Assessment of Sitka spruce natural regeneration and recruitment in Scotland through photogrammetry

Auteur : Lahaise, Yann
Promoteur(s) : Lejeune, Philippe
Faculté : Gembloux Agro-Bio Tech (GxABT)
Diplôme : Master en bioingénieur : gestion des forêts et des espaces naturels, à finalité spécialisée
Année académique : 2020-2021
URI/URL : http://hdl.handle.net/2268.2/13238

LIÈGE université
Gembloux
Agro-Bio Tech

Assessment of Sitka spruce natural regeneration and recruitment in Scotland through photogrammetry

LAHAISE YANN

TRAVAIL DE FIN D'ETUDES PRESENTE EN VUE DE L'OBTENTION DU DIPLOME DE MASTER BIOINGENIEUR EN GESTION DES FORETS ET DES ESPACES NATURELS

ANNEE ACADEMIQUE 2020-2021

(CO)-PROMOTEUR(S) : Prof. LEJEUNE Philippe

LIÈGE université
Gembloux
Agro-Bio Tech

Assessment of Sitka spruce natural regeneration and recruitment in Scotland through photogrammetry

LAHAISE YANN

TRAVAIL DE FIN D'ETUDES PRESENTE EN VUE DE L'OBTENTION DU DIPLOME DE MASTER BIOINGENIEUR EN GESTION DES FORETS ET DES ESPACES NATURELS

ANNEE ACADEMIQUE 2020 -2021

(CO)-PROMOTEUR(S) : Prof. LEJEUNE Philippe

Acknowledgements

The completion of this work could not have been completed without the help of many people to whom I wish to address my thanks.

Firstly, I would like to acknowledge Forest research for sharing the images that made possible the writing of this dissertation. A special thanks to Rubén Manso for his quality support throughout this work despite the distance, his excellent advice, but also the wise and deep review of this document.

I also thank my academic supervisor Pr. Phillipe Lejeune for his help and advice.

Thanks to Pr. Gauthier Ligot for the advice for the writing of this document and all the suggestion given through the last months.

I would like to thank my family for the support and the encouraging during all these years and for all their lovely attentions. To my father, thank you for the feedback.

To all my friend, thanks for the support and especially to Marie-Pierre for taking the time to read this document and give me her feedback.

Finally, a special thank to Sophie for her encouraging, thank you for your patience and unfailing support during all these years.

Abstract

In Scotland, Sitka spruce represents the major timber resource. The regeneration of this species is abundant and occurs widely after clearfelling. Finding way to carry out affordable inventories of regeneration would help forest managers to make decision on the future of the regeneration resource. The deployment of Unmanned aerial vehicle (UAV) allows rapid assessment of forest and regeneration and it is likely to lead of a decreasing cost for field surveys.

This current study aims to map Sitka spruce seedlings with orthophotos acquired by drones. The used method, based on RGB images, has been carried out using OTB (Orfeo ToolBox) and a object based image analysis approach. It consists in four steps: the segmentation (i), a supervised classification (random forest) (ii), the mapping of the studied sites (iii) and finally, the validation (iv) of the model through points acquired by photointerpretation. The result shows that one of the classification models reaches a global accuracy of 66.9 with pseudo independent dataset and 77.4 with an independent dataset. The results are expected to be better with images acquired during other periods (leaf-off period) in order to prevent confusion with surrounding vegetation presents on the studied sites. Despite this fact, the mapping of Sitka spruce seems promising with an RGB camera and may offer a promising potential for commercial forestry. In addition, the method may be applied in other context such as ecological restoration or forest health.

Key words : Sitka spruce; seedlings; regeneration; UAV; orthoimages; RGB; segmentation; supervised classification; random forest; Orfeo ToolBox; Scotland.

Résumé

En Écosse, l'épicéa de Sitka représente la principale ressource en bois. La régénération de cette espèce est abondante, particulièrement après une coupe à blanc. Trouver un moyen de réaliser des inventaires abordables de celle-ci aiderait les gestionnaires forestiers à prendre des décisions sur l'avenir de la ressource. L'utilisation de drones permet une évaluation rapide de la forêt et de la régénération, permettant ainsi une réduction des coûts des inventaires de terrain.

La présente étude vise à cartographier les semis d'épicéa de Sitka à partir d'orthophotos acquises par drones. La méthode utilisée, basée sur des images RGB, a été réalisée en utilisant OTB (Orfeo ToolBox) et une approche d'analyse d'image basée sur des objets. Elle consiste en quatre étapes: la segmentation (i), une classification supervisée (random forest) (ii), la cartographie des sites étudiés (iii) et enfin, la validation (iv) du modèle par des points acquis par photointerprétation. Les résultats montrent qu'un des modèles de classification atteint une précision globale de 66,9 avec un jeu de données pseudo-indépendant et de 77,4 avec un jeu de données indépendant. On s'attend à ce que les résultats soient meilleurs avec des images acquises à d'autres périodes de l'années (période hors feuille) afin d'éviter la confusion avec la végétation environnante présentes sur les sites d'étude. Malgré cela, la cartographie de l'épicéa de Sitka semble prometteuse avec une caméra RGB et offrirait donc un potentiel intéressant pour la foresterie commerciale. En outre, la méthode pourrait être utilisée dans d'autres contextes, tel que celui de la restauration écologique ou la santé des forêts.

Mots clefs : épicéa de Sitka; semis; régénération; drones ; orthoimages; RGB ; segmentation; classification supervisée ; random forest; Orfeo ToolBox; Ecosse

Contents

List of Figures

List of Tables

Abbreviations

DEM	Digital Elevation Model
EPSG	European Petroleum Survey Group
GRVI	Green Red Vegetation Index
LiDAR	Light Detection and Ranging
NDVI	Normalized Difference Vegetation Index
UAV	Unnamed Aerial Vehicle
RGB	Red, Green, Blue
OBIA	Object-Based Image Analysis
OTB	ORFEO TollBox
PA	Producer Accuracy
Qgis	Quantum Gis
RAM	Random Access Memory
RF	Random Forest
UA	User Accuracy
WGS	World Geodetic System

1 Introduction

Sitka spruce (*Picea sitchensis* (Bong.) Carr.) is currently the most spread of all non native species introduced in Europe (Øyen and Nygaard, 2020). The British Isles – Great Britain and Ireland – is the European region with more Sitka spruce forested area but the species is also found in France, Norway, Denmark, Germany, Sweden and Iceland. In total, Sitka Spruce covers about 1.3 million hectares in Europe (Nygaard and Øyen, 2017) of which about 507 000 ha are in Scotland. Sitka spruce represents the main timber resource in Scotland, amounting to 58 % of the total conifer timber in 2014 (Forestry Commission, 2020). Spontaneous natural regeneration of the species occurs widely and abundantly in Scotland shortly after clearfelling. In spite of this, restocking is typically carried out through replanting in most Sitka spruce forests (Forest Enterprise Scotland, 2017), as this allows to straightforwardly achieve the optimal spacing that maximises production and timber properties. This said, Sitka spruce natural regeneration is rarely assessed to evaluate whether or not there is a case for keeping and managing this regeneration as an alternative to replanting. Unfortunately, classic regeneration inventories are not affordable in operational forestry. Finding ways to carry out affordable surveys (e.g. Manso and McLean, 2020) would help forest managers to make decisions on the future of the regeneration resource, given the fact that an excess of seedling or an uneven distribution of the regeneration implies either additional of costs in respacing or then facing changes in wood properties (Price, 2016).

Unmanned aerial vehicles (UAVs) or drones allows rapid evaluation of forest and vegetation structure. The deployment of drones is likely to lead to a reduction of costs in fields surveys. UAVs equipped with remote sensors (e.g. Red, Green, Blue (RGB) camera, Lidar device, infrared camera, etc.) can acquire very high resolution imagery of small elements with high operational flexibility(Feduck et al., 2018; Fromm et al., 2019; Mesas-Carrascosa et al., 2014). The feasibility of using UAVs to identify seedlings of conifers has been tested in previous research with a high detection rate (Feduck et al., 2018; Goodbody et al., 2017). These studies successfully used vegetation indices based on visible reflectance, which only requires a simple spectral RGB camera – as they can be derived from true color images(Jannoura et al., 2014) –, instead of near infrared indices such as Normalize Difference Vegetation Index (NDVI). Among the many indices based on the RGB bands, the Green Red Vegetation Index (GRVI) proved to be the most promising (Feduck et al., 2018; Motohka et al., 2010).

The use of drones outside the military domain is more recent than vegetation assessment by satellite imagery or aerial photography. Generally speaking, UAVs can be used in civil applications such as scientific studies, public safety or commercial tasks (Mesas-Carrascosa et al., 2014). The development of civilian drones used for aerial photography offers new opportunities in terms of forest characterisation (Lisein, 2016). Some examples of aeroplane-mounted photography do exist where regeneration higher than 0.30 cm could be detected (Kirby, 1980), but the focus has turned to UAV-borne cameras for the reasons mentioned above.

