ISSN 1946-7664. MCFNS 2021
Vol. 13, Issue 1, pp. 15–28
Mathematical and Computational
Forestry & Natural-Resource Sciences

AVAILABLE ONLINE AT HTTP://MCFNS.COM

Submitted: Jul. 3, 2020
Accepted: Mar. 21, 2021
Published: Mar. 30, 2021
Last Correction: Mar. 22, 2021

SPATIAL ANALYSIS OF AIRBORNE LASER SCANNING POINT CLOUDS FOR PREDICTING FOREST STRUCTURE

Henrike Häbel[1], Andras Balazs[2], Mari Myllymäki[2]
[1] *Institute of Environmental Medicine, Karolinska Institutet, Nobels väg 13, SE-171 65 Solna, Sweden*
[2] *Natural Resources Institute Finland (Luke), Latokartanonkaari 9, FI-00790 Helsinki, Finland*

Abstract. The arrangement of trees with respect to each other plays a role in various forestry decisions. In this study, the arrangement of trees was summarized by three different structure indices. Their values were determined from field measurements and predicted with the well-known k-nn estimation method using data obtained by airborne laser scanning (ALS). ALS-derived predictions are often assisted by vertical summaries of the pulse returns. Our goal was to identify spatial summaries of the ALS point cloud that can improve predictions based on commonly used feature metrics. We explored the horizontal distribution of the pulse returns through canopy height models thresholded at different height levels. We introduce completely new metrics based on 1) the Euler number, which is the number of patches of vegetation minus the number of gaps, and 2) the empty-space function, which is a spatial summary function of the gap space. Data from a study site in Central Finland was available with circular field plots with 9 m radius. We find that small sample plots can be challenging. Still, we present evidence that the use of spatial feature metrics improves the prediction of forest structure indices and has potential for improvements for other forest variables related to gap structures.

Keywords: Airborne laser scanning, canopy height model, empty-space function, Euler number, forest resource prediction, spatial pattern of trees.

Background

The spatial structure of forest can be understood as the arrangement of tree locations with respect to each other. In forestry, the spatial structure plays an implicit role for example in thinning, where the aim is to produce a regular spatial pattern of trees in order to optimize the use of space for the growth of trees (e.g. Packalen et al. 2013, Pukkala 1990). The need of (first) thinning is typically earlier in clustered stands than in regular or random stands due to stronger competition between trees for growth factors and the right timing of thinning can have a great impact on growth and dynamics of a tree stand. Therefore, reliable information about the spatial structure of the forest could aid in the determination of a stand's treatment (Pippuri et al. 2012). In fact, the spatial structure of forest could be valuable also in other kinds of forestry decisions, related for example to stratifying forested areas or locating field sample plots (Packalen et al. 2013), see also Wang et al. (2020) and references therein. However, often this information is not available because the fieldwork for measuring the spatial locations of trees can be rather laborious and expensive.

There are a few studies that have aimed to classify the spatial structure of forests to regular, random and clustered forests utilizing remote sensing data. Uuttera et al. (1998), for instance, applied individual tree segmentation to high-resolution aerial photographs for this purpose, but experienced difficulties especially with clustered patterns of trees. Packalen et al. (2013) and Wang et al. (2020) tried a method based on individual tree detection as well. Packalen et al. (2013) and Pippuri et al. (2012) further aimed at classifying and predicting the spatial arrangement of trees with an area based approach for airborne laser scanning (ALS) data. The results of Pippuri et al. (2012) were promising, but Packalen et al. (2013) concluded that particularly the detection of clustered patterns appears to be difficult. Wang et al. (2020) studied the tree top patterns obtained from dense ALS data (with average pulse density 15 or 26/m^2).

Often the ALS-assisted forest inventories utilize the area-based approach that can be used even with low resolution ALS data where individual trees cannot be detected adequately. In this approach, different feature metrics calculated from the 3D ALS point cloud have been used to classify the field plots to regular, random and clustered in the works mentioned above. Prior to predicting, the field data used for training the algorithm has been classified as well. A usual approach is to summarize the structure of the field data from the complete set of locations of the trees in a field plot by summary functions, such as the Ripley's K function (Ripley 1976), or by single valued indices such as the aggregation index (Clark and Evans 1954). The function or index value is then compared to the reference arrangement called complete spatial randomness (CSR) or Poisson forest to deduce whether the pattern of trees can be regarded as random or it exhibits regularity or clustering (see e.g. Illian et al. 2008, Tomppo 1986).

The ALS feature metrics that are used in the area-based approach typically summarize the 3D point cloud vertically where examples are height, standard deviation or skewness of first returns (see e.g. Tomppo et al. 2017, Tuominen et al. 2018, 2017). However, prediction of spatial forest structure by the area-based approach can presumably benefit from metrics that summarize the spatial or horizontal distribution of the pulses. In the literature, there are such metrics for summarizing the spatial point cloud by describing the patch structure of thresholded canopy height models (CHMs) or by measuring the canopy complexity in different ways, see e.g. Kane et al. (2011, 2008), Li et al. (2014), Zhang et al. (2017) and the references therein. Pippuri et al. (2012) and Packalen et al. (2013) also utilized metrics constructed from a thresholded CHM. In order to create a thresholded CHM, Pippuri et al. (2012) set the threshold level to 5 m above ground, where pulse return values below the threshold were declared as gap and those above as canopy patch. Packalen et al. (2013) chose instead an adaptive threshold based on the maximum height of the CHM, which on average set the threshold to about 70% of the maximum height. Both Pippuri et al. (2012) and Packalen et al. (2013) used horizontal "landscape" metrics such as the mean and standard deviation of the size of patches.

In this work, we revisit the prediction of the spatial structure of boreal forest utilizing ALS data in the area-based approach. First, we propose and compare three different indices for summarizing the spatial structure of the field data. Two of these indices are based on spatial summary functions, namely the Ripley's K function and the so-called empty-space function (see e.g. Illian et al. 2008), which are commonly used to describe the structure of spatial point patterns. In addition to these rather complex spatial indices, we consider a crude measure for clustering, namely the aggregation index (Clark and Evans 1954). By using these indices we study the spatial forest structure on a continuous scale, rather than discretely as considered by Pippuri et al. (2012) and Packalen et al. (2013). Second, following the area-based approach, we predict the indices utilizing common ALS feature metrics from the practice of the management inventory (Tomppo et al. 2017) and spatial ALS features extracted from thresholded CHMs. To the best of our knowledge, some of the spatial ALS features are introduced for the first time in this context, namely the Euler number (the number of vegetation patches minus the number of gaps) and metrics based on the empty-space function also known as the spherical distribution function. Instead of thresholding the CHM only once or twice, we suggest to use several thresholds at different height levels. Finally, we illustrate the developed methodology on fitting prediction models for an example of field and ALS data from a study region in Central Finland. We show that the spatial features are useful and carefully discuss the challenges related to using rather small circular field plots with 9 m radius for model fitting.

