Identification of linear vegetation elements in a rural landscape using LiDAR point clouds

Chris Lucas[a,b,c], W Bouten[a], Zsofia Koma[a], W. Daniel Kissling[a], Arie C. Seijmonsbergen[a,*]

[a] *Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands*
[b] *Spatial Information Laboratory (SPINLab), Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam*
[c] *Geodan, President Kennedylaan 1, 1079 MB Amsterdam, The Netherlands*

Abstract

Modernization of agricultural land use across Europe is responsible for a substantial decline of linear vegetation elements such as tree lines, hedgerows, riparian vegetation and green lanes. These linear objects have an important function for biodiversity, e.g. as ecological corridors and local habitats for many animal and plant species. Knowledge on their spatial distribution is therefore essential to support conservation strategies and regional planning in rural landscapes, but detailed inventories of such linear objects are often lacking. Here, we propose a method to detect linear vegetation elements in agricultural landscapes using classification and segmentation of high resolution LiDAR point data. To quantify the 3D structure of vegetation we applied point cloud analysis to identify features based on the local neighborhood. As a preprocessing step, we removed planar surfaces, such as grassland, bare soil and water bodies, from the point cloud using scatter information. We then applied a random forest classifier to separate the remaining points into 'vegetation' and 'other'. Subsequently, a region growing algorithm allowed to segment 2D rectangular objects, which were then classified into linear objects based on their elongatedness. We evaluated the accuracy of the linear objects against a manually delineated set. This assessment showed that the majority of vegetation objects were correctly identified. These results are a promising first step for testing our method in other regions and for upscaling it to broad spatial extents. This would allow producing detailed inventories of linear vegetation elements at regional and continental scales in support of biodiversity conservation and regional planning in agricultural and other rural landscapes.

Keywords: Agricultural landscapes, Classification, LiDAR, Linear vegetation, Object recognition, Point cloud, Rectangularity, Segmentation

1. Introduction

The European landscape has dramatically changed during the Holocene as a result of human impact and climatic change (Turner, 1989; Marquer et al., 2017). Especially since the industrial revolution, landscapes

*Corresponding author. E-mail address: A.C.Seijmonsbergen@uva.nl (A. C. Seijmonsbergen).

have been deforested and reshaped into rural and agricultural landscapes. These are dominated by a mosaic of grasslands, forests and urban areas, separated or connected by linear landscape elements such as roads, ditches, tree lines, lynchets and hedgerows (Bailly et al., 2008; Meyer et al., 2012; Van der Zanden et al., 2013). The distribution, abundance and richness of species in these landscapes is related to the amount, height, length and quality of linear vegetation elements (Aguirre-Gutiérrez et al., 2016; Spellerberg and Sawyer, 1999; Croxton et al., 2005). The same holds true for the dispersal of seeds and the flow of matter, nutrients and water (Turner, 1989; Burel, 1996). Additionally, linear infrastructures such as roads and railways form barriers which lead to habitat fragmentation. In contrast, green lanes which are flanked by hedges and/or tree lines may form connecting corridors. Hence, linear vegetation elements are of key importance for biodiversity in agricultural landscapes. A wider audience has become aware that historic agricultural practices are part of the cultural heritage (Jongman, 2004; Gobster et al., 2007) and need to be conserved. However, the occurrence of green lanes and hedgerows has strongly diminished in many countries (Boutin et al., 2001; Stoate et al., 2001). This is mostly a consequence of larger agricultural fields, monocultures and a reduction in non-crop features which reduces the complexity and diversity of landscape structure (Croxton et al., 2005). Detailed knowledge of the spatial occurrence, current status, frequency and ecological functions of linear vegetation elements in a landscape is therefore of key importance for biodiversity conservation and regional planning.

The mapping of linear vegetation elements has traditionally been done with visual interpretations of aerial photographs in combination with intensive field campaigns (Aksoy et al., 2010). However, this approach is time-consuming and has limited transferability to larger areas. New methods have therefore been developed that use raster images to map linear vegetation elements by using their spectral properties in visible or infrared wavelengths, e.g. from SPOT, ASTER and Landsat imagery (Thornton et al., 2006; Vannier and Hubert-Moy, 2014; Tansey et al., 2009). This allows an automated and hierarchical feature extraction from very high resolution imagery (Aksoy et al., 2010). Despite these developments, comprehensive high-resolution inventories of linear vegetation elements such as hedgerows and tree lines are lacking at regional and continental scales. The lack of such high-resolution measurements of 3D ecosystem structure across broad spatial extents impedes major advancements in animal ecology and biodiversity science, e.g. for predicting animal species distributions (Kissling et al., 2017). On a European scale, density maps of linear vegetation elements (and ditches) have been produced at 1 km^2 resolution through spatial modeling of 200,000 ground observations (Van der Zanden et al., 2013). However, these maps strongly depend on spatial interpolation methods as well as regional environmental and socio-economic variation and therefore contain a considerable amount of uncertainty in the exact spatial distribution of linear vegetation elements in the landscape. High-resolution measurements of 2D and 3D ecosystem structures derived from cross-national remote sensing datasets are therefore needed to identify and map linear vegetation elements across broad spatial extents (Kissling et al., 2017).

An exciting development for quantifying 3D ecosystem structures is the increasing availability of high-resolution remote sensing data derived from Light Detection and Ranging (LiDAR) (Lim et al., 2003). LiDAR data have important properties which are useful for the detection, delineation and 3D characterization of vegetation, such as their physical dimensions x, y, z, laser return intensity, and multiple return information (Lefsky et al., 2002; Eitel et al., 2016). Vegetation partly reflects the LiDAR signal and usually generates multiple returns, including a first return at the top of the canopy and a last return on the underlying terrain surface. This provides valuable information for separating vegetation from non-vegetation (Lim et al., 2003). Moreover, the intensity values describe the strength of the returning light, which depends on the type of surface on which it is reflected and therefore provides information on the surface composition (Song et al., 2002). The shape and internal structure of vegetation can be analyzed by classifying information from the different return values and a variety of features, which can be calculated from the point cloud (Lim et al., 2003; Weinmann et al., 2015). Some applications of using airborne LiDAR data to quantify linear elements in agricultural landscapes already exist, e.g. the extraction of ditches in a Mediterranean vineyard landscape (Bailly et al., 2008). However, the characterization of linear vegetation elements in rural and agricultural landscapes from LiDAR point clouds is mostly lacking. Nevertheless, the increasing availability of nation-wide and freely accessible LiDAR data in several European countries provides exciting new avenues for characterizing 3D vegetation structures in agricultural landscapes (Kissling et al., 2017).