Due to the high resolution of the images acquired through UAV flights, current research on spectral image processing often involves segmentation, which is part of Object Based Image Analysis (OBIA), as a preliminary step for automatic or semi-automatic classification of the features of interest (Feduck et al., 2018). The resulting segments are essentially pixels grouped into vector objects representing land-based features. Since segmentation groups pixels with common spectral values, the objects to be classified must have different spectral signatures (Kressler et al., 2005). In a second step, the classification itself is carried out on these objects through a classification model using shape, size, spatial and spectral properties according to the objective of the classification. This approach is in opposition to pixel based classification where all pixels share the same size and there is not consideration of the neighboring pixels (GISgeography, 2021). A classification model in OBIA uses segments as the inputs and the output is the classification of those segments to represent land cover. There exist a number of classifiers, with those based on machine learning being commonly used (Kotsiantis, 2007). One specific class remote sensing imagery has to deal with is shades as they occlude the features of interest and have specific spectral properties, causing a decrease in accuracy and detection(li et al., 2005; Shahtahmassebi et al., 2013). In the case of this study, the features to detect and classify in the orthoimages are therefore Sitka spruce seedlings, shades and any other features.

The main research question of this dissertation is to fit classification models to detect natural regeneration of Sitka spruce with orthoimages acquired through UAVs in Scotland. These models would allow mapping Sitka spruce seedlings from orthoimages, which would provide forest managers with a tool for decision making in regard to natural regeneration. The technical developments are based on RGB images to avoid complex photogrammetric processing and to contribute to make the regeneration survey as affordable as possible.

2 State of the art

2.1 Sitka spruce

Sitka spruce is native of the Pacific Northwest of the USA and Canada. The distribution area of Sitka spruce in its native range is determined by full water availability during the growing season, the optimal situation being when summer precipitation is elevated and there is no marked dry period during the summer. The species was first introduced in Great Britain in 1831. It was however not until the afforestation programme that followed the First World War that the Forestry Commission, founded in 1919 for these purposes, decided to extensively plant fast growth exotic conifer species to rapidly restock British forests. Although Norway spruce was already present in Great Britain, Sitka spruce was preferred because it grows faster and it could be planted on a wider ranger of sites (Moore, 2011). The mean yield of Sitka spruce reaches 14 m^3/ha/year in Great Britain. The area of Sitka spruce forest increased from 67.000 ha in 1947 to 692.000 ha in 2007. Most of this area is located in Scotland and represents 47 % of the total forest land in the country (figure 1).

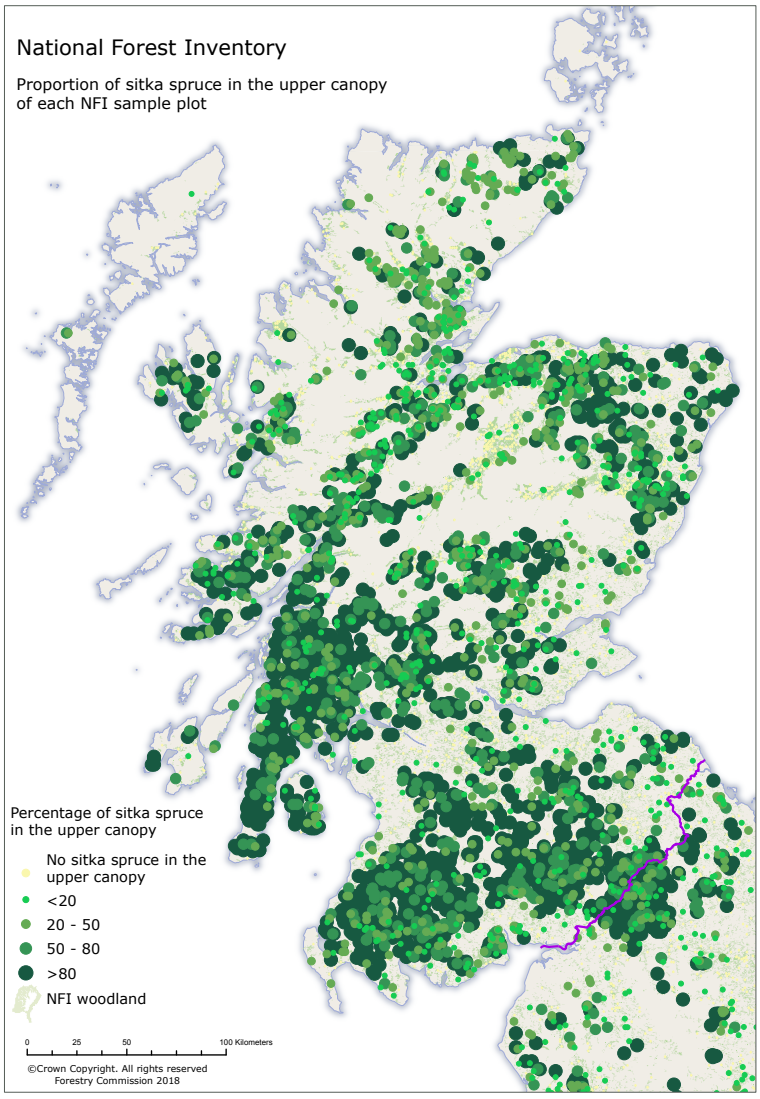

Figure 1: Percentage of Sitka spruce in the upper canopy of each National Forest Inventory sample plot in Scotland. Source: Forestry Commission

Sitka spruce trees reach 70 m in height and two meters in diameter with a life span of 500 years in their native range. In Great Britain, however, rotations usually range from 35 to 45 years for commercial stands. For an average site quality, trees reach a top height of between 16 and 23 meters at the end of the typical rotation. The diameter at breast height (1.3 m) at this time varies from 25 to 40 cm, depending on initial spacing, whether thinnings were conducted or not, and site quality.

Seedling density in naturally regenerated stands can reach 100 000 stems ha^{-1}, which could make natural regeneration a potential and cheaper alternative to replanting. The downside is that such high seedling densities need to be managed through respacing and that spatial distribution may not be perfectly uniform. Insufficient or absent respacing leads to a strong and continuous competition between the young trees, which may be detrimental from the point of view of timber assortments and timber properties (Price, 2016). As a result, managers have traditionally resorted to replanting to optimize timber quality, growth and wind stability (Forest Enterprise Scotland, 2017). The standard spacing in Scottish planted forests is 2 m $\times$ 2 m to maximise those parameters. All this been said, a comprehensive assessment of costs and outputs from respacing versus restocking has never been conducted. If the case for restocking is to be reevaluated, the first step would be to reach a better understanding on the spatial patterns of natural regeneration through an affordable survey.

2.2 Orthophotograph

An orthophoto, also known as orthophotograph, is an aerial photograph that has been orthorectified. This means that the distortion due to scale, tilt and relief present in normal aerial photographs have been corrected. An orthophoto combines the advantages of a map and those of a photograph. Contrary to a map, the terrain is represented in actual detail and not just by lines and symbols. Like in a map, however, true distance, angle and areas can be measured directly on the orthophotos. The primary inputs required for production of orthophotos is a digital elevation model (DEM) and photographs (Lillesand et al., 2015). The images can be acquired from satellite, aircraft or UAV.

Six independent parameters are needed to carried out aerial photographs for the purposes of photogrammetric mapping. Those parameters are related to position and angular orientation of each photograph. The position and angle are relative to the origin and orientation of the ground system used for mapping. Three parameters provide 3D positioning (x, y, z) information of the center of the photocoordinates axis system. The three other are the 3D rotations angle (ω,ϕ,κ) tied to the amount and direction of tilt. These rotation depends on the orientation of the platform and camera when the photos are shot. The 6 parameters are taken at the time of each photograph is taken (Lillesand et al., 2015).

The orthophotos are usually composed by several bands. For example, orthophotos with true colors are composed by RGB bands. The images are characterized by a spatial resolution depending on the pixel size. An higher resolution means smaller pixels size. The orthophotos are typically georeferenced to an Earth coordinate system, which allows locating each pixel with accuracy.

2.3 Reflectance

Sun's electromagnetic radiation is reflected and absorbed in different ways by different objects. The reflectivity properties are a function of the object's material but also of its physical and chemical state. These properties are also modified by the angle of incidence of the sun and the roughness of the surface of the material. Reflectance is not consistent across all wavelengths of the electromagnetic spectrum.

There are many definitions of reflectance. For the purposes of this dissertation, we define it *sensu* (Schaepman-Strub et al., 2006) as the percentage of radiant flux reflected over the the radiant flux incident (1).

$$Reflectance = \frac{d\phi_r}{d\phi_i} \quad (1)$$

where:

- $d\phi_r$ is derivative of the radiant flux reflected
- $d\phi_i$ is derivative of the radiant flux incident

Materials reflect in different part of the spectrum, which allows discriminating different surface features according to the spectral reflectance signature. The next paragraph will focus on the signature of vegetation and bare soil, and the properties that ultimately affect reflectivity.