Materials

Field data

A total of 2469 field plots was measured on a study site in Central Finland in 2013 following a systematic cluster sampling (see Figure 1). The land area was 5700 km^2, of which 4310 km^2 were forestry land including also poorly productive forest land and unproductive land as defined in the Finnish national forest inventories and management inventories. The topography is relatively flat with elevation values generally between 100 m and 200 m above sea level. Belonging to the southern and middle boreal vegetation zone, the forests are mainly coniferous, where Scots pine (*Pinus sylvestris* L.) and Norway spruce (*Picea abies* [L.] H. Karst.) are the most common species. The principal silvicultural system in the region has been even-aged management (Tomppo et al. 2017, Tuominen et al. 2017).

All trees with diameter at breast height (dbh) greater than 4.5 cm were measured for fixed radius plots of 9 m (ca. 254 m^2). More details on all measurements made can be found in Tomppo et al. (2017).

In this study, a subset of plots within single stands with at least ten trees with $dbh > 4.5$ cm and available ALS data were considered and only the location, and dbh of the trees were included in the analysis. This resulted in a total number of 1161 plots and 34965 measured trees (see Figure 1). Table 1 summarizes forest characteristics of the selected plots organized according to development class, which describes the development phase of the growing stock in relation to the expected rotation determined in the field (Tomppo et al. 2011, p. 40).

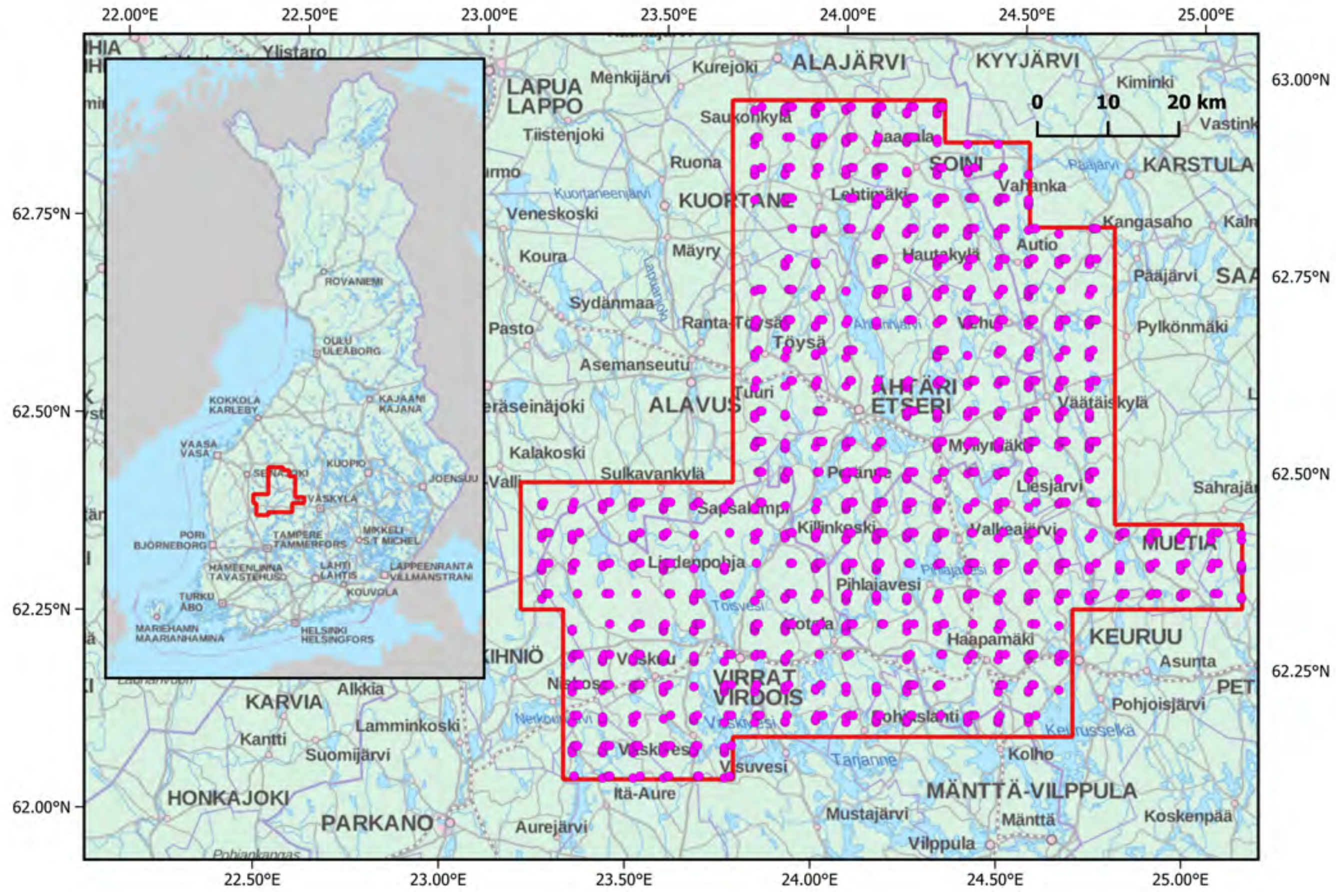

Figure 1: Location of the study site in Central Finland and locations (dots) of the 1161 plots used in the study (contains map data from the National Land Survey of Finland Topographic Database 02/2021).

Most plots (92%) were classified as either young thinning stands (Class 1), advanced thinning stands (Class 2), or mature stands (Class 3). A young thinning stand has a young growing stock at the thinning cuttings stage. Advanced thinning stands' growing stock is older and pole size larger, and the growing stock of a mature stand is either old or large enough for a regeneration cutting from the management point of view. The rest of the stands (collectively referred to as others in Table 1) were either regeneration or seedling stands or the information on the class was missing.

Validation data

In the same region in Central Finland, 30 additional plots were measured in 2014 for the purpose of model validation. The plots were of size 32 m × 32 m and subdivided into four subplots of size 16 m × 16 m (256 m^2), matching approximately with the size of the field plots. The plots were selected at locations where estimation with ALS usually results in large root mean squared errors (RMSEs) (Tomppo et al. 2017). Almost all validation plots were thinning stands, of which 17 were young (Class 1) and 10 were advanced (Class 2). Even though also smaller trees were measured, only trees with a *dbh* greater than 4.5 were included in a validation study to match the 2013 data.

ALS data

The ALS data were acquired by Blom Kartta Oy, Finland, for the operative management inventory by the Finnish Forest Centre between 28 June and 27 August 2013. The Piper Navajo airplane and the Optech Gemini ALTM scanner were used with the following parameters: flight altitude 1730 m, strip overlap 20%, pulse frequency 70,000 Hz, scanning frequency 37 Hz, half scan angle 20 degrees, pulse density $0.89/m^2$, and maximum number of observed pulse returns 4. For the analysis presented here, only the ALS data at the field plots were used.

	Development class			
	1: young	2: advanced	3: mature	other
No. plots	305	582	183	91
Mean diameter, cm	10.46 (2.08)	15.06 (3.18)	17.09 (4.42)	7.68 (1.66)
Basal area, m^2/ha	14.24 (6.06)	21.09 (6.80)	25.90 (9.23)	3.68 (1.90)
Pine, %	0.67 (0.36)	0.56 (0.33)	0.37 (0.31)	0.72 (0.40)
Spruce, %	0.14 (0.23)	0.24 (0.26)	0.38 (0.31)	0.17 (0.32)
Broadleaved, %	0.19 (0.26)	0.20 (0.20)	0.24 (0.24)	0.11 (0.25)
No. of stems/ha	1544 (701)	1128 (575)	999 (515)	711 (249)

Table 1: Average values (standard deviations) of the forest variables calculated from trees with $dbh \geq 4.5$ cm on 1161 plots summarized for development classes young thinning stands (1), advanced thinning stands (2), mature stands (3), and others (regeneration and seedling stands as well as unknown). The species percentages refer to proportions of basal area per species of the total basal area per plot.

