**GPR (%)**
main tree species (BA)	71	69	68	68

estimations evaluated with the stand-level Forest Centre holdout set.

When evaluating with the in situ measurements, main tree species were defined in the holdout set only based on basal area. The overall accuracy of the main tree species estimation was close to 70% for each dataset and algorithm combination (Table 5).

The dissimilarities between the Forest Centre data and the independent in situ measurements are noticeable from the relative biases (Table 4). The forest variable estimations evaluated using the in situ holdout set had higher biases than the estimations evaluated with the Forest Centre holdout set. Especially, the biases of LAI were clearly different.

4. Discussion

4.1. Estimation accuracies on stand-level

The relative root-mean-square errors were all good or at least satisfactory (rRMSE below 30%) for the forest variable estimations evaluated with the Forest Centre holdout set. According to Uuttera et al. (2002), from the perspective of practical forestry the accuracy requirement for stand-level inventory is 10–20% (rRMSE) for mean height and 15–25% (rRMSE) for basal area. In this respect, the estimations of this study were successful. Moreover, the small relative biases were a proof of similarity between training and holdout sets, and therefore the regression models did not make incorrect assumptions on the data.

The estimation accuracies were very similar between the dataset and algorithm combinations for mean height, basal area and stem biomass, albeit small differences were noticeable especially between the two algorithms. For instance, the rRMSE for stem biomass were the same for AISA and Sentinel-2 datasets, with only differences in the regression algorithm. In general, GPR was able to predict more accurately. Also with LAI estimations, GPR provided the smallest rRMSE values, but more importantly, the hyperspectral data clearly had an effect on the estimations. The calculated LAI values depend on species-specific values (e.g., SLA), and it is known that hyperspectral data is capable of distinguishing different vegetation species (Féret and Asner, 2011). Based on this, it can be argued that the higher spectral resolution of AISA data had a more positive effect to the LAI estimation accuracies than the higher spatial resolution. In addition, as the estimation accuracies were approximately the same for the other variables of interest regardless of the used dataset, it follows that on stand-level, spatial resolution does not have a significant effect, whereas spectral resolution does, especially for variables with species-specific information. This result is in line with what would be expected: as the used reflectances were stand zonal mean values, the effect of higher spatial resolution decreases, since the energy is averaged within the same area.

In addition, the main tree species estimations succeed very well on stand-level. The estimation accuracies increased when we used the stands, which had a species dominance of over 75%. For estimations based on basal area and LAI, the increase was on average three and eleven percentage points, respectively. When the species dominance was not noticed, estimations based on basal area provided better accuracies for each dataset and algorithm combination. On the other hand, with the species dominance taken into account, estimations based on LAI were better. The reason for this is probably the accuracy of the reference data. According to the Finnish Forest Centre, basal area has good accuracy ($\pm$ 3 m^2/ha), but the accuracy of LAI, calculated using allometric relations, is lower. Although by using only the stands with a clear species dominance, we decreased the number of stands by more than 50%. It is reasonable to use only the stands where a dominant species can be found, since if no clear dominant species exists, one cannot be meaningfully predicted. The hyperspectral AISA data provided better overall accuracies for all estimations of the dominant species. In fact, species identification is known to be one of the application examples of hyperspectral imaging systems (Féret and Asner, 2011; Piiroinen et al., 2015), as hundreds of narrow spectral bands help to retrieve information from specific leaf constituents.

4.2. Estimation accuracies on plot-level

When evaluating with the in situ holdout set, the forest variable estimations were made on plot-level. Compared to the stand-level (chapter 4.1) estimations, the accuracies were lower. The estimation accuracies of mean height, basal area, LAI and main tree species could be compared between the two holdout sets. In general, the rRMSE values were considerably higher, and the biases were high for basal area and LAI. The large biases showed that the training set and in situ holdout set lacked similarity. In order to make the forest variable estimations to succeed, the training data ought to be comprehensive, unbiased and similar with the holdout set (Hyvönen, 2002). In this study, the training set can be considered as comprehensive, but it did not share the same variable distribution as the in situ data. As a result, the regression models ended up making incorrect assumptions on the data, and we obtained poor estimation accuracies and large biases. Especially LAI estimations had a large bias, which can be explained with different LAI values. In the process of calculating effective LAI for all stands, we only applied shoot-level clumping correction and left the clumping caused by other structural levels unaccounted for. This yielded to a slight overestimation of effective LAI values (Stenberg et al., 2014). As these values were used in the training phase, the bias on stand-level estimations was small and on plot-level large. Furthermore, the large bias of basal area estimations can be explained with similar standard deviation but different mean between the training set and in situ holdout set (data not shown).