In general, healthy vegetation very efficiently absorbs electromagnetic energy in the visible region. Our eyes discern vegetation as green because the pigment present in plants (the chlorophyle) strongly absorbs light at wavelengths around 0.45 (blue) and 0.67 µm (red), and strongly reflects at 0.5 µm (green) (figure 2). Although the frequencies in the near infrared are not considered in this dissertation, it should be mentioned, for completeness, that reflectance sharply increases between 0.7 and 1.3 µm (Lillesand et al., 2015) due to the internal structure of the leaves. As this structure differs between species, near infrared wavelengths can, for example, be used to discriminate plant species.

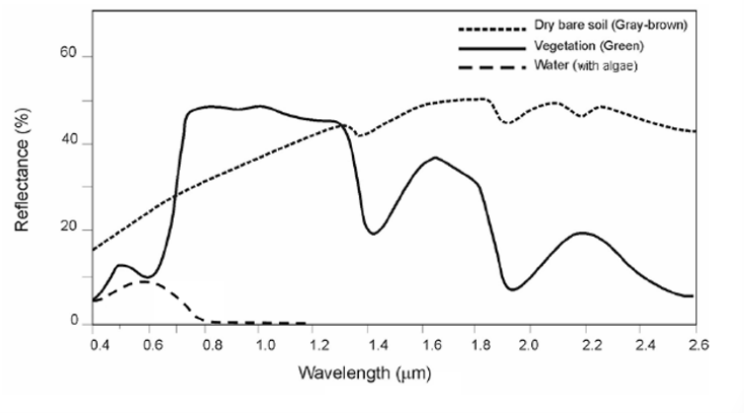

Figure 2: Spectral reflectance curves for three material. Source: Lillesand et al.

Bare soil generally presents high reflectance values, even greater in the near-infrared and shortwave infrared. Soil reflectance is modified by:

- Moisture content
- Soil texture (rate of sand, silt, and clay)
- Surface roughness
- Presence of iron oxide
- biological matter content

2.4 Segmentation

The segmentation process is the first step in the OBIA approach and a previous step to the classifier algorithm that will lead to a classification of good quality on images with high resolution. The high complexity of those images (i.e. acquired by drones) hinders the automated or semi-automated analysis, like OBIA or classification of pixels, and therefore requires new methods. Pixel based classifiers do not consider neighbouring pixels, which leads to classification errors by not including the spatial context of the target pixel. A solution for this issue is to separate images into smaller segments that group neighbour pixels with homogeneous spectral information (i.e. segmentation) (Willhauck, 2000). Moreover, once the segments are created, the classification computing time decreases compared to pixel based classification methods. In addition, the segmentation allows to reduce misclassification of pixels at the edge between different classes and to prevent the so called speckle effect (Kressler et al., 2005). The speckle effect, also knows as "salt and pepper" effect, is due to local spatial heterogeneity between bordering pixels. Despite the similarity, as a result, close neighbors pixels are classified into different classes despite their similarity (Kelly et al., 2011).

The segments are grouped on the basis of different parameters which may vary according to the dataset and the field of application of the classification. These parameter could be shape, size, spectral value, colour, texture, compactness, *etc.* An accurate segmentation is necessary to achieve a correct classification on the basis that mistakes cannot be corrected at a later stage (Kressler et al., 2005).

2.5 Supervised classification

The purpose of supervised classification in machine learning is to build a concise model with the predictor features of a training dataset. The class labels of this dataset are known and are distributed by the model according to the predictor features. The training is used a reference by the algorithm to assign class labels to another dataset where the predictor features are known and the value of the class labels are unknown. The supervised classification is then validated with data containing the same predictor features and whose labels are also known. This allows to compare the prediction (i.e. class label) of the model to a correct classification. Throughout the supervised classification process the predictor features should remain the same. There exist several classification algorithms (Kotsiantis, 2007). The unsupervised classification is not developed in this document but the algorithm used in this type of classification is similar to the one applied here.

2.5.1 Random forest

Random forest (RF) is an algorithm used to build a classification model, for example in supervised classification, and probably one of the most widely used due to its multiple advantages (Havryliuk et al., 2018; Pelletier et al., 2016). An RF is a collection of tree classifiers (Breiman, 2001). Each tree is built with an input sample subsets selected by bootstrapping, a selection method where any given sample has the same probability to be selected and replaced. Contrary to a classical decision tree, a subset of the input variables is randomly chosen at each "node" (minimal decision unit within the tree). The number of variables selected is mainly the square root of the total number of inputs variable (Pelletier et al., 2016). Each tree in the RF "votes" for a class (e.g. assigns a class to a given segment). After generating of a large number of trees, most popular class ("majority of vote") is assigned to the segment (Breiman, 2001).

2.6 Indices

Several indices can be used to detect vegetation or other features such as shades in classification. This section will focus on indices based on visible wavelength such as the RGB bands applied in this study. These indices differ according to the context of application (Feduck et al., 2018; Goodbody et al., 2017; Motohka et al., 2010)

2.6.1 Green red vegetation index

The Green Red Vegetation Index (GRVI) can be used to discriminate vegetation from bare soil. An advantage of this index is that only red and green band are needed(2). Recall that an RGB camera is all it is needed, compared to Normalized Difference Vegetation Index (NDVI), where an infrared camera is required (Feduck et al., 2018; Motohka et al., 2010).

$$GRVI = \frac{\rho_{green} - \rho_{red}}{\rho_{green} + \rho_{red}} \quad (2)$$

where :

- ρ is the reflectance and it ranges from 0 to 255 for each bands

The GRVI ranges from -1 to 1 and value comprises between 0 and 1 represents healthy vegetation. In the context of this study, formula was modified to obtain:

$$GRVI_m = \frac{\rho_{green} - \rho_{red}}{\rho_{green} + \rho_{red}} * 1000 + 1000$$

This modification allows extending the range of this index from 0 to 2000, which in turns allows for a higher ranging than the original formula. As a result, healthy vegetation ranges between 1000 and 2000 in the $GRVI_m$.

2.6.2 Shades and brightness

A shadow materialises when an item totally or partially obstructs light directly from the light source. There exist two types of shadows with different brightness: cast and self shadow. When a shadow is projected by the object in the direction of the light source, it is called a cast shadow. In contrast, self shadows are the part of an object not lit by direct light. The cast shadow itself is divided into two parts: umbra and penumbra (figure 3) depending on whether the light is totally or moderately blocked by its object (Arevalo et al., 2006). The radiating or the reflecting light of an object is an aspect of visual perception called brightness. The values of brightness are commonly lower for cast shadow than self shadow because the former gathers more electromagnetic radiation from neighboring illuminated objects than the latter.

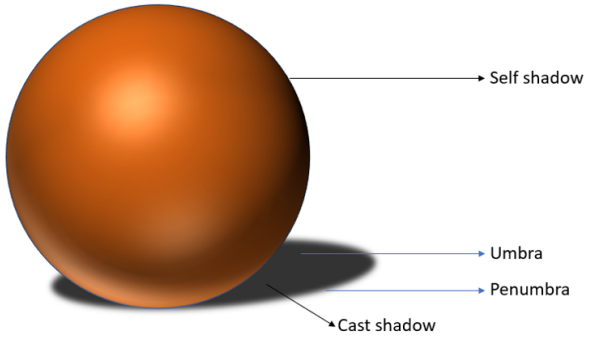

Figure 3: Type of shadow, adapted from Arevalo et al.

As previously mentioned, shades are one of the main issues in remote sensing imagery, being the most frequent type of error observed in remote sensing data. If shade is not properly dealt with, it may lead to incorrect classification in land cover assessments and, in consequence, to a considerable reduction in the accuracy of data extraction and change detection. Moreover, season, time of acquisition and latitude have all an impact on shades due to the variation of light incidence (Shahtahmassebi et al., 2013).

In an RGB color space, brightness can be captured as an index based on the reflectance of the three bands and ranges from 0 to 255.

$$Brightness = \frac{G + R + B}{3} \quad (3)$$

Given the close relationship between brightness and shade, this index (3) can be used to improve the segmentation process and highlight shadows (Thanh Ngo, 2015).

3 Material

3.1 Study area

For the purposes of this study, the three sites that had originally been selected by Forest Research in a preliminary study on this topic (Manso and McLean, 2020) were used (figure 4). These sites consist of former Sitka spruce planted stands where clearcutting had been applied in recent years and natural regeneration had subsequently established.

The sites were located at the forest of Ae, in Dumfries and Galloway, Southwest Scotland. The specific coordinates (reference system Wgs 84) were 55.22, -3.51 (site 1), 55.12, -3.63 (site 2) and 55.31, -3.51 (site 3). Each site consists of different cohorts of seedlings (table 2). A fourth site was surveyed at the beginning of July 2021 in Stirlingshire (figure 4), near Aberfoyle (56.18, -4.56). The seedlings in this site were between 5 and 6 years old.