Here, we present a transparent and accurate method for classifying linear vegetation elements from LiDAR point clouds in an agricultural landscape. We develop the method using free and open source data and analysis tools and apply it for characterizing various linear vegetation elements in a rural landscape of the Netherlands containing agricultural fields, grasslands, bare soil, roads and buildings. While the identification of linear objects (e.g. the automated delineation of roads) is often based on raster-based remotely sensed imagery (Quackenbush, 2004), the detection of linear vegetation objects is more complex due to their 3-dimensional shape, size and variety. We therefore use a method which allows us to directly classify the point cloud using machine learning algorithms (Yan et al., 2015). Using fourteen features based on echo, local geometric and local eigenvalue information of the LiDAR point cloud, we apply a machine learning algorithm to classify the vegetation points in the point cloud. We then use a region growing algorithm to segment the classified vegetation points into rectangular objects, and apply elongatedness as a criterion to classify linear objects. The accuracy of the method is tested against manually annotated datasets based on high resolution orthophotos and field surveys. Our method provides a promising first step for upscaling the detection of linear vegetation objects in agricultural landscapes to broad spatial extents.

2. Data and study area

2.1. LiDAR and orthophoto data

Raw LiDAR point cloud data were retrieved from "Publieke Dienstverlening op de Kaart" (PDOK), an open geo-information service of the Dutch government.[1] The data are part of the "Actueel Hoogtebestand Nederland 3" (AHN3) dataset, which is collected between 2014 and 2019. The density of the LiDAR data is around 10 pulses/m^2 and includes multiple discrete return values (which can result into effective point densities of over 20 point/m^2) as well as intensity data. The dataset is collected in the first quarter of each year when deciduous vegetation is leafless (AHN, 2016). Nevertheless, the return signal is sufficiently strong to retrieve a useful scan of the vegetation cover. Freely available very high resolution (VHR) true color orthophotos from PDOK with a resolution of 25cm were consulted for validation purposes.[2]

All data were analyzed using free and open source software.[3] The scripting was performed in Python (3.6.5) using the NumPy (1.14.2) (Walt et al., 2011), SciPy (1.1.0) (Jones et al., 2001), pandas (0.22.0) (McKinney et al., 2010), scikit-learn (0.19.1) (Pedregosa et al., 2011), and CGAL (4.12) (CGAL Project, 2018) libraries. PDAL (1.7.2) (PDAL contributors, 2018) was used for preprocessing and downsampling data. CloudCompare (v2.10alpha) (CloudCompare, 2018) was used for visualizing the point cloud and for the manual classification.

2.2. Study area

The case study area is located in a rural landscape in the center of the Netherlands (figure 1). The area is about 1.6 km from east to west and 1.2 km from north to south, spanning an area of almost 2 million square meters. The point density of the point cloud in the area is 22.49 points/m^2. The area contains numerous linear vegetation elements of varying geometry, ranging from completely straight to curved, isolated or connected to other linear or nonlinear objects. Examples of vegetation and non-vegetation elements are planted forest patches, hedges, green lanes, isolated farms, ditches, a river, dykes and a road network (figure 1). This heterogeneity within a small area ensured that both the classification of vegetation and delineation of linear objects can be efficiently trained and tested.

3. Method

The workflow consisted of three main routines: feature extraction, vegetation classification, and linear object segmentation (figure 2). Each of these is explained in more detail below.

[1] https://www.pdok.nl/nl/ahn3-downloads
[2] https://www.pdok.nl/nl/service/wms-luchtfoto-beeldmateriaal-pdok-25-cm-rgb
[3] https://github.com/clucas111/delineating-linear-elements

Figure 1: Location of the study area in the central part of The Netherlands. The true color aerial photo (PDOK) in the center shows several linear objects in the rural landscape related to agricultural fields, grasslands, bare soil and infrastructure such as (un)paved roads and farmhouses. The numbered photos show a selection of the variety of linear vegetation elements, such as (1) green lanes, (2) planted high tree lines along ditches, (3) low and high shrubs/copse, (4) hedges and (5) rows of fruit trees and willows which are slightly separated from each other.

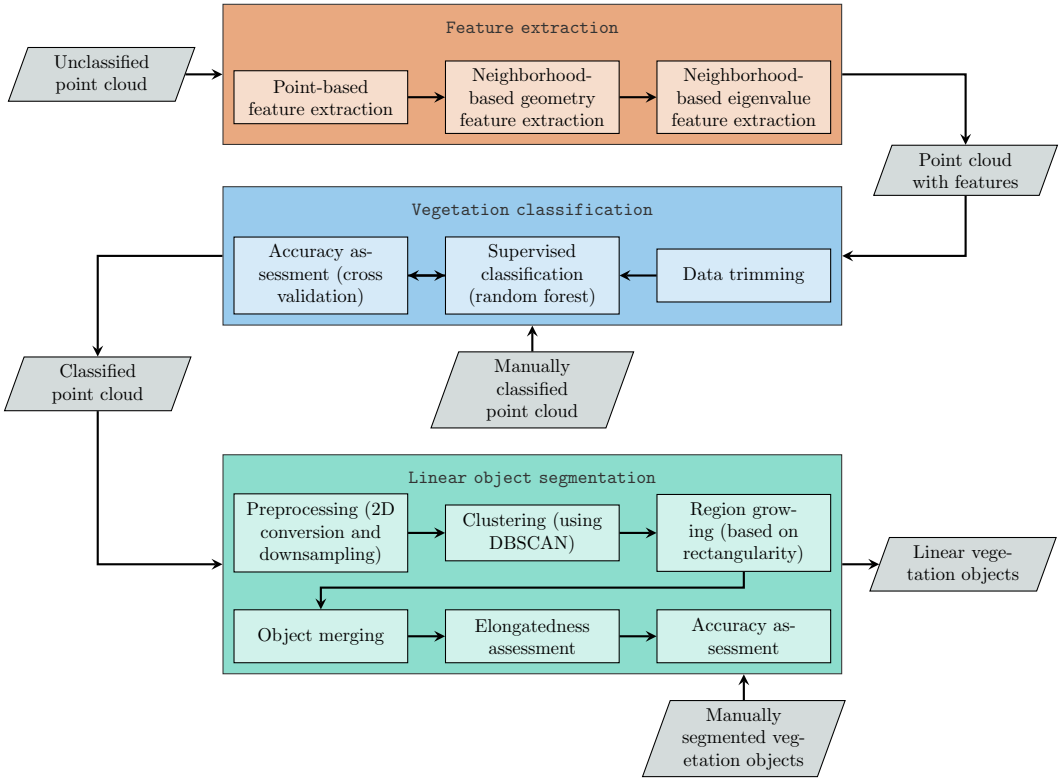

Figure 2: Workflow for feature extraction (orange), vegetation classification (blue), and linear objects segmentation (green). Computational steps are represented as rectangles and datasets as parallelograms.

3.1. Feature extraction

The relevance of the various input features has been extensively studied to separate urban from vegetation objects (Chehata et al., 2009; Guo et al., 2011; Mallet et al., 2011), but concise information for vegetation classification is scarce. After reviewing relevant literature, we selected fourteen features (table 1) which were later used for classification. These features are grouped in point-based and neighborhood-based features and reflect information from echo and local neighborhoods (geometric and eigenvalue based), respectively. The qualities of these features are considered to be efficient for discriminating vegetation objects from point clouds (Chehata et al., 2009).