The main tree species estimations were also poor compared to the estimations evaluated with stand-level data. One issue was the fact that we used only those plots that intersected the used stands. However, some of the stands included several plots. Therefore, a misclassification was unavoidable. In addition, the remote sensing data were from four years later than the in situ measurements, thus adding a potential source of errors to the analysis.

The stands in the forest data are delineated subjectively to be relatively homogeneous. However, local differences in reflectance and transmittance within a stand are inevitable due to gaps in the canopy. In boreal forests, the structure and optical properties of understory can have a clear effect on the forest spectra that can vary due to, for example, fertility type or moisture conditions of the local forest site (Rautiainen et al., 2018). Our results indicate that a small number of in situ plots that locate within larger stands were not representative of the properties of a natural stand. The machine learning models trained with the stand-level data produced poor estimates for the small in situ plots.

4.3. Comparison with similar studies

Very few research papers are available on estimating forest variables on stand-level in boreal forest using machine learning and passive optical remote sensing data. Mutanen et al. (2016) used Gaussian process regression to estimate tree height from stand-wise forest maps and Landsat 8 image. Their study area was located near the Hyytiälä forestry field station. They obtained RMSE of 5.06 m for the tree height estimations; the estimation accuracy was worse than our stand-level estimation (RMSE = 2.60 m), but better than our estimation on plot-

level (RMSE = 6.14 m).

In Finland, stand-level estimations of forest structure variables utilizing space- or airborne imagery were studied more in the beginning of 2000s (e.g., Anttila and Lehikoinen, 2002; Hyvönen, 2002; Maltamo et al., 2003). At present, research concentrates on utilizing airborne laser scanning (ALS) data. Tuominen et al. (2017) compared different remote sensing data in the estimation of forest variables and found that space- and airborne optical imagery together with 3D information from ALS or digital aerial photogrammetry provided better accuracies than the 2D imagery alone. They used the k-nearest neighbor (k-NN) method for the forest variable estimations, and as field data, they utilized National Forest Inventory (NFI) sample plots with a fixed radius of 9 m. As remote sensing data, they used features that coincided with the sample plots. Using Landsat 8 data (bands 1–7), they obtained rRMSE of 34% and 45% for mean height and basal area estimations, respectively. Our best accuracies with spaceborne imagery were on stand-level 17% and 18%, and on plot-level 37% and 47%, respectively. Utilizing aerial imagery (four spectral bands, 0.3 m spatial resolution) produced better accuracy for mean height (rRMSE = 31%) and slightly worse accuracy for basal area (rRMSE = 46%). Our values using airborne data (128 bands, 0.7 m spatial resolution) were better: on stand-level 15% and 17%, and on plot-level 39% and 45%, respectively. However, the aerial data we used had a higher spectral resolution. Hyvönen (2002) used also the k-nearest neighbor method for forest variable estimations. He studied stand-level forest variables using nonparametric kNN method, Landsat TM imagery and forest inventory information. Comparable variables from his study were mean height and basal area. For mean height the rRMSE was 34% (RMSE = 4.55 m) and for basal area 41% (RMSE = 6.53 m^2/ha). These were weaker than the stand-level estimations by us: rRMSE of 15% (RMSE = 2.60 m) and 17% (RMSE = 3.84 m^2/ha), respectively. Anttila and Lehikoinen (2002) estimated forest variables on stand-level with semiautomatic segmentation in Finland, and found that rRMSE for mean height and basal area were 17% (RMSE = 3.2 m) and 35% (RMSE = 8.0 m^2/ha), respectively. These results were outperformed by our stand-level results. Maltamo et al. (2003) determined stand characteristics using field data and airborne video imagery. They developed a new method, which combined pattern recognition of single trees from video imagery and the theoretical diameter distribution to determine stand characteristics. They obtained rRMSE of 19% (RMSE = 3.64 m^2/ha) for the basal area estimations, which is poorer than our stand-level estimation accuracy.