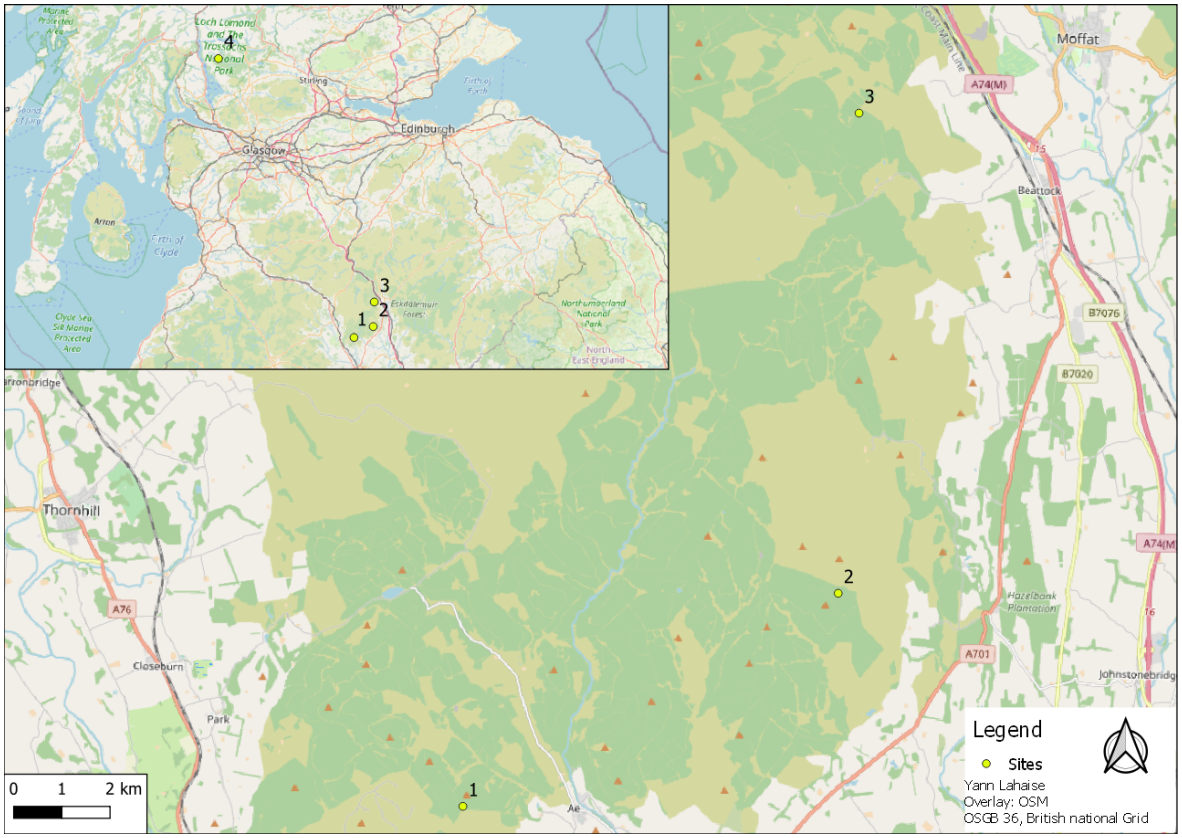

Figure 4: Localisation of the studied sites.

3.2 Equipment

A contractor was in charge of flight and data collection. The flights were conducted under suitable conditions (no wind, sunny and dry). The device used was a DJI inspire 2 and the details about this UAV are in the table 1. No information about the camera used to acquire the images was recorded by the contractor.

According to Manso and McLean, the horizontal datum used throughout the data gathering phase of the survey was OSGB36 (OSTN15). Data has been rendered in OSGB36 Datum, British National Grid. The vertical datum for all data is Ordnance Datum. OSTN15 defines OSGB36 National Grid in conjunction with the National GPS Network.

The other only information on data collection available was the approximate dates when the photos were acquired by the drones (mid August 2019, sites 1, 2 and 3; and beginning of July, site 4).

Table 1: UAV specifications

UAV specifications: DJI inspire 2 (t650)	
Description	Quadcopter
Battery	4280 mAh
Vehicle Weight with Battery	3440g
Platform Estimated Flight Time	27 min
Maximum Speed	94 kph(Sport mode)

3.3 Data

Based on the photos acquired by the drones, Forest Research had produced orthophotos for each site (figure 5,6,7), 8). These are composed of the three RGB bands. The height of the seedlings varied from 0.5 m to 3 m. The resolution of the initial orthoimages ranged from 0.004 m to 0.013 m (table 2).

Table 2: Resolution of the orthoimages, ages of natural regeneration and surface for the sites studied

	Resolution (m)	Ages of natural regeneration (years)	Surface (ha)
Site 1	0.004	4	1.65
Site 2	0.011	1-2	4.53
Site 3	0.006	6	2.65
Site 4	0.013	5-6	2.15

The 4 orthoimages exhibited very different characteristics, with variations in the presence and colour of grass and other non relevant vegetation such as bushes, the size of the Sitka spruce crowns, as well as the presence of remnants from clearcutting. The age of the young seedlings varies among the different images (table 2).

Site 1 (Figure 5) shows a large amount of remnants and easily detectable regeneration. Grass are present in the plot and it is difficult to discern this vegetation type from regeneration in some cases (figure 5).

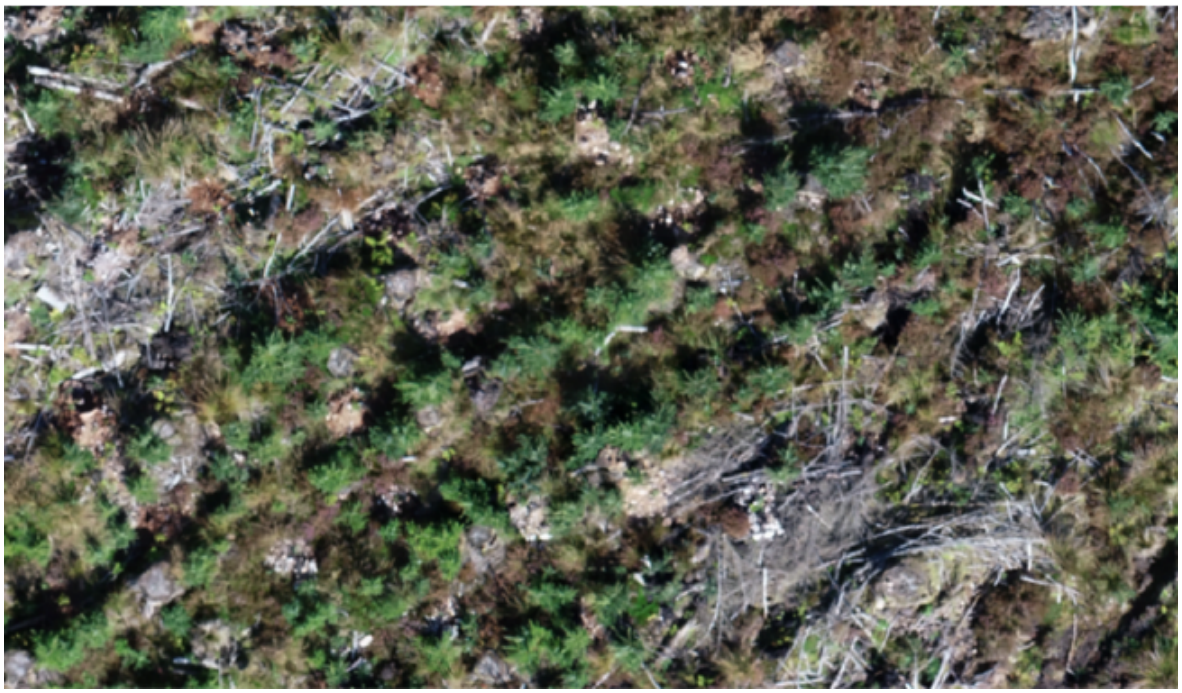

Figure 5: Orthophoto of the site 1. Remnants and regeneration are easily detectable.

Site 2 also has a significant amount of slash and most stumps from the previous cut are visible on the image, resulting in at least an equivalent number of small shadows (figure 6). The regeneration, although younger than in site 1 or 3, can still be distinguished, albeit less clearly than on the other two sites.