3.1.1. Point-based features

The point-based features represent information from each single point (table 1). The point cloud $\mathcal{P}$ is a set of points $\{p_1, p_2, \ldots, p_n\} \in \mathbb{R}^3$, where each point p_i has x, y and z coordinates. In addition, an intensity value (I), a return number (R), and a number of returns (R_t) of the returned signal are stored. We used R_t as well as the normalized return number R_n as echo-based features (table 1). Since the available LiDAR data were lacking the information required to do a radiometric correction of the intensity data we omitted this feature for the classification (Kashani et al., 2015).

3.1.2. Neighborhood-based features

In addition to point-based features, we computed features based on the local neighborhoods of the points. We defined a neighborhood set $\mathcal{N}_i$ of points $\{q_1, q_2, \ldots, q_k\}$ for each point p_i, where $q_1 = p_i$, by using the k-nearest neighbors method with $k = 10$ points. In this way a k of 10 results in a neighborhood of ten points, one of which is the concerned point itself. Based on these neighborhoods we then computed four geometric features: height difference, height standard deviation, local radius and local point density (table 1).

In addition to these geometric features we further calculated eight eigenvalue-based features (table 1), which are used to describe the distribution of points of a neighborhood in space (Hoppe et al., 1992; Chehata et al., 2009). We used the local structure tensor to estimate the surface normal and to define surface variation (Pauly et al., 2002). The structure tensor describes the dominant directions of the neighborhood of a point by determining the covariance matrix of the x, y and z coordinates of the set of neighborhood points and computing the eigenvalues ($\lambda_1, \lambda_2, \lambda_3$, where $\lambda_1 > \lambda_2 > \lambda_3$) of this matrix and ranking them based on the eigenvalue values. Hence, the magnitude of the eigenvalues of this covariance matrix describe the spread of points in the direction of the eigenvector. The eigenvector belonging to the third eigenvalue is equal to the normal vector ($\vec{N} = (N_x, N_y, N_z)$) (Pauly et al., 2002). The points are linearly distributed if the eigenvalue of the first principle direction is significantly larger than the other two ($\lambda_1 \gg \lambda_2 \approx \lambda_3$), planarly distributed if the eigenvalues of the first two principle directions are about equal and significantly larger than the third ($\lambda_1 \approx \lambda_2 \gg \lambda_3$), and the points are scattered in all directions if all eigenvalues are about

Table 1: The features used for classification, split into two main groups: point-based and neighborhood-based. The point-based features are based on echo information and the neighborhood-based features are based on the local geometry and eigenvalue characteristics.

Feature group	Feature	Symbol	Formula	Reference
Point				
- Echo	Number of returns	R_t		
	Normalized return number	R_n	R/R_t	Guo et al. (2011)
Neighborhood				
- Geometric	Height difference	Δ_z	$\max_{j:\mathcal{N}_i}(q_{z_j}) - \min_{j:\mathcal{N}_i}(q_{z_j})$	Weinmann et al. (2015)
	Height standard deviation	σ_z	$\sqrt{\frac{1}{k}\sum_{j=1}^{k}(q_{z_j} - \overline{q_z})^2}$	Weinmann et al. (2015)
	Local radius	r_l	$\max_{j:\mathcal{N}_i}(\lvert p_i - q_j \rvert)$	Weinmann et al. (2015)
	Local point density	D	$k/(\frac{4}{3}\pi r_l^3)$	Weinmann et al. (2015)
- Eigenvalue	Normal vector Z	N_z		Pauly et al. (2002)
	Linearity	L_λ	$\frac{\lambda_1-\lambda_2}{\lambda_1}$	West et al. (2004)
	Planarity	P_λ	$\frac{\lambda_2-\lambda_3}{\lambda_1}$	West et al. (2004)
	Scatter	S_λ	$\frac{\lambda_3}{\lambda_1}$	West et al. (2004)
	Omnivariance	O_λ	$\sqrt[3]{\lambda_1\lambda_2\lambda_3}$	West et al. (2004)
	Eigenentropy	E_λ	$-\lambda_1\ln(\lambda_1) - \lambda_2\ln(\lambda_2) - \lambda_3\ln(\lambda_3)$	West et al. (2004)
	Sum of eigenvalues	$\sum_\lambda$	$\lambda_1 + \lambda_2 + \lambda_3$	Mallet et al. (2011)
	Curvature	C_λ	$\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}$	Pauly et al. (2002)

equal ($\lambda_1 \approx \lambda_2 \approx \lambda_3$). These properties (linearity, planarity and scatter), as well as some additional features (omnivariance, eigenentropy, sum of eigenvalues and curvature), are quantified using formulas (table 1).

3.2. Vegetation classification

The fourteen features (two echo, four geometric, and eight eigenvalue features) served as input for the vegetation classification. For this we first trimmed irrelevant points, then used a supervised classification to classify the remaining points, and finally assessed the accuracy.

3.2.1. Data trimming

To facilitate efficient processing, points that clearly did not belong to tall vegetation were removed from the dataset. This was done to remove non-vegetation and low-stature vegetation (e.g. grasses, herbs, agricultural fields) and to allow identification of linear vegetation elements with a certain height (i.e. composed of shrubs and trees). The points to be removed were characterized by a locally planar neighborhood and selected on the basis of the scatter feature (table 1). Points with a very low scatter value ($S_\lambda < 0.03$) were

removed, as this threshold was conservative and allowed to reduce thee data size, while still preserving all points characterizing tall vegetation.

3.2.2. Supervised classification

For the vegetation classification, we used a random forest classifier because it provides a good trade-off between classification accuracy and computational efficiency (Breiman, 2001; Weinmann et al., 2015). The random forest algorithm creates a collection of decision trees, where each tree is based on a random subset of the training data (Ho, 1998). Random forest parameters such as the maximum number of features, minimal samples per leaf, minimal samples per split and the ratio between minority and majority samples were optimized using a grid search. During the grid search a range of applicable values were chosen for each parameter and all combinations were tested and evaluated for performance using cross validation. To save time first a course grid (meaning the values of the parameters were in a larger range and more distance in between) was created to identify the region of best performance and subsequently a finer grid was made to find the best performing parameter set in that region (Hsu et al., 2003).

As the trimmed point cloud was imbalanced (i.e. it included a lot more *vegetation* than *other* points) and imbalanced training data can lead to undesirable classification results (He and Garcia, 2009), we used a balanced random forest algorithm. In this algorithm the subsets are created by taking a bootstrap sample from the minority class and a random sample from the majority class with a size based on the size of the minority class sample (Chen et al., 2004). By employing enough trees all majority class data are eventually used, while still maintaining a balance between the two classes. The decision trees were created using a Classification and Regression Tree (CART) algorithm (Breiman et al., 1984).