Methods

The presented methodology can be divided into three different pipelines. The first is for calculating the forest structure indices from ground level measurements, the second covers the processing of the ALS point cloud data, and the third deals with the model fitting (see Figure 2).

Prior to introducing the studied forest structure indices, the ALS feature metrics, and the modelling steps, we first describe necessary background for the spatial analysis. Finally, we explain how the forest structure indices are predicted for field plots using the ALS data.

All computations were conducted with the statistical software R (version 3.4.4.) and mainly using the packages `spatstat` (Baddeley et al. 2015), `spatialgraphs` (Rajala 2017), and `lidR` (Roussel and Auty 2018).

Preliminaries on spatial statistics

In this application, tree locations are mathematically expressed as a point pattern with a finite number of n trees observed on a field plot $W \subset \mathbb{R}^2$. Each observed point pattern is interpreted as a realization of a planar point process, which is assumed to be translation and rotation invariant with intensity λ. Here, λ can be interpreted as the tree density per square meter.

A point pattern is called completely spatially random (CSR) if there is no interaction between the points. Comparing to the CSR case, interaction between the points may result in either larger inter-point distances and regular patterns or smaller inter-point distances and clustered patterns. Regularity and clustering may also occur in the same pattern, but at different distances. Due to the small field plot size in this study, distances only up to 4.5 meters were taken into account and, thus, the spatial structure of forests, clustering or regularity, was considered only within this range.

Let us now consider a random set Ξ of discs with a random radius $\mathcal{R}$ centered at random locations forming a point pattern in an observation window W. For instance in this application, Ξ consists of the canopy patches. The empty-space function F then gives the cumulative distribution function of the distance r from an arbitrary location s in the 'empty' space $W \backslash \Xi$ to the nearest point in the random set Ξ (see Figure 3).

In the case of the Boolean model, which serves as a reference model with discs located uniformly on W, a theoretical F-function for all distances $r > 0$ is given by (Chiu et al. 2013, pp. 87)

$$F_{\text{theo}}(r) = 1 - \exp(-\lambda\pi r(2\mathbb{E}[\mathcal{R}] + r)), \qquad (1)$$

where the area fraction of the discs $p = 1 - \exp(-\lambda\pi\mathbb{E}[\mathcal{R}^2])$ can be used to calculate the expected number of disc centers per unit area (λ). Thus, λ in (1) can be replaced by $-\log(1-p)/(\pi\mathbb{E}[\mathcal{R}^2])$.

The empty-space function F can be defined for a point pattern accordingly. The theoretical F-function in the CSR case is

$$F_{\text{theo}}(r) = 1 - \exp(-\lambda\pi r^2) \qquad (2)$$

for distances $r \geq 0$. The point density λ is usually estimated by the number of points observed in W divided by the area of W, i.e. $n/|W|$. In order to obtain an unbiased estimate for the empty-space function, the spatial Kaplan-Meier estimator was used to correct for unobserved points outside the observation window W (Baddeley and Gill 1997).

The empty-space function F is also often called the spherical contact distribution function and denoted by H_s (Chiu et al. 2013, pp. 42, 87, 115).

Forest structure indices

Aggregation index The spatial forest structure has long been quantified by the aggregation index R (Clark

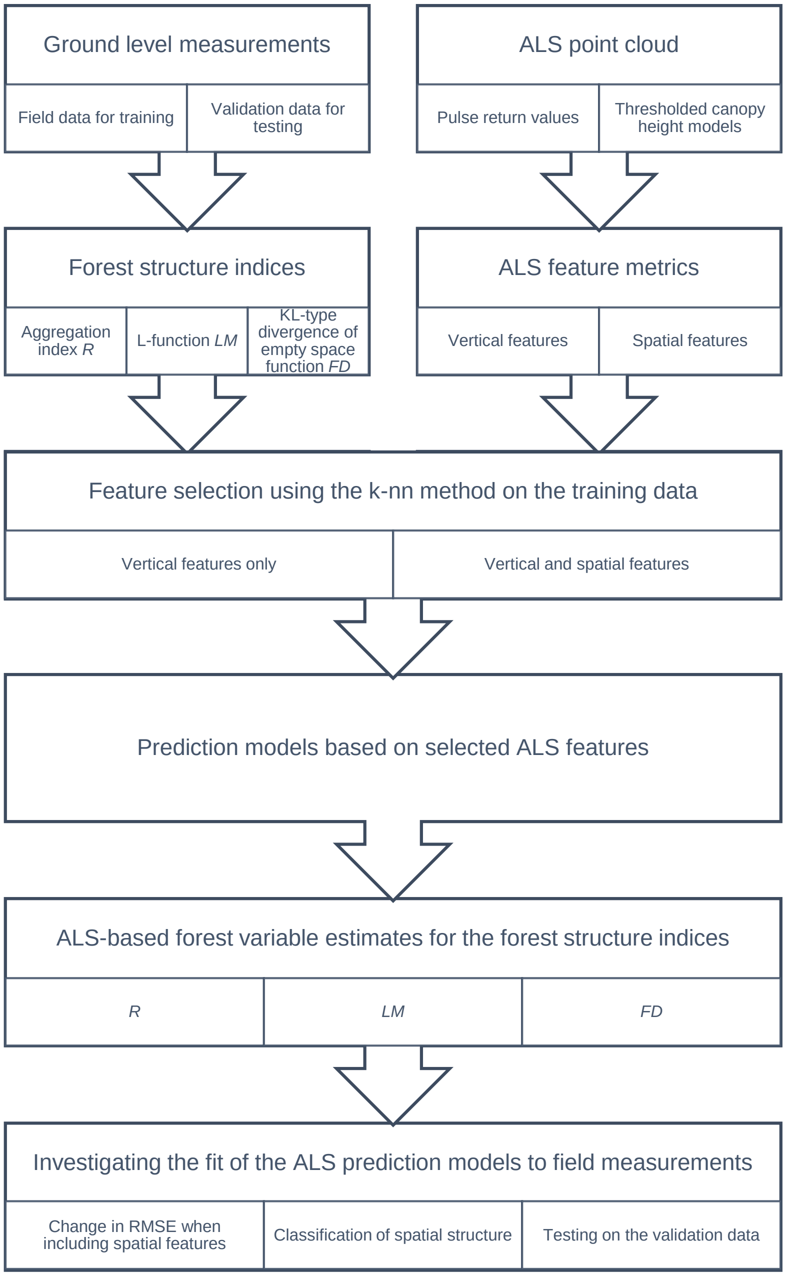
Ground level measurements
Field data for training
Validation data for testing
ALS point cloud
Pulse return values
Thresholded canopy height models
Forest structure indices
Aggregation index *R*
L-function *LM*
KL-type divergence of empty space function *FD*
ALS feature metrics
Vertical features
Spatial features
Feature selection using the k-nn method on the training data
Vertical features only
Vertical and spatial features
Prediction models based on selected ALS features
ALS-based forest variable estimates for the forest structure indices
R
LM
FD
Investigating the fit of the ALS prediction models to field measurements
Change in RMSE when including spatial features
Classification of spatial structure
Testing on the validation data

Figure 2: Flowchart over the methodology divided into three pipelines. In the first, ground level measurements from the field data (training set) and validation data (testing set) are summaries into three different forest structure indices. The second pipeline covers the processing of the ALS point cloud data based on the actual pulse return values and thresholded canopy height models. The latter is used to define new spatial features. Both spatial features and features from vertical summaries of the pulse return values make up for the ALS feature metrics used in the prediction model fitting. The third pipeline includes the model fitting steps of feature selection, obtaining estimates of the forest structure indices and goodness-of-fit analyses.