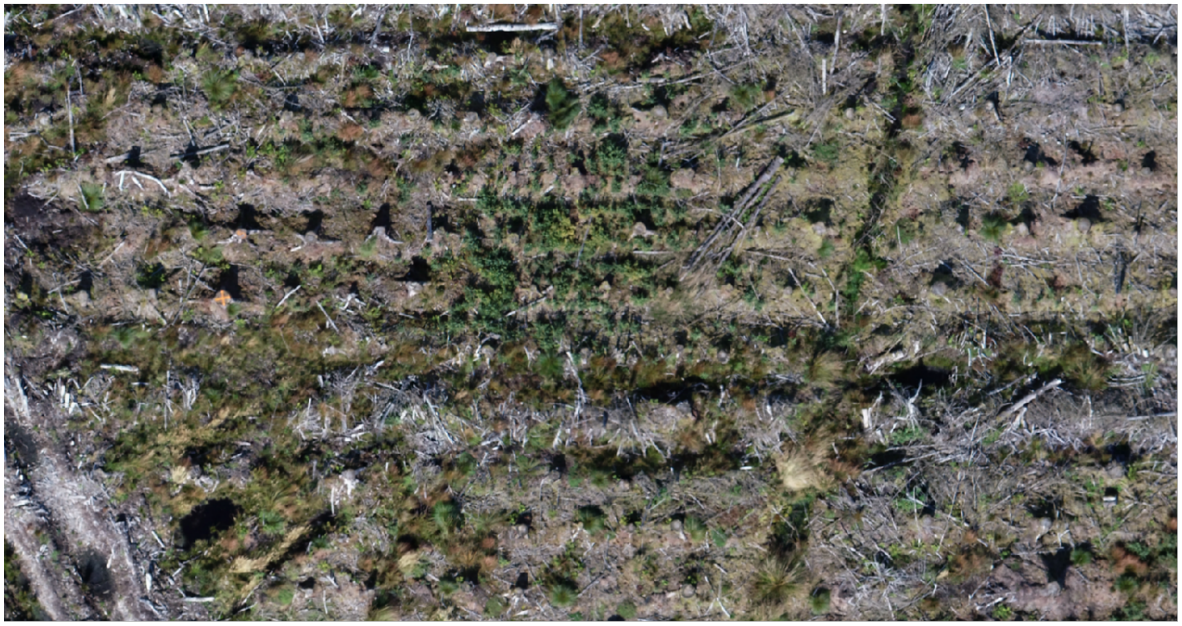

Figure 6: Orthophoto of the site 2. The regeneration is less distinguished due to a smaller size.

In contrast, site 3 shows little evidence of remnant from previous harvesting operations and grass and bushes are easily discernible from the Sitka spruce regeneration, given the predominantly brown colour of the former, which makes manual classification easier. Regeneration is very obvious although the seedlings cast are large (figure 7).

Figure 7: Orthophoto of the site 3. The regeneration is abundant and shades present large area.

Site 4 presents characteristics that are comparable to those of site 3 in terms of vegetation and remnants from clearfelling, although there seems to be a lower proportion of shades (figure 8).

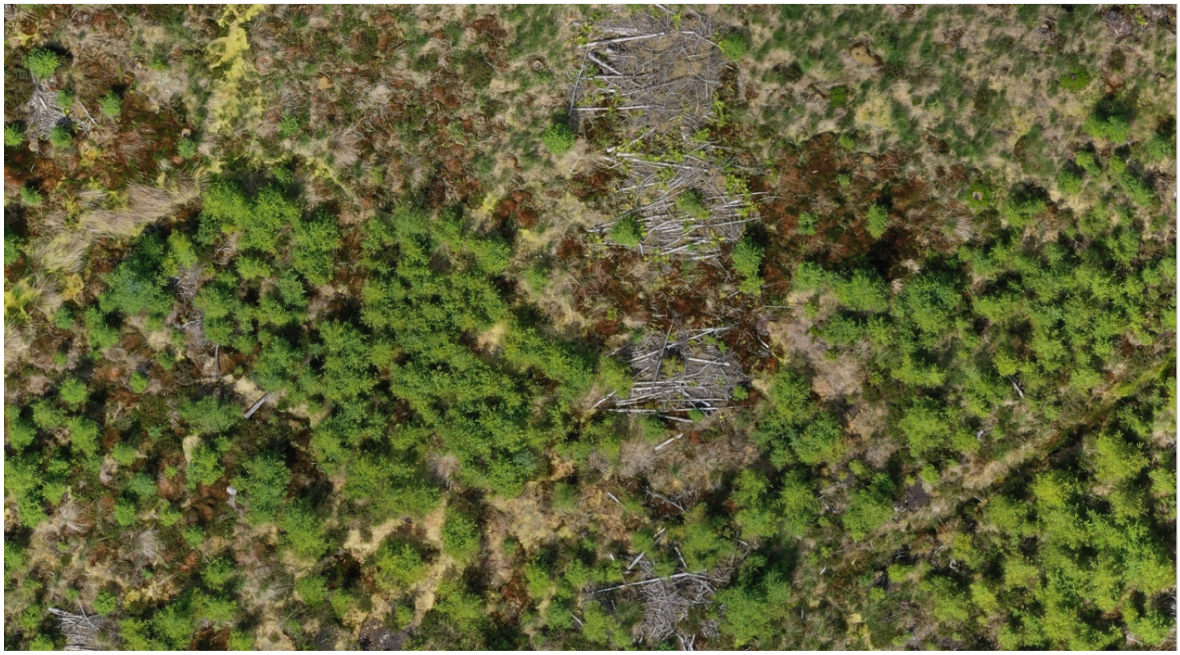

Figure 8: Orthophoto of the site 4. The characteristics of vegetation are similar to the site 3 except concerning the proportion of shades.

3.4 Software

All analysis and image manipulation was computed in R (R Core Team, 2019) with Orfeo ToolBox (OTB), an open-source project for processing remotely sensed imagery (Grizonnet et al., 2017) and in

Qgis software (QGIS Development Team, 2021). The reference system used was the OSGB36 Datum (British National Grid), which corresponds to the EPSG: 27700 in R and Qgis.

4 Methods

A method was established for the purposes of this dissertation that consists of four main steps (figure 9). The fist step deals with the preparation of the data, where the images were first segmented into groups of pixels according their spectral value. Based on this segmentation, a sample of polygons were manually assigned by photo interpretation into three classes of interest (i.e. Sitka spruce seedlings, shades and all other features). The second step consists in using these segments or polygons to train a RF algorithm with the aim of building classification models. In the third step, these classification models were applied over all the segments of the different images to produce predictions of the three classes, which resulted in classification maps. Finally, the different classifications were evaluated with a sample of independent segments to compute confusion matrices and the Cohen's Kappa index (global performance) (Cohen, 1960).

The code developed to carry out these analyses is available in the appendix.

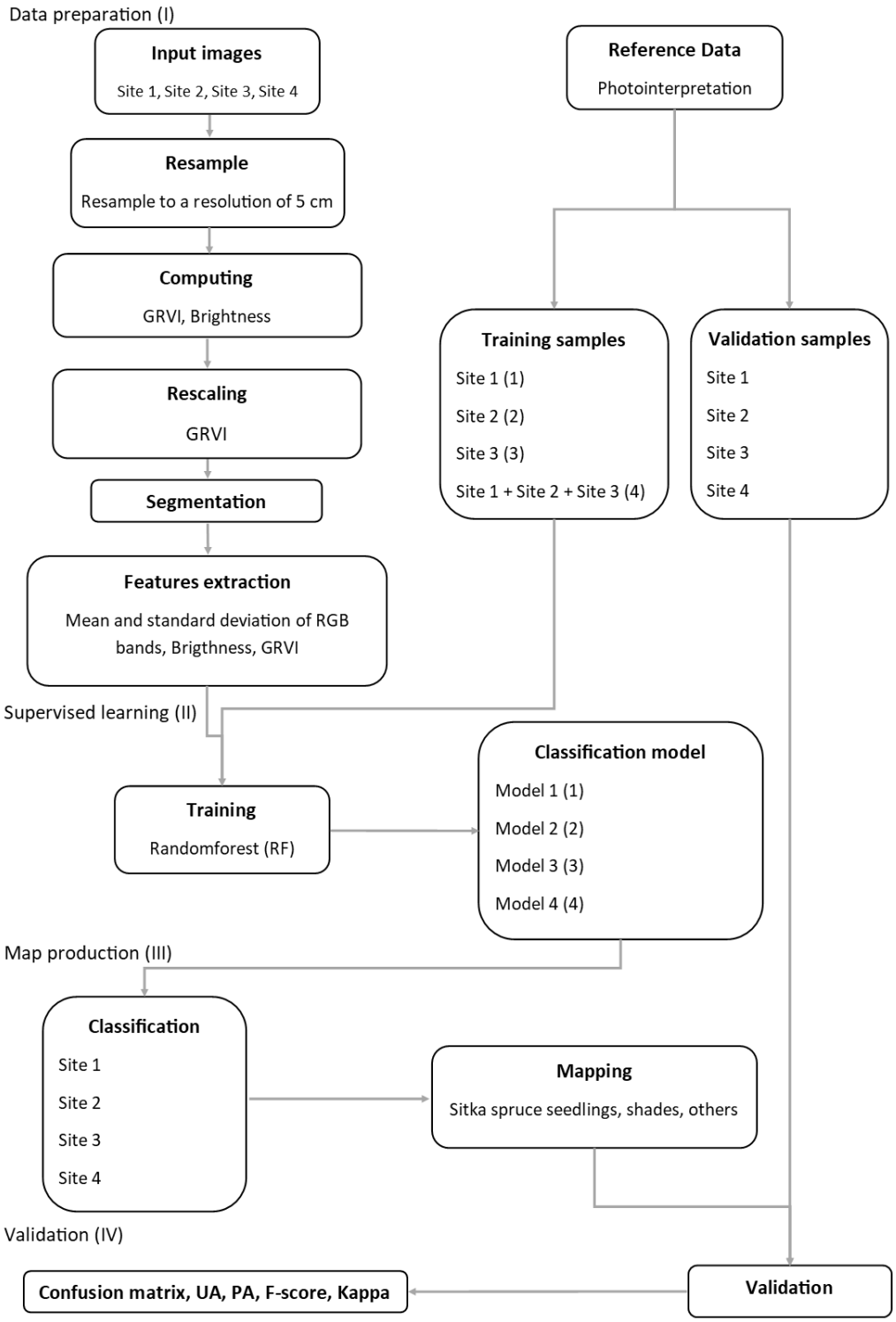

Figure 9: Flowchart of the main steps developed in the method.