3.2.3. Accuracy assessment

For the accuracy assessment of the vegetation classification, a manual annotation of the trimmed point cloud into *vegetation* (e.g. trees and shrubs) and *other* (e.g. buildings, ditches, railroad infrastructure) classes was done using an interpretation of the point cloud and high resolution aerial photos. This resulted in a ground truth dataset of 101226 points of *vegetation* and 57752 points of *other*. To allow for a good assessment of the performance, while considering the imbalance in the dataset, we used the receiver operating characteristic (ROC) curve (Bradley, 1997), the Matthews correlation coefficient (MCC) (Matthews, 1975) and the geometric mean (Kubat et al., 1998) as accuracy metrics. These metrics evaluate the performance of a classifier well, even when dealing with an imbalanced dataset (Kohavi et al., 1995; Sun et al., 2009; López et al., 2013). To create a ROC curve, the true positive (TP) rate is plotted against the false positive (FP) rate at various decision thresholds. The area under a ROC curve (AUROCC) is a measure for the performance of the classifier (Bradley, 1997). The MCC analyzes the correlation between the observed and

the predicted data and is defined as:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1}$$

where TN are the true negatives and FN the false negatives. The geometric mean is defined as:

$$\text{Geometric mean} = \sqrt{\frac{TP}{TP + FP} \times \frac{TN}{TN + FN}} \tag{2}$$

The MCC, AUROCC and the geometric mean were obtained using a 10-fold cross validation. This is done by splitting the data into 10 randomly mutually exclusive subsets and using a subset as testing data and a classifier trained on the remaining data (Kohavi et al., 1995).

3.3. Linear object segmentation

For the third part of the workflow (linear object segmentation), we applied a preprocessing step, clustered the points, applied a region growing algorithm, merged nearby and aligned objects, evaluated their elongatedness, and finally assessed the accuracy (figure 2).

3.3.1. Preprocessing

Since we defined linearity as a purely two dimensional property, the point cloud was converted to 2D by removing the z-coordinate of the vegetation points. In addition, the data were spatially downsampled to 1 meter distance between vegetation points using Poisson sampling. Without losing too much precision (figure 3), this substantially decreased computation time.

3.3.2. Clustering

After reducing the amount of points, we clustered the remaining points together using a DBSCAN clustering algorithm (Ester et al., 1996). This algorithm is able to quickly cluster points together based on density and removes outlying points in the process. This decreased the processing time needed in the subsequent region growing step, since the amount of possible neighboring points is reduced.

3.3.3. Region growing

Region growing is an accepted way of decomposing point clouds (Rabbani et al., 2006; Vosselman, 2013) into homogeneous objects. Normally, regions are grown based on similarity of the attributes of the points. Here, regions were grown based on a rectangularity constraint. The rectangularity of an object is described as the ratio between the area of an object and the area of its minimum oriented bounding box (MOBB) (Rosin, 1999). The MOBB (figure 4b) is computed using rotating calipers (Toussaint, 1983). Here first a convex hull (figure 4a) is constructed using the QuickHull algorithm (Preparata and Shamos, 1985) and then the MOBB can be found by rotating the system by the angles the edges of the convex hull make with the x-axis and checking the bounding rectangles of each rotation, as the minimum oriented bounding box has a side collinear with one of the edges of the convex hull (Freeman and Shapira, 1975). The area of the

Figure 3: The vegetation points of a piece of tree line within the research area before (blue) and after (red) downsampling the point cloud, plotted on top of the high resolution orthophoto, in RD coordinates.

object can be calculated by computing the concave hull of the set of points belonging to the object (figure 4c). This hull is found by computing an alpha shape of the set of points (Edelsbrunner et al., 1983). This shape is created by computing a Delaunay triangulation of the points (Delaunay, 1934) and removing the triangles with a circumradius higher than $1/\alpha$, where α is a parameter which consequently influences the amount of triangles removed from the triangulation and thus the shape and area of the alpha shape. Higher alphas lead to more complex shapes, while lower ones to more smooth shapes.

For each cluster, points with the minimum x-coordinate and its ten closest neighbors were used as the starting region. Subsequently for each point the eight nearest neighbors are considered for growth (figure 5). Points were added as long as the region's rectangularity did not drop below a set threshold. An analysis of this threshold value on a subset of the data showed the best performance of the algorithm when this value was between 0.5 and 0.6, with marginal differences in performance in between these values, so we set it at 0.55. After a region is grown, the growing procedure was repeated for the next region until the entire cluster is segmented into rectangular regions.

3.3.4. Object merging

The resulting objects can be fragmented, for example, as the result of minor curves in the linear elements or small interruptions in vegetation. These objects were merged if they were in close proximity, faced a

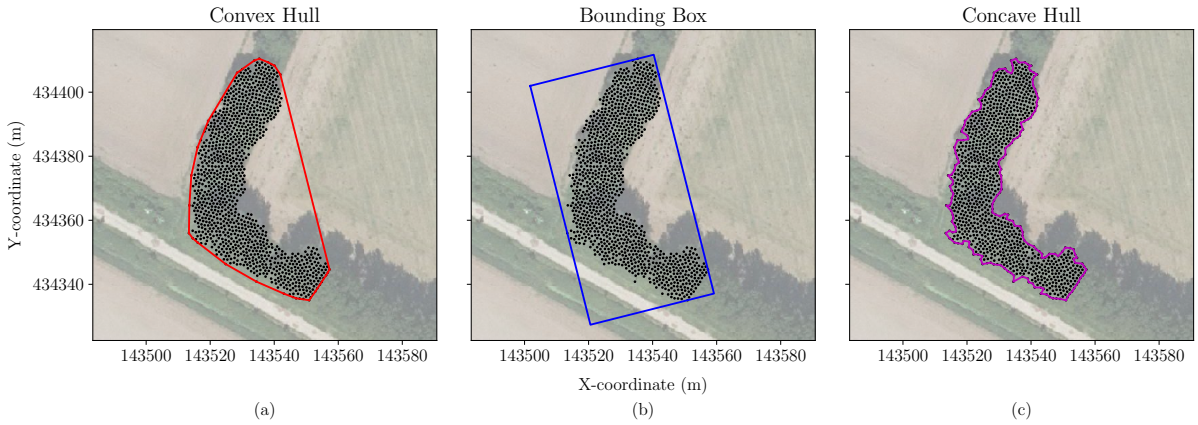

Figure 4: The downsampled vegetation points of a piece of tree line within the study area plotted on top of the high resolution orthophoto in Rijksdriehoek-coordinates, showing the different hulls used during the region growing algorithm: (a) the convex hull, (b) the minimal oriented bounding box, and (c) the concave hull. During the region growing the rectangularity is calculated by dividing the area of the concave hull (c) by the area of the bounding box (b).