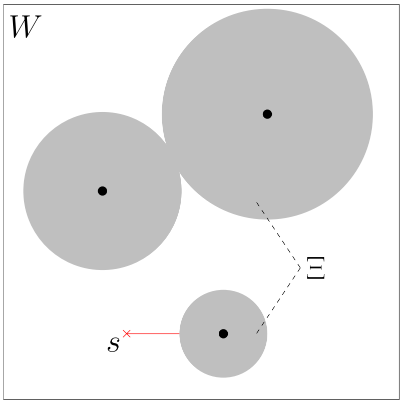

Figure 3: Schematic illustration of the random set Ξ (e.g. canopy patches) in an observation window W and an arbitrary location s in $W \setminus \Xi$ and its shortest distance to Ξ (red solid line).

and Evans 1954). It gives information about the spatial structure of trees with locations $(x_1, \ldots, x_n)$ based on their nearest neighbors $\mathrm{nn}(x_i)$, $i = 1, \ldots, n$, and is estimated by

$$R = \frac{2}{\sqrt{n|W|}} \sum_{i=1}^{n} \|x_i - \mathrm{nn}(x_i)\|, \tag{3}$$

where $R \approx 1$ in the CSR case, $R > 1$ indicates regularity and $R < 1$ clustering. In theory, the aggregation index can obtain values between zero and 2.1491. For the plots in Figure 4, $R_{\text{reg}} = 1.17$ and $R_{\text{clu}} = 0.76$.

***L*-function** A commonly used second-order characteristic for point patterns is Ripley's K-function (Ripley 1976): $\lambda K(r)$ gives the expected number of points (trees) of X within a ball $b(o, r)$ without counting o itself given that there is a point of X in o.

The L-function is a variance stabilizing transformation of the K-function and is given by

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \quad \text{for}\, r \geq 0. \tag{4}$$

In the CSR case, $L(r) - r = 0$ for all $r \geq 0$. This fact can be used in a test for CSR based on the test statistic

$$\tau = \max_{r \leq r_t} |\widehat{L}(r) - r| \tag{5}$$

with Ripley's isotropic edge corrected estimator $\widehat{L}$ and user-specified maximum radius r_t. The CSR hypothesis can be rejected at a 5% significance level if

$$\tau > \frac{1.45\sqrt{|W|}}{n}, \tag{6}$$

where $|W|$ denotes the area of W (Chiu et al. 2013, pp. 57 f., 139 ff.). The rule in (6) may depend on the choice of r_t if chosen too small and it deserves mentioning that 80% of the values for τ were obtained at inter-point distances r smaller than 3.47 m for the 2013 data. Furthermore, a sensitivity analysis in the validation data from 2014 with r_t ranging from 3 to 11 meters showed no important differences to the final choice of $r_t = 4.5$ m.

In order to determine the degree of regularity or clustering in the pattern, we used $LM = \widehat{L}(r_\tau) - r_\tau$, where r_τ denotes the distance at which the maximum difference from zero occurs, and negative and positive values indicate regularity and clustering, respectively. For the plots in Figure 2, $LM_{reg} = -1.03$ and $LM_{clu} = 1.31$.

KL-type divergence In addition to the L-function, we define a new forest structure index based on a Kullback-Leibler-type (KL-type) divergence of the estimated empty-space function $\widehat{F}$ from its theoretical counterpart $\widehat{F}_{\text{theo}}$ for point patterns (2) as

$$FD = D_{KL}(\widehat{F} \| \widehat{F}_{\text{theo}}) = \int_0^{r_t} \widehat{F}(r) \log\left(\frac{\widehat{F}(r)}{\widehat{F}_{\text{theo}}(r)}\right) dr, \tag{7}$$

for a chosen upper limit $r_t > 0$ of considered distances. D_{KL} is a simpler version of the cumulative Kullback-Leibler information (Crescenzo and Longobardi 2015). We refer to this new index FD as the KL-type divergence. $FD \approx 0$ in the CSR case, $FD > 0$ indicates regularity and $FD < 0$ clustering. For the plots in Figure 4, $FD_{\text{reg}} = 23$ and $FD_{\text{clu}} = -33$.

ALS feature metrics

The ALS feature metrics included in our study were divided into two groups. The first group includes the vertical features (features 1-62 in Table 2). The second group is formed by the spatial features extracted from thresholded CHMs (features 63-98 in Table 2) including new features (features 79-98) based on the Euler number and empty-space function F.

The ALS feature metrics were determined for the circles with a radius of 9 m covering the respective field plots. They were calculated from the CHMs that were first determined for slightly larger circles, with a buffer zone of 3 m, using the R package `lidR` (Khosravipour et al. 2014, Roussel and Auty 2018). Beforehand, first pulse returns below 1.3 m, but above the ground, were set to ground level.

Definition of spatial ALS features

The spatial ALS features are defined based on a thresholded CHM. In particular, each CHM was divided

Vertical features	
1	Height of the canopy
2	Minimum height of first returns
3	Maximum height of first returns
4	Mean height of first returns
5	Standard deviation of heights of first returns
6	Skewness of heights of first returns
7	Kurtosis of heights of first returns
8	Width of range of heights of first returns
9-15	The features similar to 2-8 for last returns
16	Proportion of canopy returns, all pulse returns
17-27	Percentiles (5, 10, 20, …, 90, 95%) for first returns
28-38	Same features as 17-27 for last returns
39-49	Cumulative proportions of foliage returns 0-5, 5-10, 10-20, …, 90-95%
50-60	Same features as 39-49 for last returns
61-62	Mean intensity (first and last returns)
Spatial features at 80, 60, 40, 20% levels of the maximum height	
63-66	Number of patches
67-70	Average size of patches in number of pixels
71-74	Standard deviation of size of patches
75-78	Average number of same pixel type in a 4-neighbourhood
79-82	Euler number for TCHMs
83-86	Integrated deviation of F-function from theoretical reference
87-90	KL-type divergence of F-function from theoretical reference
91-94	Pairwise integrated difference of F-functions between TCHMs
95-98	Pairwise KL-type divergence of F-functions between TCHMs

Table 2: ALS feature metrics calculated on the basis of common summaries of pulse return values (vertical features) and spatial information from thresholded canopy height models (TCHM). The foliage returns were in the range of the height of the first and last returns, respectively.

into two regions according to a threshold of $q \cdot hmax$ for $q = 80, 60, \ldots, 20\%$ given the maximum height $hmax$ of the CHM. Values above the threshold formed the canopy patches at height level q. Values below the threshold are referred to as gaps or empty space.