4.1 Segmentation

The orthoimages were segmented using the spectral value of the following five bands as a variable :

- Red band
- Green band
- Blue band
- Brightness
- GRVI

The RGB bands and brightness ranged from 0 to 255. The GRVI was rescaled from 0 to 512 (table 3), which results in a greater weight of the GRVI value in the segmentation process to ensure that Sitka spruce seedlings are correctly detected. This is because the weight of each variable used in the segmentation process is proportional to its range of variation. It is therefore recommended to scale the range of values of the different bands in order to control the weight given to the different variables.

Table 3: Range of spectral band used for segmentation

Spectral band	Initial range	Segmentation range
Red	0-255	0-255
Green	0-255	0-255
Blue	0-255	0-255
Brightness	0-255	0-255
GRVI	Site 1 : 0-1714	0-512
	Site 2 : 0-1666	0-512
	Site 3 : 0-1777	0-512
	Site 4 : 0-1428	0-512

The segmentation was carried out with the algorithm meanshift of OTB (OTB, 2021). Three parameters are necessary to compute the segmentation, which are spatial radius, range radius and the minimum size. These parameters need to be tuned to avoid oversegmenting or undersegmenting the scene. This was done visually, as any statistical assessment of this part would beyond the scope of this dissertation. The first parameter corresponds to spatial radius of the neighbourhood and the second one defines the radius (expressed in radiometry unit) in the multispectral space. Those were fixed to 10 and 30, respectively, after performing several preliminary tests. The third parameter is the minimum number of pixels that defines the minimum size of segments and it was set to 32 (i.e. 0.08 m^2), also after some initial tests. These tests allowed visually evaluating the parameters chosen for the segmentation 10.

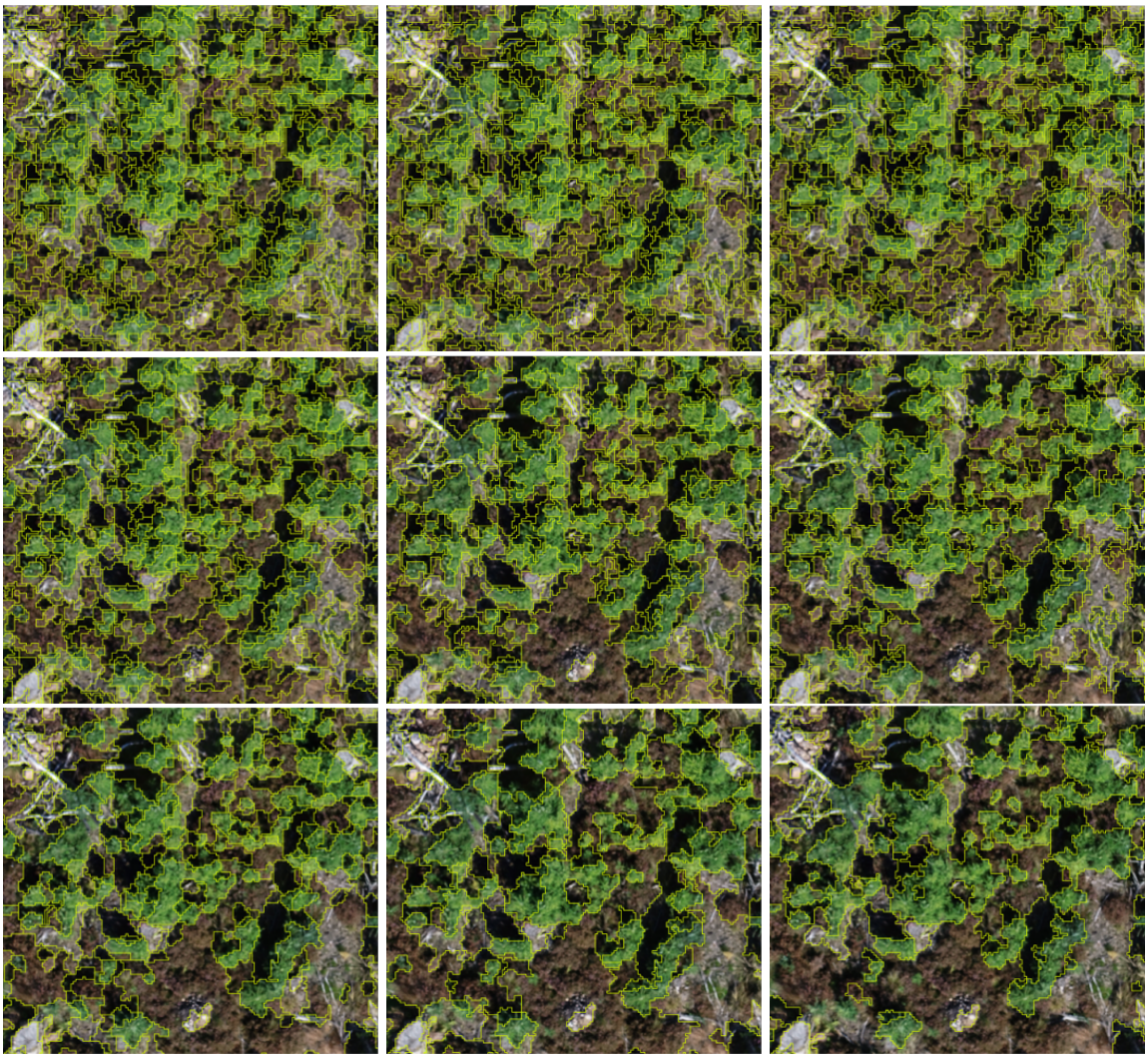

Figure 10: Spatial radius and range radius parameters variation for segmentation. From right to left spatial radius was set to 5,10,20 and from top to bottom range radius was set to 15,30,50. The central image shows the result of the parameters chosen.

To reduce computing time, the orthoimages were resampled with a bilinear interpolation to a resolution of 5 cm. The computation time for segmentation varied from several minutes to several hours according to the computer processor, random access memory (RAM) available and the number of pixels composing the images.

4.2 Classification models

Based on the segmentation, an OBIA was performed. Three separate models were first obtained for sites 1, 2 and 3, respectively, using the training data from each site (specific models). A fourth model (global model) was trained with all the training data the three sites. Each of the specific models can then be used to obtain predictions (i.e. classified segments) in their respective sites, whereas the global model can be used for prediction purposes across all three sites.

As a preliminary step to train the models, the spectral statistics per segment were calculated on the basis of the different spectral bands created beforehand. This process was computed through the command ObjectsradiometricsStatistics from OTB. This command calculates and adds statistic information about the different bands on each segment.

The next step was to prepare the training data for the classification models. Separate Shapefiles were created with the Qgis environment (one for each site). A series of points were manually edited and

classified through photointerpretation in these shapefile. These points were distributed according to a code corresponding to one of the three classes of interest (Sitka spruce seedlings, shades and all other features). As a reminder, the third class is all the features which are not Sitka spruce seedling or shades. It could be logging residues from clear-cutting, logging corridor, grass or bush vegetation vegetation, forestry roads or tree stumps. The following table shows the number of points for the training data on each classes and each sites (table 4). The training data for the model 4 is the set of training data for the other sites compiled. An intersection was then carried out between the point dataset and the segment dataset to create the training data for modelling. One point represents one segment in the training.

Table 4: Training points used for each models distributed by classes, the training points for the model 4 are the training point of the three first models gathered.