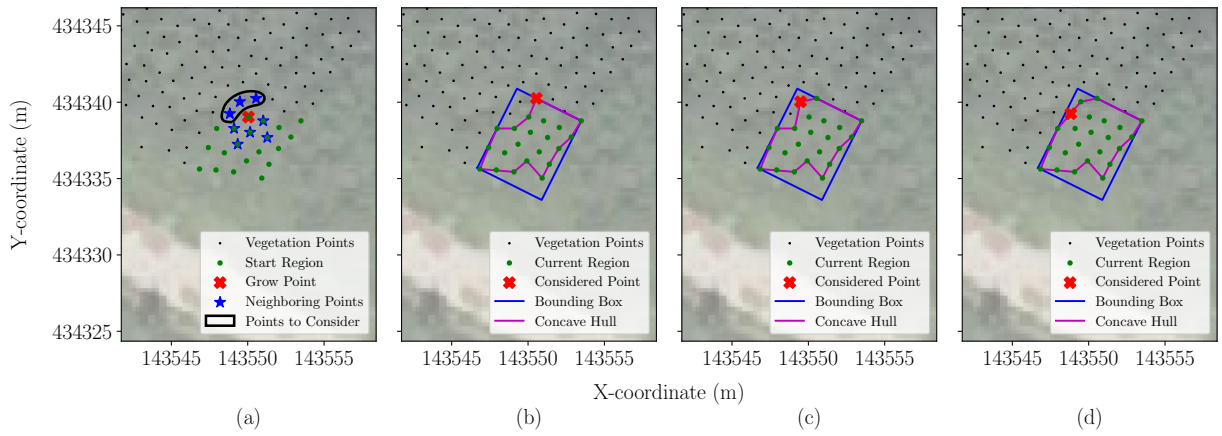

Figure 5: An example of the region growing process for one point. First the eight nearest neighbors are computed and the neighbors which are not already part of the region are considered to be added to the region (a). A bounding box and concave hull is computed for the region with the considered point added and the rectangularity is calculated (b). If this rectangularity is above a certain threshold the point is added to the region. Subsequently the process is repeated for the other potential points (c, d). When all potential points are checked, a next point of the region is check for nearest neighbors and the whole process is repeated until all points of the region, including the ones which are added during the growing process, have been checked.

similar compass direction, and were aligned. The compass direction was determined by computing the angle between one of the long sides of the minimum bounding box and the x-axis. The alignment was checked by comparing the angle of the line between the two center points with the directions of the objects. Once merged the lengths of the objects were added and the maximum of the widths taken as the new width.

3.3.5. Elongatedness

The merged objects were assessed for linearity by evaluating the elongatedness of an object, which is defined as the ratio between its length and width (Nagao and Matsuyama, 2013). The definition of a linear object is not clearly defined and consequently somewhat arbitrary. After analyzing the results using different values we set the minimum elongatedness at 1.5 and a maximum width of 60 meters, because these values made for a good extraction of linear elements, while excluding large forest-like vegetation.

3.3.6. Accuracy assessment

The accuracy of the delineated linear objects was assessed by calculating the user's, producer's and overall accuracy, as well as the harmonic mean of the precision and recall (F1), and MCC scores (Congalton and Green, 2008). We manually segmented the vegetation data into linear and nonlinear objects, after converting the classified vegetation points into polygons using an alpha shape. Consequently this assessment evaluates the accuracy of the segmentation given the accuracy of the vegetation points. By differencing of the automated and manually constructed data we created a map and confusion matrix detailing the accuracy.

4. Results

4.1. Vegetation classification

The vegetation classification resulted in a map showing three different classes (figure 6): points removed during preprocessing (e.g. low-stature vegetation, bare soil, water bodies), points classified as non-tall-vegetation structures (e.g. building edges, ditches and railroad infrastructure) and points belonging to tall vegetation. The accuracy assessment of this classification showed a producer's accuracy of 0.98 for *vegetation* and of 0.85 for *other*. The AUROCC of 0.98 showed that the *vegetation* and *other* class were well separated, and this was also supported by an MCC value of 0.76 (indicating of a positive correlation between the predicted and observed classes) and the geometric mean of 0.90.

4.2. Linear object segmentation

The vegetation was segmented using a region growing algorithm, the resulting objects were filtered for linearity and the results were compared with the manual segmentation (figure 7). This resulted in areas that were correctly classified as linear vegetation elements (true positives) and the regions that were accurately classified as nonlinear vegetation objects (true negatives). Some nonlinear areas were misclassified as linear (false positives), and some linear regions were classified as nonlinear (false negatives). However, the confusion matrix (table 3) showed an overall good accuracy of 0.90, an F1-score of 0.82, and an MCC of 0.76. Hence,