Following Packalen et al. (2013) and other work on quantifying canopy patch characteristics, we calculated the number of patches (features 63-66 in Table 2), the average patch size (features 67-70), standard deviation of the patch size (features 71-74), and the average number of pixels in a 4-neighborhood of the same type as the focal pixel (either patch or gap pixel, features 75-78).

We additionally included the Euler number (features 79-82) to the set of features. It gives the number of canopy patches minus the number of gaps and therewith it is an easy measure of the canopy complexity. As an illustration for why spatial ALS features at different height levels and including the Euler number can be meaningful especially in combination, let us consider the regular and clustered plots presented in Figure 4. All plots have approximately the same number of trees (around 25) and almost no gaps at the 80% height level, but the regular pattern of trees has more canopy patches (15) than the clustered one (8). At the 40% height level, the largest difference between the two plots can be observed in the Euler number. Due the differences in spatial forest structure, the regular plot has only one canopy patch and shows many gaps resulting in a very low Euler number of -14 whereas there are still 4 canopy patches on the clustered plot and only a few gaps resulting in an Euler number of -3.

In order to include information about the gaps or empty space, we introduce spatial ALS features based on the empty-space function F. For each thresholded CHM, F was estimated as the empirical cumulative distribution function of distances from all empty space pixels to the nearest canopy pixel. The estimator for F_{theo} was based on equation (1), where the random radius $\mathcal{R}$ was determined by the average of the largest distance between any two pixel of the canopy patches.

We used two alternative ways to summarize differences between the estimated F-function and its theoretical counterpart to single numbers. Namely, we consid-

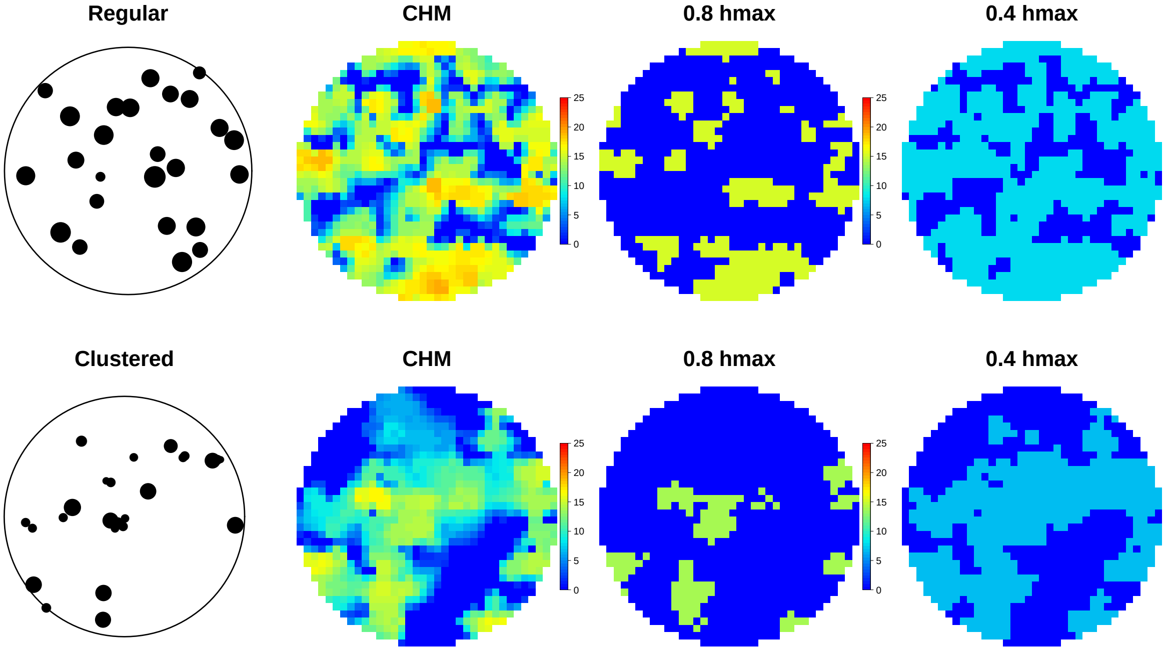

Figure 4: Examples of canopy height models (CHM) and thresholded CHMs for a regular (top row) and a clustered (bottom row) pattern of trees. The thresholds were selected at 80% and 40% of the maximum height (hmax) of each CHM. The points of the patterns of trees on the left have been scaled according to their estimated tree height.

ered the integrated squared difference

$$\begin{aligned} D_I &= D_I(\widehat{F}, \widehat{F}_{\text{theo}}) \\ &= \text{sgn} \cdot \int_0^{r_t} (\widehat{F}(r) - \widehat{F}_{\text{theo}}(r))^2 dr \end{aligned} \qquad (8)$$

and the proposed KL-type divergence D_{KL} (see equation (7)) with a chosen upper limit $r_t > 0$ (see Section). The larger the absolute value of D_I or D_{KL}, the larger the difference to the CSR case in terms of space around a random location in the empty space. To make a differentiation between regularity and clustering possible also by D_I, the integral in (8) was multiplied by the sign of the maximal difference to F_{theo} (sgn). Thus, for a height level q, positive and negative signs of both feature metrics relate to regular and clustered patterns of trees that have heights larger or equal to $q \cdot hmax$, respectively. It should be noted that the values of D_{KL} tend to be generally smaller in their absolute value than the values of D_I.

The summaries D_I and D_{KL} were also used to compare height layers with each other (features 91-98). It can be expected that the higher layers appear more regular than the lower layers, but that the difference is larger for clustered than for regular plots. For instance, for the example in Figure 4, $D_{KL}(F^{(q=0.8)} \| F^{(q=0.4)}) \approx 40$ for the regular example, but 83 for the clustered plot. This indicates that the upper layer appears more regular than the lower layer in both cases, but that the difference is larger for the clustered plot.

Feature selection and prediction of indices

Feature selection and prediction of the forest structure indices were carried out using a genetic algorithm along with the k-nearest neighbor method (k-nn method) as described by Tomppo and Halme (2004) and Tuominen et al. (2018, 2017) for continuous and discrete variables respectively.

In a preceding step, all features were standardized to have the same variation. Then the genetic algorithm implemented in the `Genalg` package in R (Willighagen and Ballings 2015) was used to select the relevant features $f_{l,\cdot}$, $l = 1, \ldots, m$, and to determine the weights ω_l for them according to a fitness function based on plot-level RMSE and absolute bias, which we set to have equal importance. The selection of the optimal k among the tested values (3-6) was included in the routine. The forest variable y of the plot p was predicted using the set of k nearest neighboring plots I_p with $p \notin I_p$:

$$\widehat{y}_p = \sum_{i \in I_p} w_i y_i \quad \text{with} \quad w_i = d_{i,p}^{-g} / \sum_{j \in I_p} d_{j,p}^{-g}, \qquad (9)$$

where the weights w_i were determined by the distance between each neighbor i and the plot p in the feature space, namely

$$d_{i,p}^2 = \sum_{l=1}^{m} \omega_l^2 (f_{l,i} - f_{l,p})^2, \tag{10}$$

and the factor g. Different values for g (0-3) were tested, where $g > 0$ implies that neighbors with smaller distances get larger weights. The final values for k and g are given in Table 3.