The results of this study are in line with previous studies where Gaussian process and support vector regression techniques have been used together with earth observation data. GPR was able to provide more accurate estimations than SVR also in previous studies (e.g., Pasolli et al., 2008; Verrelst et al., 2012b). Earlier studies have also shown that GPR was computationally faster than SVR. With our implementation, however, SVR turned out to be much faster. For example, when evaluating with the Forest Centre holdout set, GPR was six times slower in estimating basal area when the AISA dataset was used. There were also differences in running durations between the forest variables. The relative differences in time consumption between the two methods compared to the earlier studies can be considered software-dependent. For instance, Verrelst et al. (2012b) used Matlab implementation.

4.4. Study results compared to study objectives

Based on the previous Finnish studies, our forest structure variable estimations were successful. The stand-level estimation accuracies were good when compared to any previous study listed in this paper. The highest accuracies in the reference data were for the mean height and basal area, which were the same variables that were estimated with the lowest rRMSE. The known accuracy, according to the Finnish Forest Centre, is for mean height ± 2 m and for basal area ± 3 m^2/ha. The other variables of interest are modeled from the basic attributes and have lower accuracy. We estimated all variables of interest with good or reasonable accuracy on stand-level. In addition, this study showed that on stand-level the additional value of hyperspectral data is related to variables with species-specific information. In general, Sentinel-2 data with 10-meter spatial resolution provided similar accuracy compared to the hyperspectral data.

4.5. Stand- and plot-level differences

This study showed that plot-level data remains difficult to upscale to stand-level data with reasonable accuracy, and practices on how the stand- and plot-level data could be used interchangeably in forest variable estimations require further research. Stand-level data is always modeled in some way, and the most straightforward method to validate forest variable estimations is using independent field measurements.

Stand-level estimations in this study, especially mean height, basal area and LAI, were good. However, the rRMSE values for the plot-level estimations were surprisingly large. This can be attributed to the dissimilarities between the variable distributions of stand- and plot-level data and possible spectral neighborhood effect. In addition, the temporal differences between the used remote sensing data (summer 2017), the independent field measurements (summer 2013) and the epoch of the used forest data (beginning of the year 2018) possibly contributed towards poor model transferability for both GPR and SVR models.

The reliability of the forest data is slightly unclear. The accuracy of mean height and basal area are known to be good, but other data used in this study, especially the variables depending on species-specific stratum information, have weaker quality. Furthermore, updating of the forest data is dependent on growth models and reports on occurred forest operations. According to the Finnish Forest Centre, growth modeling has weaker reliability in seeding stands than in more mature stands. Fortunately, in this study the number of seeding stands was very low. The growth models work well at least five years after a successful inventory. As the area around Hyytiälä forestry field station was inventoried in 2015, the growth modeling should be accurate, if all occurred forest operations have been reported. Nevertheless, this study has shown that the modeled stand data is difficult to compare with the direct measurements of the in situ plots.

5. Conclusions

In this study, we performed estimations of forest variables on stand- and plot-level including stem biomass, basal area, mean height, leaf area index (LAI) and main tree species. The estimations were performed using two machine learning regression algorithms; Gaussian process (GPR) and support vector (SVR) regressions. The algorithms were trained with stand-level data and evaluated with stand- and plot-level holdout sets. In machine learning the representativeness of the training data is in high value. In this study, the stand-level data was difficult to compare with plot-level data, as the former is modeled data and the latter is direct measurement. On stand-level, the algorithms showed good performance, and proved themselves as potential tools for forest variable estimation in the Finnish boreal forest. In general, GPR slightly outperformed SVR. The estimations of main tree species, mean height and basal area were the most accurate. We found that on stand-level spatial resolution has smaller effect on forest variable estimation than spectral resolution. Variables related to species-specific information, i.e. main tree species and LAI, benefited from the higher spectral resolution.

Declaration of Competing Interest

None

Acknowledgements

Funding was supported by the Academy of Finland grants 311928 and 317387. We are grateful to Titta Majasalmi and Miina Rautiainen

(Aalto University) for providing the in situ measurements from Hyytiälä. In addition, we thank the Finnish Forest Centre for the additional and more detailed information on the forest resource data.

References

Anttila, P., Lehikoinen, M., 2002. Kuvioittaisten puustotunnusten estimointi ilmakuvilta puoliautomaattisella latvusten segmentoinnilla. Metsätieteen Aikakauskirja 2002 (3), 381–389. https://doi.org/10.14214/ma.6178. (in Finnish).