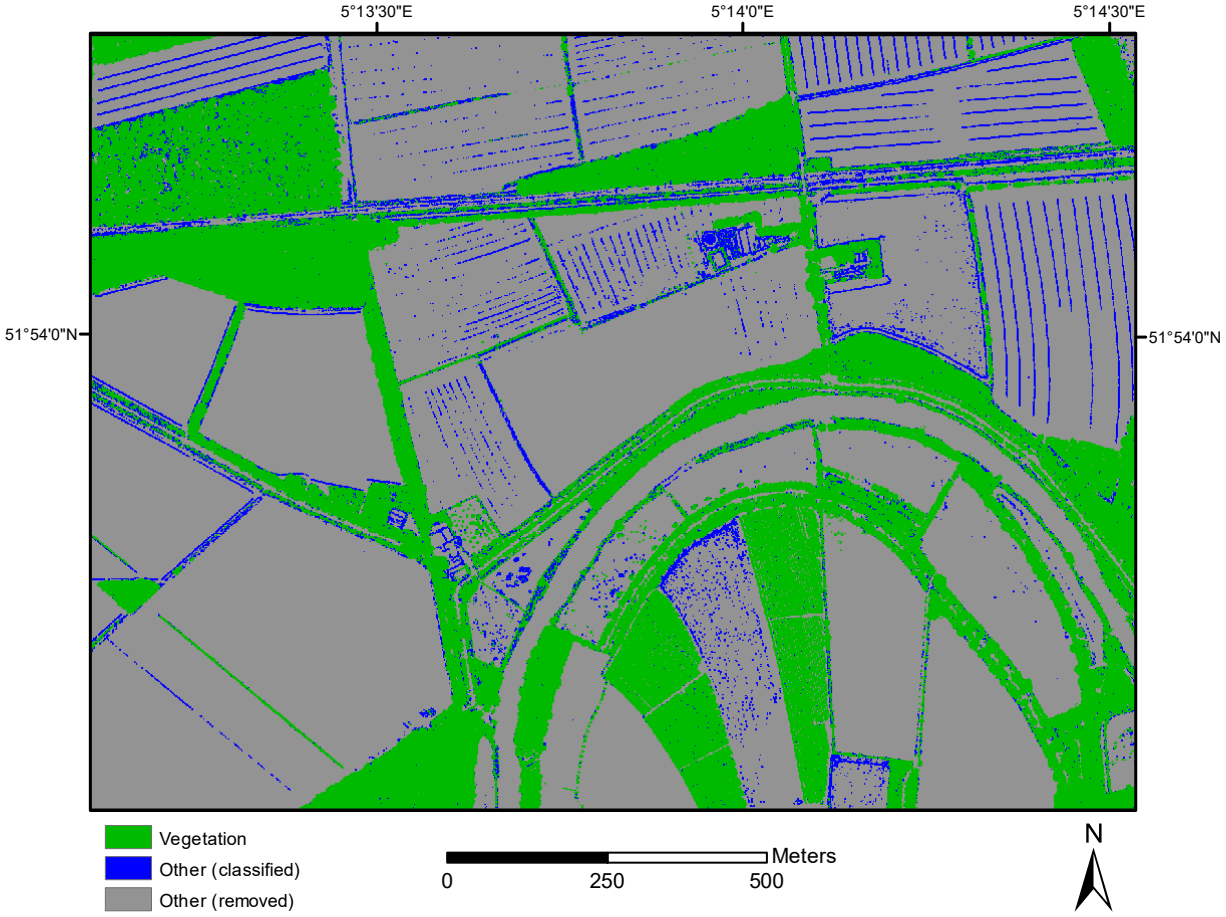

Figure 6: Results of the vegetation classification. Grey areas represent data points which were removed during the preprocessing (e.g. grasslands, agricultural fields, bare soil, water bodies). Blue represents points which were not removed during the preprocessing, but classified as *other* (mainly building edges, ditches and railroad infrastructure). Green represents points which were not removed during the preprocessing, but classified as tall vegetation.

Table 2: A confusion matrix showing the predicted classes against the actual classes of the points. These are accumulated over the 10 fold cross-validation.

		Predicted	
		Vegetation	Other
Actual	Vegetation	974177	22908
	Other	8171	47999

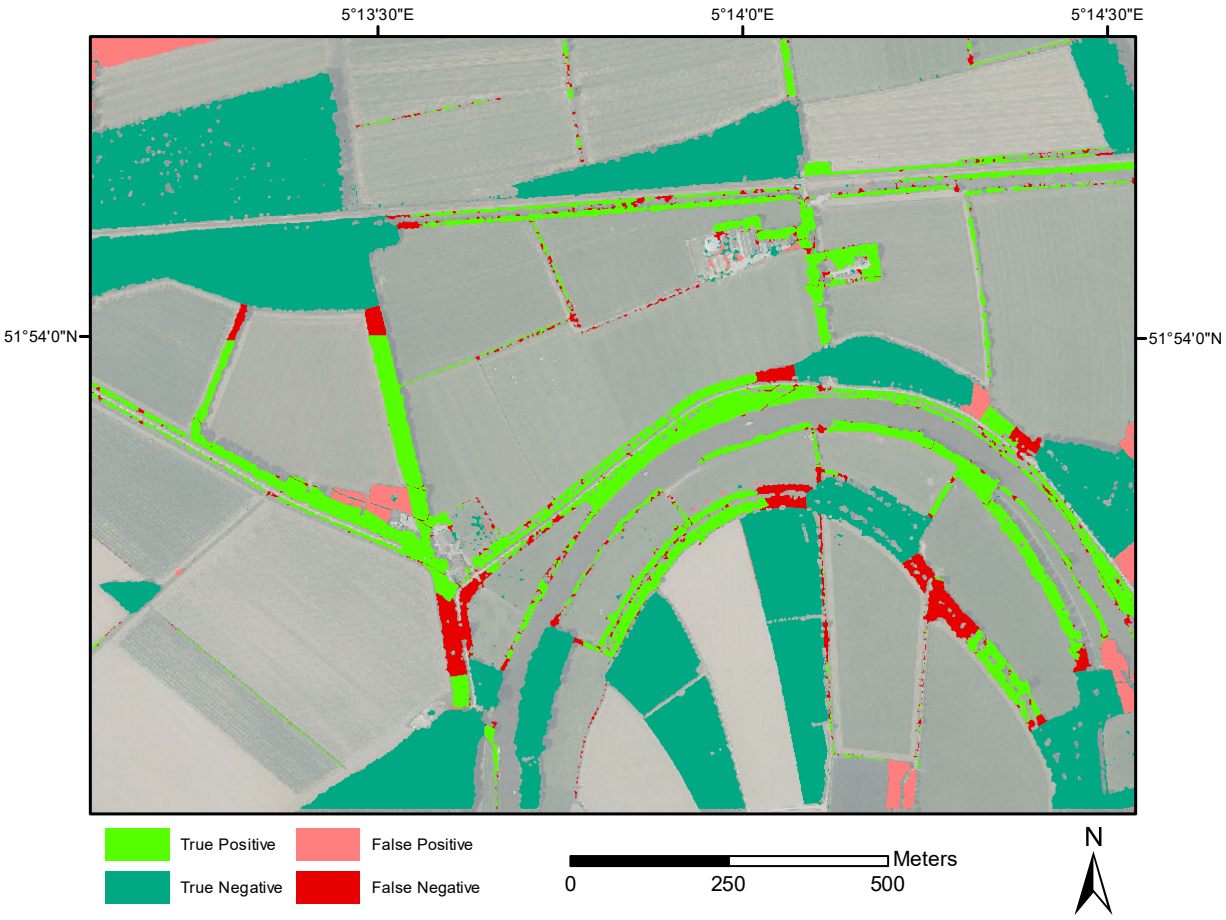

Figure 7: Results of the segmentation of linear elements. The correctly delineated areas are shown in green, where true positive (light green) means correctly classified as a linear element and true negative (dark green) as a nonlinear element. The misclassified areas are shown in red, where false positive (light red) means classified as a linear elements, while actually being a nonlinear element, and false negative (dark red) classified as a nonlinear object, while actually being a linear.

the majority of linear vegetation elements were successfully segmented, with user's and producer's accuracies of 0.85 and 0.80, respectively. Non-linear objects were successfully separated well, with user's and producer's accuracies of 0.92 and 0.94, respectively.

5. Discussion

5.1. Vegetation classification

Trimming the data based on the scatter feature proved an efficient step to reduce the computation time needed to classify the vegetation. A substantial part of the point cloud (about 22%) corresponding mainly to smooth and planar areas such as bare soil, grassland and water bodies, was removed. This preprocessing step made the dataset imbalanced, but proper steps could be taken to handle the problems when classifying such datasets. When analyzing the classification statistics it is important to take this

Table 3: Confusion matrix listing the automatically segmented against the manually annotated set of linear and non-linear vegetation objects in area (m^2).

		Predicted	
		Linear	Nonlinear
Actual	Linear	116483.76	28385.56
	Nonlinear	20201.53	336754.65

filtering step into consideration. The removed points are the ones easy to classify as not belonging to tall vegetation. Consequently, the remaining points that do not belong to tall vegetation (*other*) share many similarities with the tall vegetation points and are therefore harder to classify correctly. The filtered points are not part of the accuracy assessment and therefore these statistics might give a distorted, more negative picture of the accuracy. If the filtered points would be part of the assessment it would show even higher accuracy values, but it would give less information on how the classification actually went.