Results

Relevance of spatial features

The spatial features were sufficiently correlated with the field data, such that seven spatial features were selected by the genetic algorithm for R as well as for FD and two for LM. The spatial features made up 47%, 39%, and 22%, respectively, of all selected features. The spatial features selected most often were the average number of pixels of same type in a 4-neighborhood (features 75-78 in Table 2) and F-function based features (features 83-98 in Table 2). The integrated difference measure D_I in (8) was chosen mostly for the comparison to the theoretical F-function on the 80, 40, and 20% height level and the KL-type measure D_{KL} in (7) for the comparison of the F-function of two different height levels.

Comparing the predicted values of the indices obtained with spatial features to predicted indices obtained without them, we observed a 8.4% reduction of RMSE for the forest structure index FD. There was a small improvement for the other variables R and LM as well, where the RMSE was reduced by 3.3% and 2.1%, respectively.

Prediction of forest structure indices

The spatial features improved the predictions and all estimates of the studied forest structure indices were practically unbiased. However, the RMSEs were rather large (see Table 3).

There was a tendency to overestimate the negative values and underestimate the positive values, i.e. more patterns of trees appeared random and fewer forest were classified as regular or clustered based on the ALS data (see Figure 5).

Classification of forest structure

To compare our results to Pippuri et al. (2012) and Packalen et al. (2013), the forest structure indices were used in classifications of field plots into regular, random, and clustered patterns. In contrast to these previous works, we used simple cut-off values in a more practical approach. These cut-off values were 0.85 and 1.15 for R and ± 15 for FD. For LM, we used the rule given in (6). The overall accuracy was 62.2% for R, 60.6% for LM, and 59.3% for FD with a Cohen's kappa of 0.31, 0.14, and 0.23, respectively. Due to the relatively simple spatial structure of tree center points on mature field plots, it was expected to obtain the best results for all considered measures of regularity for this development class. Interestingly, R and FD obtained their highest values for advanced thinning plots, whereas LM performed best on mature plots. All in all, the overall accuracy, Cohen's kappa, and differentiation of clustered and regular patterns were the best for R followed by LM. However, more regular plots were missclassified as clustered (and vice versa) than for FD.

Index	g	SD	RMSE	Bias
R	0.9	0.202	0.178	0
LM	0.8	1.018	0.843	0
FD	1.8	21.411	16.453	0

Table 3: Outcome of the genetic algorithm for feature selection and prediction of the forest structure indices R, LM, and FD for the 2013 data including spatial features. In all cases $k = 6$ neighbors were considered, but different scaling values g were given to the weights. SD is the standard deviation of the field data-based values.

Validation study

The data from 2014 was used in a validation study. The features and weights selected for the 1161 circular plots measured in 2013 were used to predict the forest structure indices of the 120 subplots of size 256 m^2 measured in 2014 by the k-nn method. When classifying the spatial forest structure with the same rules as applied to the 2013 data (see Section), the overall accuracy was 55.8% for R, 39.2% for LM, and 68.3% for FD. The values for Cohen's kappa were 0.07, 0.01, and 0.01, respectively. Upon recalling that LM performed best for mature plots and that the validation data only contained thinning plots, it is not surprising that the overall accuracy for LM was lower than for the 2013 test data. Also not surprising, FD achieved the highest overall accuracy as this summary appeared to perform best for detecting clustered forests. However, with the ALS-based FD no clustered plots and only four regular were classified correctly, but 78 out of 94 (83%) random plots were in agreement with the field data classification. The ALS-based R and LM prediction classified around 62% of the regular plots correctly, but failed to detect any prominent number of clustered patterns.

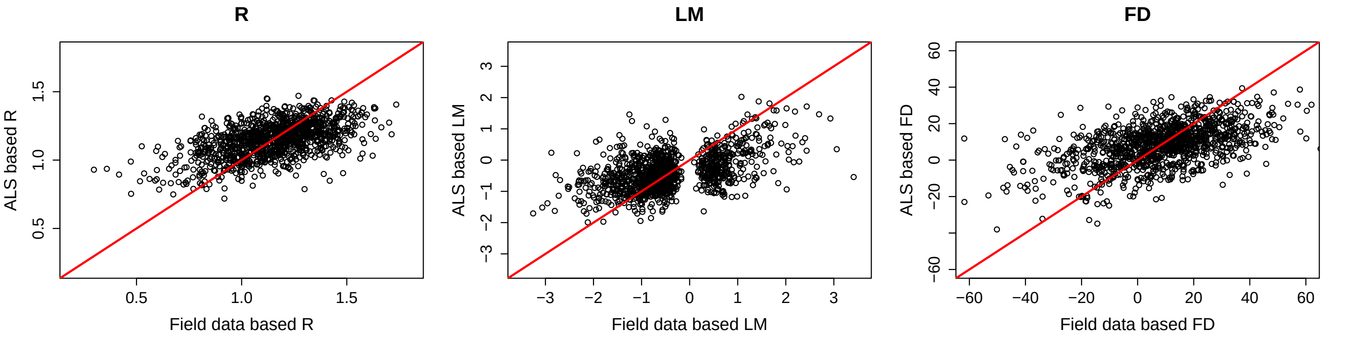

Figure 5: ALS-based forest variable estimates versus field data-based values for the forest structure indices R, LM and FD.

Discussion

The primary objective of this study was to identify novel spatial ALS features that can improve the prediction of spatial forest structure. The relevance of our proposed spatial features was evaluated by examining their correlation to the field data measured as proportion of selected features. Furthermore, estimates based on prediction models with and without spatial features were compared in term of their RMSEs. The goodness-of-fit was further analyzed by plotting the predicted versus the field-based values of the forest structure indices and by studying confusion matrices of classifications derived from the estimates. As a secondary objective, we examined challenges related to small field plots and low resolution ALS data. The following discussion focuses first on the classification of spatial forest structure into regular, random, and clustered arrangements of trees. Then, the focus is shifted towards the new spatial features and what type of ALS metrics appear best for the prediction and classification.

Spatial forest structure classification

The three field data-based classifications agreed on 205 regular, 373 random, and 64 clustered patterns for the field data, which means only a 44% agreement and, hence, shows that the ground truth of the forest structure classification differs for each forest structure index. Classifying the original 32 m × 32 m field plots in an additional significance test for complete spatial randomness with the global envelope test based on the F- and Ripley's K-function (Mrkvička et al. 2017, Myllymäki et al. 2017, Myllymäki and Mrkvička 2020), the CSR hypothesis was rejected for all plots leading to the same classification as with FD. Testing the 16 m × 16 m subplots, however, 69 regular, 42 random patterns, and 9 clustered patterns were obtained. Figure 6 shows this phenomena more precisely: it shows for the proportion of rejections of CSR among 32 circular sample plots with radius 5-16 m placed in the centre of the 32 m × 32 m plots.