Atzberger, C., 2004. Object-based retrieval of biophysical canopy variables using artificial neural nets and radiative transfer models. Remote Sens. Environ. 93 (1–2), 53–67. https://doi.org/10.1016/j.rse.2004.06.016.

Bishop, C.M., 2006. Pattern recognition and machine learning. Information Science and Statistics. Springer-Verlag, Berlin, Heidelberg 738 pp.

Camps-Valls, G., Bruzzone, L., Rojo-Álvarez, J.L., Melgani, F., 2006. Robust Support vector regression for biophysical variable estimation from remotely sensed images. Ieee Geosci. Remote. Sens. Lett. 3 (3), 339–343. https://doi.org/10.1109/LGRS.2006.871748.

Chen, J.M., Black, T.A., 1992. Defining leaf area index for non-flat leaves. Plant Cell Environ. 15 (4), 421–429. https://doi.org/10.1111/j.1365-3040.1992.tb00992.x.

Féret, J., Asner, G.P., 2011. Spectroscopic classification of tropical forest species using radiative transfer modeling. Remote Sens. Environ. 115 (9), 2415–2422. https://doi.org/10.1016/j.rse.2011.05.004.

Gómez-Chova, L., Muñoz-Marí, J., Laparra, V., Malo-López, J., Camps-Valls, G., 2011. A review of kernel methods in remote sensing data analysis. In: In: Prasad, S., Bruce, L., Chanussot, J. (Eds.), Optical Remote Sensing. Augmented Vision and Reality 3. Springer, Berlin, Heidelberg, pp. 171–206. https://doi.org/10.1007/978-3-642-14212-3_10.

Hultquist, C., Chen, G., Zhao, K., 2014. A comparison of Gaussian process regression, random forests and Support vector regression for burn severity assessment in diseased forests. Remote. Sens. Lett. 5 (8), 723–732. https://doi.org/10.1080/2150704X.2014.963733.

Hyvönen, P., 2002. Kuvioittaisten puustotunnusten ja toimenpide- ehdotusten estimointi k-lähimmän naapurin menetelmällä Landsat TM -satelliittikuvan, vanhan inventointitiedon ja kuviotason tukiaineiston avulla. Metsätieteen Aikakauskirja 2002 (3), 363–379. https://doi.org/10.14214/ma.6177. (in Finnish).

Majasalmi, T., Rautiainen, M., Stenberg, P., Lukeš, P., 2013. An assessment of ground reference methods for estimating LAI of boreal forests. For. Ecol. Manage. 292, 10–18. https://doi.org/10.1016/j.foreco.2012.12.017.

Majasalmi, T., Rautiainen, M., Stenberg, P., Manninen, T., 2015. Validation of MODIS and GEOV1 fPAR products in a boreal forest site in Finland. Remote Sens. (Basel) 7 (2), 1359–1379. https://doi.org/10.3390/rs70201359.

Maltamo, M., Tokola, T., Lehikoinen, M., 2003. Estimating stand characteristics by combining single tree pattern recognition of digital video imagery and a theoretical diameter distribution model. For. Sci. 49 (1), 98–109. https://doi.org/10.1093/forestscience/49.1.98.

Maltamo, M., Packalén, P., Kallio, E., Kangas, J., Uuttera, J., Heikkilä, J., 2011. Airborne laser scanning based stand level management inventory in Finland. Proceedings of SilviLaser 2011, 11th International Conference on LiDAR Applications for Assessing Forest Ecosystems 1–10 Hobart, Australia: Conference Secretariat. 777 pp.

Markiet, V., Hernández-Clemente, R., Mõttus, M., 2017. Spectral similarity and PRI variations for a boreal forest stand using multi-angular airborne imagery. Remote Sens. (Basel) 9 (10). https://doi.org/10.3390/rs9101005. 17 pp.

Mutanen, T., Sirro, L., Rauste, Y., 2016. Tree height estimates in boreal forest using gaussian process regression. International Geoscience and Remote Sensing Symposium (IGARSS) 1757–1760. https://doi.org/10.1109/IGARSS.2016.7729450.

Nichol, C.J., Huemmrich, K.F., Black, T.A., Jarvis, P.G., Walthall, C.L., Grace, J., Hall, F.G., 2000. Remote sensing of photosynthetic-light-use efficiency of boreal forest. Agric. For. Meteorol. 101 (2–3), 131–142. https://doi.org/10.1016/S0168-1923(99)00167-7.