Nevertheless, the majority of points were correctly classified (table 2). Most of the incorrectly classified points are dispersed and surrounded by correctly classified points. These scattered points did not have a major impact on the subsequent segmentation of linear objects.

A limitation in our accuracy assessment is the lack of a validation dataset that was taken from a completely different area. Such a validation dataset is an effective measure against overfitting. To prevent overfitting we instead used a cross validation which, while considered an effective approach, is often less preferred than a validation dataset. (need ref)

5.2. Linear object segmentation

The comparison of the manual with automated delineation shows that linear objects were accurately extracted (figure 7), which is supported by the accuracy scores. The largest errors are found at places where a nonlinear element transitions into a linear one. This is a consequence of the rectangularity sometimes only falling below the threshold after a while of growing in a wrong direction.

(needs more work)

It is important to keep in mind that the definition of a linear element is not clearly defined and is therefore somewhat arbitrary. The accuracy assessment was based on our interpretation of the linear elements in the area.

(needs more work)

6. Concluding remarks

At present, LiDAR datasets still differ in quality, content and accessibility across and within countries. Therefore, object identification methods developed should overcome these inconsistencies. The quality of the AHN3 dataset of the Netherlands is sufficient to correctly identify linear vegetation objects with our method. In addition, multi-temporal LiDAR datasets can effectively be analyzed for change in the spatial distribution of linear vegetation objects using such a generic classifier. Initiatives to upscale the classification of linear vegetation objects, reed beds and selected forest metrics to national and European scale, based on classification of LiDAR point clouds, and using efficient cloud computing facilities are being made (Kissling et al., 2017).

The ecological value of providing such large a dataset of linear vegetation objects lies in the broad extent and fine-scale locational details, which is a powerful quality that can be used in the (3D) characterization of ecosystem structure. Existing ecosystem and biodiversity assessment projects, such as the MAES (Mapping and Assessment of Ecosystems and their Services) project (Maes et al., 2013), the SEBI (Streamlining European Biodiversity Indicators) project (Biała et al., 2012), and the high nature value farmland assessment (Paracchini et al., 2008) on a European scale and assessments of Planbureau voor de Leefomgeving (PBL) on a national level (Bouwma et al., 2014), could profit from the new details.

Acknowledgments

This work is part of the eEcoLiDAR project, eScience infrastructure for Ecological applications of LiDAR point clouds (Kissling et al., 2017), funded by Netherlands eScience Center (`https://www.esciencecenter.nl`).

References

Aguirre-Gutiérrez, J., Kissling, W. D., Carvalheiro, L. G., WallisDeVries, M. F., Franzén, M., Biesmeijer, J. C., 2016. Functional traits help to explain half-century long shifts in pollinator distributions. Scientific reports 6, doi: 10.1038/srep24451.

AHN, 2016. Inwinjaren AHN2 & AHN3. `http://www.ahn.nl/common-nlm/inwinjaren-ahn2--ahn3.html` [Online: accessed April 2017].

Aksoy, S., Akçay, H. G., Wassenaar, T., 2010. Automatic mapping of linear woody vegetation features in agricultural landscapes using very high resolution imagery. IEEE Transactions on Geoscience and Remote Sensing 48 (1), 511–522, doi: 10.1109/TGRS.2009.2027702.

Bailly, J., Lagacherie, P., Millier, C., Puech, C., Kosuth, P., 2008. Agrarian landscapes linear features detection from lidar: application to artificial drainage networks. International Journal of Remote Sensing 29 (12), 3489–3508.

Biała, K., Condé, S., Delbaere, B., Jones-Walters, L., Torre-Marín, A., 2012. Streamlining european biodiversity indicators 2020. Tech. Rep. 11/2012, European Environment Agency.

Boutin, C., Jobin, B., Bélanger, L., Baril, A., Freemark, K., 2001. Hedgerows in the farming landscapes of canada. Hedgerows of the World: their ecological functions in different landscapes, 33–42.

Bouwma, I., Sanders, M., op Akkerhuis, G. J., Onno Knol, J. V., de Wit, B., Wiertz, J., van Hinsber, A., 2014. Biodiversiteit bekeken: hoe evalueert en verkent het PBL het natuurbeleid? Tech. Rep. 924, Planbureau voor de Leefomgeving.

Bradley, A. P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition 30 (7), 1145–1159.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A., 1984. Classification and regression trees. CRC press.

Burel, F., 1996. Hedgerows and their role in agricultural landscapes. Critical reviews in plant sciences 15 (2), 169–190.

CGAL Project, 2018. CGAL User and Reference Manual, 4th Edition. CGAL Editorial Board.

Chehata, N., Guo, L., Mallet, C., 2009. Airborne lidar feature selection for urban classification using random forests. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 38 (3), 207–2012.

Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn imbalanced data. Tech. Rep. 666, Department of Statistics, UC Berkeley.

CloudCompare, 2018. CloudCompare (version 2.10alpha) [GPL software]. `http://www.cloudcompare.org/` [Online: accessed June 2018].

Congalton, R. G., Green, K., 2008. Assessing the accuracy of remotely sensed data: principles and practices. CRC press.

Croxton, P., Hann, J., Greatorex-Davies, J., Sparks, T., 2005. Linear hotspots? the floral and butterfly diversity of green lanes. Biological conservation 121 (4), 579–584, doi: 10.1016/j.biocon.2004.06.008.

Delaunay, B., 1934. Sur la sphere vide. Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk 7 (793-800), 1–2.

Edelsbrunner, H., Kirkpatrick, D., Seidel, R., 1983. On the shape of a set of points in the plane. IEEE Transactions on information theory 29 (4), 551–559.

Eitel, J. U., Höfle, B., Vierling, L. A., Abellán, A., Asner, G. P., Deems, J. S., Glennie, C. L., Joerg, P. C., LeWinter, A. L., Magney, T. S., et al., 2016. Beyond 3-d: The new spectrum of lidar applications for earth and ecological sciences. Remote Sensing of Environment 186, 372–392.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. Vol. 96. pp. 226–231.

Freeman, H., Shapira, R., 1975. Determining the minimum-area encasing rectangle for an arbitrary closed curve. Communications of the ACM 18 (7), 409–413.

Gobster, P. H., Nassauer, J. I., Daniel, T. C., Fry, G., 2007. The shared landscape: what does aesthetics have to do with ecology? Landscape ecology 22 (7), 959–972, doi: 10.1007/s10980-007-9110-x.

Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. ISPRS Journal of Photogrammetry and Remote Sensing 66 (1), 56–66.

He, H., Garcia, E. A., 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21 (9), 1263–1284, doi: 10.1109/TKDE.2008.239.

Ho, T. K., 1998. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence 20 (8), 832–844.

Hoppe, H., DeRose, T., Duchampt, T., McDonald, J., Stuetzle, W., 1992. Surface reconstruction from unorganized points. Computer Graphics 26, 2.

Hsu, C., Chang, C., Lin, C., 2003. A practical guide to support vector classification.