Class	Code	Model 1	Model 2	Model 3	Model 4
Seedlings	1	287	238	216	741
Shades	2	234	248	151	633
Others	3	256	219	212	687
Total		777	705	579	2061

The models were trained with the RF algorithm using the training data and applied over the segments previously created. The variables used in the process were the mean and standard deviation of the five bands used in the segmentation. RF was chosen because it has been applied in many previous studies to deal with remote sensing data in the field of land cover mapping (Havryliuk et al., 2018; Oddi et al., 2021; Pelletier et al., 2016). Moreover, this classifier is less sensitive to overfitting than others. Another advantage is that it is relatively robust to aberration and noise in the data (Breiman, 2001).

Some parameters in the RF algorithm had to be tuned. One of them, the number of trees, has an impact on the accuracy and the prediction time. Typically, the accuracy increases and reaches an asymptote for a certain number of trees whereas the running time increases linearly with the number of trees. This parameters was set to 120 after some preliminary tests, and maximises the accuracy without compromising computing time. Another parameter that needed tuning was the maximal depth and it was set to 10. The tree building stops when this quantity is reached. Finally, the size of the randomly selected subset of features at each tree node was set to default value. This means that the number of features selected is the square root of the variables, 3 in this case, as the total number of variables was 10.

To assess the importance of variables used in RF, the mean decrease in Gini coefficient was computed for all models. This metric measures the contribution of each variable to the homogeneity of the nodes in the resulting of RF (Martinez-Taboada and Redondo, 2020) and it is a fundamental outcome of an RF. The variables are usually plotted in descending importance to facilitate its understanding.

4.3 Model evaluation and validation

In order to evaluate and validate the models, we used confusion matrices, which consist in a comparison between predictions provided by the models and validation data. The confusion matrix allows easily detecting correct and incorrect predictions (i.e. correct and incorrect classification). In addition, other metrics such as the producer's accuracy (PA), the users accuracy (UA) or the F-score can be calculated from the confusion matrix. PA is the probability that a segment that belongs to a given class in reality was correctly classified. The user accuracy UA is the probability that a segment predicted for a class is really in that class. In practice, this is the fraction of correctly predicted value to the total number of values predicted to be in a class. The F-score takes into account the producer and user accuracy.

In addition to the confusion matrix, the coefficient kappa (Cohen, 1960) was computed. This statistic represents the global accuracy of the model:

$$\kappa = \frac{N \sum_{i=1}^{l} x_{ii} - \sum_{i=1}^{l} (x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{l} (x_{i+} * x_{+i})}$$

where:

- N is the total number of observations,
- l is the number of lign and colomns of the confusion matrix,
- x_{ii} is the number of observations in lign and colomn i,
- x_{i+} is the marginal sum of the lign i,
- x_{+i} is the marginal sum of colomn i

Landis and Koch established a qualitative scale of classification depending on the value of Kappa coefficient (Landis and Koch, 1977). This qualitative scale will allow determining the quality of the classification (table 5).

Table 5: Agreement measures for categorical data from Landis and Koch

Kappa	Strength of agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

The reference datasets involved in the classification models and their evaluation/validation consists of the already described training data (used to train the supervised models) and the validation data. The validation data can in turn be grouped into two categories: a pseudo-independent dataset and a completely independent dataset. The pseudo independent dataset consists of the new points from the images with which the training was carried out (site 1, 2 and 3). These points were randomly chosen in R and they were different from the training data, leading to three separate shapefiles in Qgis (one corresponding to each sites). The totally independent dataset was edited using site 4, with randomly selected points. The next table shows the number of points used for validation for each sites (table 6).

Table 6: Validate points edited and classified for the different sites. The validate points were used to asses robustness of the models.

Class	Code	Site 1	Site 2	Site 3	Site 4
Seedlings	1	61	53	138	108
Shades	2	118	67	139	62
Others	3	367	433	205	525
Total		546	553	482	695

The specific models and the global model were evaluated following the aforementioned methodology using the pseudo-independent validation datasets. Additionally, the best specific model and the global model were validated with the totally independent data coming from site 4.

5 Results

5.1 Segmentation

The following figures are presented to give an idea of the quality of the segmentation according to the different parameters chosen. The numbers of polygons varied in the different images (table 7). The segmentation seemed to be visually accurate across the different sites. Some inaccuracies have been find for the segmentation, for example, small trees grouped with ground. These segments remain rare and and they are not expected to decrease the accuracy of the classification.

Table 7: Number of polygons provided by the segmentation and total surface for the different sites.

	Polygons(n)	Surface(ha)
Site 1	39 115	1.65
Site 2	102 530	4.53
Site 3	55 558	2.65
Site 4	43 578	2.15

Figure 11 shows the segmentation for all the sites of this study, which seems to be correct and in accordance with the orthoimages. A big number of polygons can be observed with the remnant area due to the variation between shades and wood material for these areas. The segmentation for the site 3 and site 4 shows large polygons, especially for the shade area. It is possible to see that some seedlings are clearly divided into multiple segments, as their size is enough to create some variation into the spectral value within the crown. This does not represent an issue for the classification. As a final remark, the shades of the sites 4 are less pronounced than in the site 3 and smaller in size.

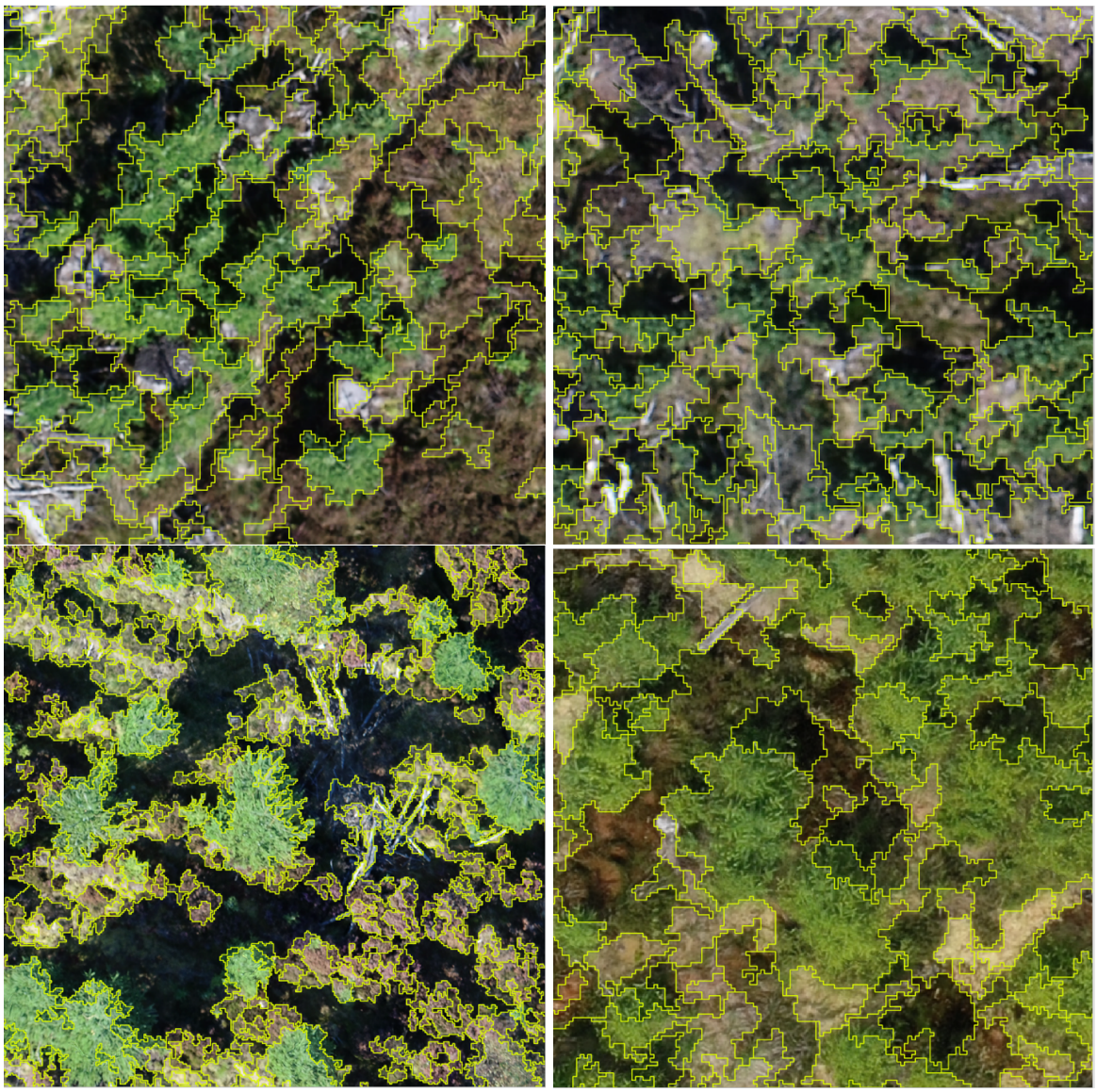

Figure 11: Segmentation in yellow color in overlay of the orthoimages processed for the site 1 and site 2 on the top respectively on the left and right and for the site 3 and site 4 at the bottom respectively on left and right.