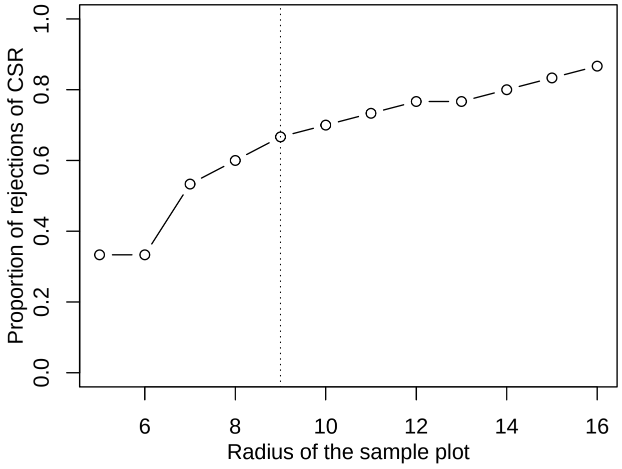

Figure 6: Proportion of rejections of the CSR in 32 circular sample plots with different radius. See text for details. The vertical line indicates the 9 m radius of the field plots used in the main study.

The increasing trend along the radius means that the larger the plot size, the better a pattern can be deduced to deviate from the CSR case. This is a common feature of statistical tests. These classification and test results lead to two conclusions. First, the quality of the differentiation of regular and especially clustered patterns from random patterns depends on the field plot size. Second, among the classifiers studied here no index was best for detecting both regular and clustered patterns at the same time. R and LM appeared more suitable for regular patterns and FD for clustered when the field plot was a circular plot with 9 m radius.

We had not expected that the aggregation index R would perform well in the classification of the 2013 data as it is a rather crude summary of the spatial distribution. In fact, this simplicity might facilitate its prediction, especially for small field plots. The validation study has shown, however, that even though R was predicted more accurately, it may not be the best classifier for the spatial structure of trees in all possible forest scenarios especially with clustered patterns of trees. Therefore, we expect that better results may be obtained with FD or other summaries for larger field plots. Our study showed that quantifying clustering and regularity of forests from small field plots is itself a difficult task. It is not only statistically difficult to separate clustered and regular forests from CSR based on a pattern of only a few trees, but also different indices can lead to different classifications.

Spatial ALS features

In this study, the spatial ALS features describing the canopy complexity were the most informative. The empty-space function based feature metrics proved to be a relevant addition to the 4-neighborhood based summary inspired by Packalen et al. (2013) (features 75-78 in our Table 2) as both were selected by the genetic algorithm used in the forest variable prediction. In a side study not presented here, we implemented the original set of features by Packalen et al. (2013) (their Table 2), but found that merging the neighborhood counts for patch and gap pixels into one feature led to smaller RMSEs. We also found that the Euler number was selected more often, together with the number of patches (features 67-70), if features 75-78 were not included. Consequently, the Euler number in combination with the number of patches appeared to contain valuable information for the prediction of the studied forest structure indices, but the same information was apparently better captured by the the 4-neighborhood based features in our data. In the same side study, we further included spatial ALS features to predict other forest variables such as breast-height diameter distributions, development class and species mingling. Not surprisingly, the species related forest variables were poorly predicted with and without spatial ALS features. Even though, the generic algorithm picked spatial ALS features as well, no prominent improvement was achieved compared to a prediction without them. On the other hand, we obtained small improvements in RMSE for the structural variables with using the spatial features.

The spatial ALS features were selected by the genetic algorithm for the prediction in the k-nn method of all studied forest structure indices. The original aim of this study was to predict even the degrees of clustering and regularity. Hence, we summarized the spatial forest structure by the continuous indices. However, it appeared that this task was a bit too ambitious given the relatively small field plots. Larger plots should result in smaller RMSEs. It seems that the size of the field plots used for training the algorithm should preferably be larger than 256 m^2 if aiming at prediction of forest structure variables similar to those considered here. As forestry decisions are often done for stands with uniform structure, it could be interesting to study the spatial structure for forest stands, instead of field plots of fixed size (see e.g. Tuominen and Haapanen 2011).

Packalen et al. (2013) also used the so-called L-function in the classification rule of the field data. The resolution of their ALS data was slightly lower, but still comparable, and they used a smaller number (79 in total) of larger plots of size 20 m × 20 m and 30 m × 30 m. Still, we obtained comparable results with those obtained by the AREA method in Packalen et al. (2013) for the 2013 data.

Pippuri et al. (2012) also had slightly lower resolution ALS data and chose the so-called (spatial) t_N-index for the classification of 28 microstands in Southern Finland of sizes between 0.2 and 0.7 ha. They achieved better classification results, however, this could be due to the large size of their field plots and greater differences in tree density between regular and clustered patterns compared to our study.

All of the mentioned work with predicting spatial forest structure employed rather low resolution ALS data. Given that this was the case, and that the forest structure seemed to be more accurately estimated for advanced thinning plots than mature plots, one of the reasons why regular plots have been missclassified as clustered could be the large canopies of mature trees which may have led to a patch structure resembling a clustered pattern. On the other hand, the main reason for clustered plots being missclassified as regular plots appeared to be an extremely large variation in tree size, where small trees where covered by large trees resulting in a gap structure of a regular pattern. Higher resolution ALS data might lead to better results, although similar problems exist even then.

All in all, the presented study can be interpreted as a pilot study on these aspects, giving indications for future studies. It would be interesting to investigate the usefulness of the spatial features at several height layers in the future not only for spatial structure indices using larger field plots, but also for other forest variables related to structural properties of the forests horizontally and vertically (see e.g. Sverdrup-Thygeson et al. 2016). After appropriate validation with the field data, wall-to-wall ALS data could be used to predict the spatial structure in a complete study region (see e.g. Luther et al. 2019, Schumacher and Nord-Larsen 2014).

Conclusion

New spatial ALS features were found for the prediction of forest structure variables in ALS-assisted inventories. Their application was focused on the prediction and classification of the spatial structure of forests. For the classification, three different spatial summary statistics referred to as spatial forest structure indices were used and their properties discussed. In general, we can recommend to include spatial ALS features from CHMs thresholded at more than one height level. For our data, relatively complex features were most informative that describe the spatial structure of the gaps with the empty-space function or that take the pixel type in a 4-neighborhood into account. The empty-space function appeared to be especially useful as a measure of differences in the gap structure at two different height levels. A simple alternative to describe the canopy complexity is offered by the Euler number, which is the number of vegetation patches minus the number of gaps.

This study has shown the potential of spatial analysis of ALS point clouds through CHMs thresholded at several height levels. Concurrently, there is a need for further investigation with higher resolution data or larger sample plots. Furthermore, in order to find the best alternatives for spatial structure classification of forests for certain needs, e.g. for finding clustered plots with a need of thinning, it would be interesting to compare the classifications given by different indices to classifications made in the field in future studies.

In conclusion, the presented methodology for spatial ALS features is practical and general enough to be applied to other forest variables such as conventional forest inventory attributes and different three-dimensional remote sensing scenarios.

Abbreviations

ALS: Airborne laser scanning
CSR: Complete spatial randomness
CHM: Canopy height model
dbh: Diameter at breast height measured at 1.3 m
RMSE: Root mean squared error

Declarations

Availability of data and material

The data is not owned by the authors and thus cannot be shared. R codes for index and feature calculations are available upon request.

Competing interests

The authors declare that they have no competing interests.

Funding

HH and AB were financially supported by the Academy of Finland (Project Number 304212) and MM similarly by the Academy of Finland (Project Numbers 295100, 306875 and 327211).