Jones, E., Oliphant, T., Peterson, P., et al., 2001. SciPy: Open source scientific tools for Python. `http://www.scipy.org/` [Online: accessed April 2017].

Jongman, R., 2004. Landscape linkages and biodiversity in european landscapes. The new dimensions of the European landscape. Springer, Dordrecht, 179–189.

Kashani, A. G., Olsen, M. J., Parrish, C. E., Wilson, N., 2015. A review of lidar radiometric processing: From ad hoc intensity correction to rigorous radiometric calibration. Sensors 15 (11), 28099–28128.

Kissling, W. D., Seijmonsbergen, A., Foppen, R., Bouten, W., 2017. eecolidar, escience infrastructure for ecological applications of lidar point clouds: reconstructing the 3d ecosystem structure for animals at regional to continental scales. Research Ideas and Outcomes 3, e14939, doi: 10.3897/rio.3.e14939.

Kohavi, R., et al., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. Vol. 14. Stanford, CA, pp. 1137–1145.

Kubat, M., Holte, R. C., Matwin, S., 1998. Machine learning for the detection of oil spills in satellite radar images. Machine learning 30 (2-3), 195–215.

Lefsky, M. A., Cohen, W. B., Parker, G. G., Harding, D. J., 2002. Lidar remote sensing for ecosystem studies: Lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies, can accurately estimate vegetation structural attributes and should be of particular interest to forest, landscape, and global ecologists. AIBS Bulletin 52 (1), 19–30.

Lim, K., Treitz, P., Wulder, M., St-Onge, B., Flood, M., 2003. Lidar remote sensing of forest structure. Progress in physical geography 27 (1), 88–106.

López, V., Fernández, A., García, S., Palade, V., Herrera, F., 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250, 113–141.

Maes, J., Teller, A., Erhard, M., Liquete, C., Braat, L., Berry, P., Egoh, B., Puydarrieux, P., Fiorina, C., Santos, F., et al., 2013. Mapping and assessment of ecosystems and their services. Tech. Rep. EUR 27143 EN, Joint Research Center - Institute for Environment and Sustainability.

Mallet, C., Bretar, F., Roux, M., Soergel, U., Heipke, C., 2011. Relevance assessment of full-waveform lidar data for urban area classification. ISPRS Journal of Photogrammetry and Remote Sensing 66 (6), S71–S84, doi: 10.1016/j.isprsjprs.2011.09.008.

Marquer, L., Gaillard, M.-J., Sugita, S., Poska, A., Trondman, A.-K., Mazier, F., Nielsen, A. B., Fyfe, R. M., Jönsson, A. M., Smith, B., et al., 2017. Quantifying the effects of land use and climate on holocene vegetation in europe. Quaternary Science Reviews 171, 20–37.

Matthews, B. W., 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405 (2), 442–451.

McKinney, W., et al., 2010. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. Vol. 445. Austin, TX, pp. 51–56.

Meyer, B. C., Wolf, T., Grabaum, R., 2012. A multifunctional assessment method for compromise optimisation of linear

landscape elements. Ecological Indicators 22, 53–63.
Nagao, M., Matsuyama, T., 2013. A structural analysis of complex aerial photographs. Springer Science & Business Media.
Paracchini, M. L., Petersen, J.-E., Hoogeveen, Y., Bamps, C., Burfield, I., van Swaay, C., 2008. High nature value farmland in europe. Tech. Rep. EUR 23480 EN, Joint Research Center - Institute for Environment and Sustainability.
Pauly, M., Gross, M., Kobbelt, L. P., 2002. Efficient simplification of point-sampled surfaces. In: Proceedings of the conference on Visualization'02. IEEE Computer Society, pp. 163–170.
PDAL contributors, 2018. PDAL: The Point Data Abstraction Library. `https://pdal.io/`.
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, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.
Preparata, F. P., Shamos, M., 1985. Computational geometry: an introduction. Springer Science & Business Media.
Quackenbush, L. J., 2004. A review of techniques for extracting linear features from imagery. Photogrammetric Engineering & Remote Sensing 70 (12), 1383–1392.
Rabbani, T., Van Den Heuvel, F., Vosselmann, G., 2006. Segmentation of point clouds using smoothness constraint. International archives of photogrammetry, remote sensing and spatial information sciences 36 (5), 248–253.
Rosin, P. L., 1999. Measuring rectangularity. Machine Vision and Applications 11 (4), 191–196.
Song, J.-H., Han, S.-H., Yu, K., Kim, Y.-I., 2002. Assessing the possibility of land-cover classification using lidar intensity data. International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences 34 (3/B), 259–262.
Spellerberg, I. F., Sawyer, J. W., 1999. An introduction to applied biogeography. Cambridge University Press, doi: 10.1086/393452.
Stoate, C., Boatman, N., Borralho, R., Carvalho, C. R., De Snoo, G., Eden, P., 2001. Ecological impacts of arable intensification in europe. Journal of environmental management 63 (4), 337–365.
Sun, Y., Wong, A. K., Kamel, M. S., 2009. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 23 (04), 687–719.
Tansey, K., Chambers, I., Anstee, A., Denniss, A., Lamb, A., 2009. Object-oriented classification of very high resolution airborne imagery for the extraction of hedgerows and field margin cover in agricultural areas. Applied geography 29 (2), 145–157, doi: 10.1016/j.apgeog.2008.08.004.
Thornton, M. W., Atkinson, P. M., Holland, D., 2006. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. International Journal of Remote Sensing 27 (3), 473–491, doi: 10.1080/01431160500207088.
Toussaint, G. T., 1983. Solving geometric problems with the rotating calipers. In: Proc. IEEE Melecon. Vol. 83. p. A10.
Turner, M. G., 1989. Landscape ecology: the effect of pattern on process. Annual review of ecology and systematics, 171–197.
Van der Zanden, E. H., Verburg, P. H., Mücher, C. A., 2013. Modelling the spatial distribution of linear landscape elements in europe. Ecological indicators 27, 125–136.
Vannier, C., Hubert-Moy, L., 2014. Multiscale comparison of remote-sensing data for linear woody vegetation mapping. International journal of remote sensing 35 (21), 7376–7399, doi: 10.1080/01431161.2014.968683.
Vosselman, G., 2013. Point cloud segmentation for urban scene classification. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 1 (2), 257–262.
Walt, S. v. d., Colbert, S. C., Varoquaux, G., 2011. The numpy array: a structure for efficient numerical computation. Computing in Science & Engineering 13 (2), 22–30.
Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. ISPRS Journal of Photogrammetry and Remote Sensing 105, 286–304.
West, K. F., Webb, B. N., Lersch, J. R., Pothier, S., Triscari, J. M., Iverson, A. E., 2004. Context-driven automated target detection in 3d data. In: Defense and Security. International Society for Optics and Photonics, pp. 133–143.
Yan, W. Y., Shaker, A., El-Ashmawy, N., 2015. Urban land cover classification using airborne lidar data: A review. Remote Sensing of Environment 158, 295–310.