Pasolli, L., Blanzieri, E., Melgani, F. 2008. Estimating Biophysical Parameters from Remotely Sensed Imagery with Gaussian Processes. IGARSS 2008 – 2008 IEEE International Geoscience and Remote Sensing Symposium, Boston, MA, 2008, pp. II-851–854. DOI: 10.1109/IGARSS.2008.4779128.

Pasolli, L., Melgani, F., Blanzieri, E., 2010. Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing data. Ieee Geosci. Remote. Sens. Lett. 7 (3), 464–468. https://doi.org/10.1109/LGRS.2009.2039191.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. Available from: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Piiroinen, R., Heiskanen, J., Mõttus, M., Pellikka, P., 2015. Classification of crops across heterogeneous agricultural landscape in Kenya using AisaEAGLE imaging spectroscopy data. Int. J. Appl. Earth Obs. Geoinf. 39, 1–8. https://doi.org/10.1016/j.jag.2015.02.005.

Rabe, A., Van Der Linden, S., Hostert, P. 2009. Simplifying Support vector machines for regression analysis of hyperspectral imagery. WHISPERS' 09–1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. Article number 5289090, 4 pp. DOI: 10.1109/WHISPERS.2009.5289090.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian processes for machine learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, USA 248 pp.

Rautiainen, M., Lukeš, P., Homolová, L., Hovi, A., Pisek, J., Mõttus, M., 2018. Spectral properties of coniferous forests: a review of in situ and laboratory measurements. Remote Sens. (Basel) 10 (207). https://doi.org/10.3390/rs10020207. 28 pp.

Smola, A., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

Stenberg, P., Mõttus, M., Rautiainen, M., Sievänen, R., 2014. Quantitative characterization of clumping in Scots pine crowns. Ann. Bot. 114 (4), 689–694. https://doi.org/10.1093/aob/mct310.

Tuia, D., Verrelst, J., Alonso, L., Perez-Cruz, F., Camps-Valls, G., 2011. Multioutput Support vector regression for remote sensing biophysical parameter estimation. Ieee Geosci. Remote. Sens. Lett. 8 (4), 804–808. https://doi.org/10.1109/LGRS.2011.2109934.

Tuominen, S., Pitkänen, T., Balázs, A., Kangas, A., 2017. Improving Finnish multi-source national forest inventory by 3D aerial imaging. Silva Fenn. 51 (4). https://doi.org/10.14214/sf.7743. 21 pp.

Uuttera, J., Hiltunen, J., Rissanen, P., Anttila, P., Hyvönen, P., 2002. Uudet kuvioittaisen arvioinnin menetelmät – arvio soveltuvuudesta yksityismaiden metsäsuunnitteluun. Metsätieteen Aikakauskirja 2002 (3), 523–531. https://doi.org/10.14214/ma.6189. (in Finnish).

Varvia, P. 2018. Uncertainty quantification in remote sensing of forests. Dissertation. University of Eastern Finland, Faculty of Science and Forestry, Department of Applied Physics. Kuopio. 44 pp. http://urn.fi/URN:ISBN:978-952-61-2867-2.

Verrelst, J., Alonso, L., Camps-Valls, G., Delegido, J., Moreno, J., 2012a. Retrieval of vegetation biophysical parameters using Gaussian process techniques. Ieee Trans. Geosci. Remote. Sens. 50 (5), 1832–1843. https://doi.org/10.1109/TGRS.2011.2168962. PART 2.

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J.P., Camps-Valls, G., Moreno, J., 2012b. Machine learning regression algorithms for biophysical parameter retrieval: opportunities for Sentinel-2 and -3. Remote Sens. Environ. 118, 127–139. https://doi.org/10.1016/j.rse.2011.11.002.

Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J.P., Veroustraete, F., Clevers, J.G.P.W., Moreno, J., 2015. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties – a review. Isprs J. Photogramm. Remote. Sens. 108, 273–290. https://doi.org/10.1016/j.isprsjprs.2015.05.005.

Verrelst, J., Malenovský, Z., Van der Tol, C., Camps-Valls, G., Gastellu-Etchegorry, J.-, Lewis, P., North, P., Moreno, J., 2019. Quantifying Vegetation Biophysical Variables from Imaging Spectroscopy Data: A Review on Retrieval Methods. Surv. Geophys. 40 (3), 589–629. https://doi.org/10.1007/s10712-018-9478-y.