Author's contributions

All authors were involved in planning and conducting the study. AB was responsible for the feature selection and prediction of indices, while HH was the main person responsible for everything else. MM was the principal investigator of this study, had acquired the funding, supervised HH.

Acknowledgements

The authors would like to thank Kai Mäkisara for assisting with preparing the data for the study, and Jussi Peuhkurinen (Oy Arbonaut Ltd) and Annika Kangas together with other colleagues at Luke for valuable discussions. The Finnish Forest Centre provided the ALS data. The work was mainly done during the time when HH was at Natural Resources Institute Finland (Luke). The Background map series of National Land Survey of Finland were used for Figure 1; these open data is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

References

Baddeley, A., and R. Gill, 1997. Kaplan-Meier estimators of interpoint distance distributions for spatial point processes. Annals of Statistics 25(1):263–292.

Baddeley, A., E. Rubak, and R. Turner, 2015. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, London.

Chiu, S. N., D. Stoyan, W. S. Kendall, and J. Mecke, 2013. Stochastic Geometry and its Applications. 3^{rd} edition. Wiley.

Clark, P., and F. Evans, 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35(4):445–453.

Crescenzo, A. D., and M. Longobardi, 2015. Some properties and applications of cumulative Kullback-Leibler information. Applied Stochastic Models in Business and Industry 31(6):875–891.

Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan, 2008. Statistical Analysis and Modelling of Spatial Point Patterns. Statistics in Practice, 1^{st} edition. John Wiley & Sons, Ltd, Chichester.

Kane, V. R., R. F. Gersonde, J. A. Lutz, R. J. McGaughey, J. D. Bakker, and J. F. Franklin, 2011. Patch dynamics and the development of structural and spatial heterogeneity in pacific northwest forests. Canadian Journal of Forest Research 41(12):2276–2291.

Kane, V. R., A. R. Gillespie, R. McGaughey, J. A. Lutz, K. Ceder, and J. F. Franklin, 2008. Interpretation and topographic compensation of conifer canopy self-shadowing. Remote Sensing of Environment 112(10):3820–3832.

Khosravipour, A., A. Skidmore, M. Isenburg, T. Wang, and Y. Hussin, 2014. Generating pit-free canopy height models from airborne lidar. Photogrammetric Engineering & Remote Sensing 80:863–872.

Li, W., Z. Niu, S. Gao, N. Huang, and H. Chen, 2014. Correlating the horizontal and vertical distribution of lidar point clouds with components of biomass in a *Picea crassifolia* forest. Forests 5(8):1910–1930.

Luther, J. E., R. A. Fournier, O. R. van Lier, and M. Bujold, 2019. Extending ALS-based mapping of forest attributes with medium resolution satellite and environmental data. Remote Sensing 11(9).

Mrkvička, T., M. Myllymäki, and U. Hahn, 2017. Multiple Monte Carlo testing, with applications in spatial point processes. Statistics and Computing 27(5):1239–1255.

Myllymäki, M., T. Mrkvička, P. Grabarnik, H. Seijo, and U. Hahn, 2017. Global envelope tests for spatial processes. Journal of the Royal Statistical Society Series B - Statistical Methodology 79(2):381–404.

Myllymäki, M., and T. Mrkvička, 2020. GET: Global envelopes in R. arXiv:1911.06583 [stat.ME] .

Packalen, P., J. Vauhkonen, E. Kallio, J. Peuhkurinen, J. Pitkänen, I. Pippuri, J. Strunk, and M. Maltamo, 2013. Predicting the spatial pattern of trees by airborne laser scanning. International Journal of Remote Sensing 34(14):5154–5165.

Pippuri, I., E. Kallio, M. Maltamo, H. Peltola, and P. Packalen, 2012. Exploring horizontal area-based metrics to discriminate the spatial pattern of trees and need for first thinning using airborne laser scanning. Forestry 85(2):305–314.

Pukkala, T., 1990. A method for incorporating the within-stand variation into forest management planning. Scandinavian Journal of Forest Research 5(1-4):263–275.

Rajala, T., 2017. spatgraphs: Graph Edge Computations for Spatial Point Patterns. URL `https://CRAN.R-project.org/package=spatgraphs`. R package version 3.2-1.

Ripley, B. D., 1976. The second-order analysis of stationary point processes. Journal of Applied Probability 13(2):255–266.

Roussel, J.-R., and D. Auty, 2018. lidR: Airborne LiDAR Data Manipulation and Visualization for Forestry Applications. URL `https://CRAN.R-project.org/package=lidR`. R package version 1.4.2.

Schumacher, J., and T. Nord-Larsen, 2014. Wall-to-wall tree type classification using airborne lidar data and CIR images. International Journal of Remote Sensing 35(9):3057–3073.

Sverdrup-Thygeson, A., H. O. Ørka, T. Gobakken, and E. Næsset, 2016. Can airborne laser scanning assist in mapping and monitoring natural forests? Forest Ecology and Management 369:116–125.

Tomppo, E., 1986. Models and methods for analysing spatial patterns of trees. Ph.D. thesis, Communicationes Instituti Forestalis Fenniae 138.

Tomppo, E., and M. Halme, 2004. Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach. Remote Sensing of Environment 92(1):1–20.

Tomppo, E., J. Heikkinen, H. M. Henttonen, A. Ihalainen, M. Katila, H. Mäkelä, T. Tuomainen, and N. Vainikainen, 2011. Designing and Conducting a Forest Inventory - case: 9th National Forest Inventory of Finland. Springer, Dordrecht, The Netherlands.

Tomppo, E., N. Kuusinen, K. Mäkisara, M. Katila, and R. E. McRoberts, 2017. Effects of field plot configurations on the uncertainties of ALS-assisted forest resource estimates. Scandinavian Journal of Forest Research 32(6):488–500.

Tuominen, S., and R. Haapanen, 2011. Comparison of grid-based and segment-based estimation of forest attributes using airborne laser scanning and digital areal imagery. Remote Sensing 3:945–961.

Tuominen, S., R. Näsi, E. Honkavaara, A. Balazs, T. Hakala, N. Viljanen, I. Pölönen, H. Saari, and H. Ojanen, 2018. Assessment of classifiers and remote sensing features of hyperspectral imagery and stereo-photogrammetric point clouds for recognition of tree species in a forest area of high species diversity. Remote Sensing 10(5).

Tuominen, S., T. Pitkänen, A. Balazs, and A. Kangas, 2017. Improving Finnish multi-source national forest inventory by 3D aerial imaging. Silva Fennica 51(4):Article ID 7743.

Uuttera, J., A. Haara, T. Tokola, and M. Maltamo, 1998. Determination of the spatial distribution of trees from digital aerial photographs. Forest Ecology and Management 110(1):275–282.

Wang, X., G. Zheng, Z. Yun, and L. M. Moskal, 2020. Characterizing tree spatial distribution patterns using discrete aerial lidar data. Remote Sensing 12(4):Article 712.

Willighagen, E., and M. Ballings, 2015. genalg: R Based Genetic Algorithm. URL https://CRAN.R-project.org/package=genalg. R package version 0.2.0.

Zhang, Z., L. Cao, and G. She, 2017. Estimating forest structural parameters using canopy metrics derived from airborne lidar data in subtropical forests. Remote Sensing 9(9):Article 940.