5.2 Classification models

The mean decrease Gini was used to asses the variable importance within the RF classification (figure 12). The importance of the variables varies slightly across the different models. Despite this, the mean of band green (B2) is the main variable in all the models, except the global, where the brightness band (B4) is prevalent. The mean decrease Gini plot shows that the mean computed for the bands have more importance than the standard deviation.

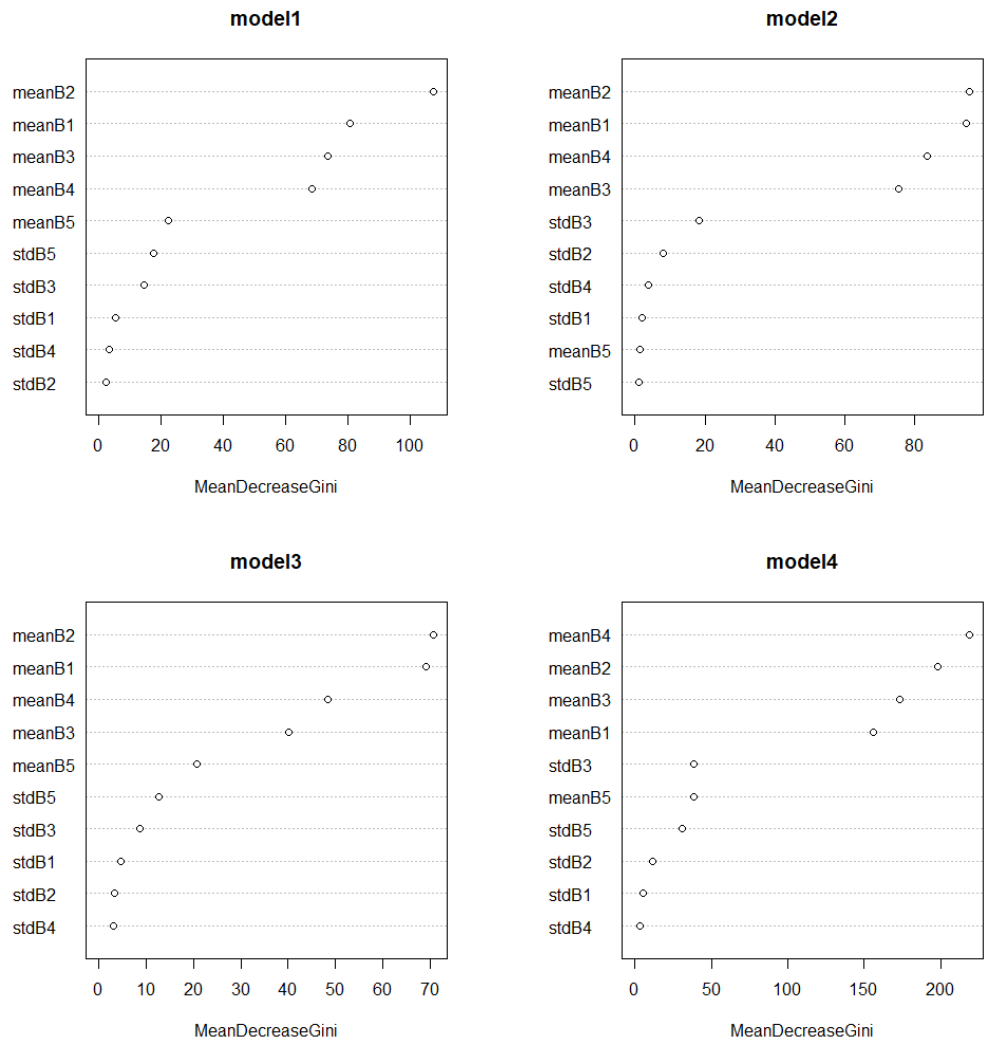

Figure 12: Mean decrease Gini plot for the four models, B1 to B5 referred to the band number in the segmentation meaning respectively, red, green, blue, brightness and GRVI. Std corresponds to standard deviation.

In order to illustrate the outputs from the classification, the three specific models were used to classify a small sample area of their respective sites. Also for these purposes, model 3 was in addition used to classify a fragment of site 4. The global model was applied over the four sites to compare to the specific models.

The results of the classification with the specific model for the site 1 seem accurate when the prediction are compared to the orthoimages. In the area of the figure 13, the seedlings are easily discriminate from the rest of the vegetation, which is mostly brown. In other part of this site however growing grass vegetation was sometimes confused with Sitka spruce seedlings. The shades are in general correctly defined.

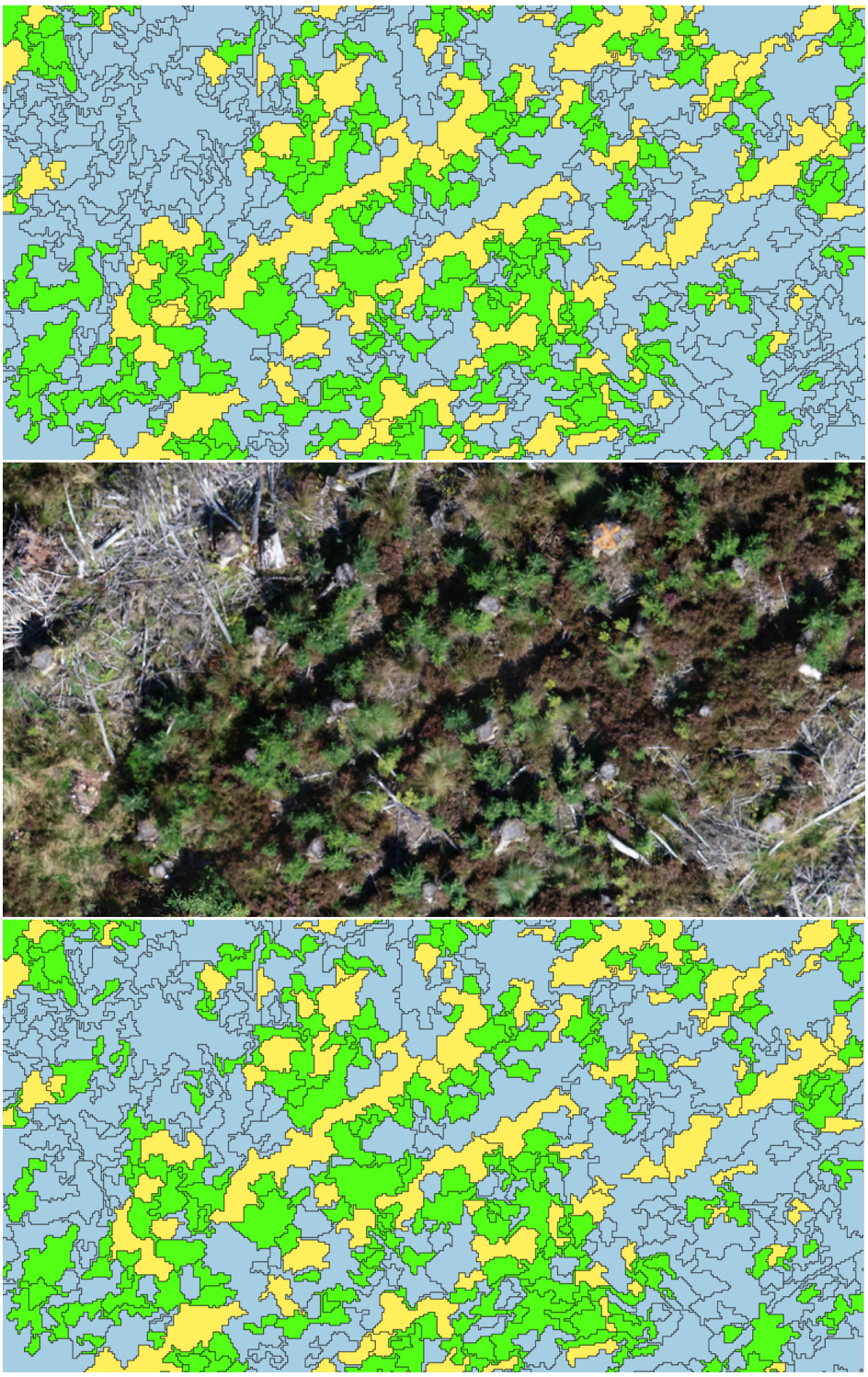

Figure 13: Predictions of the model 1 and model 4 for a small area of the site 1. From top to bottom predictions model 1, orthoimage, predictions model 4. The legend is the following: green for Sitka spruce predictions, yellow represents shade predictions and blue is the others features predictions.

When the orthoimage is compared to the prediction (model specific) in site 2, some misclassifications are clearly identified on figure 14. On the right, There is a large dark area classified as Sitka spruce that very obviously corresponds to the third. Many other errors of this nature can easily be spotted at first glance. The model seems to overestimate the amount of segments classified as seedlings according to this figure.

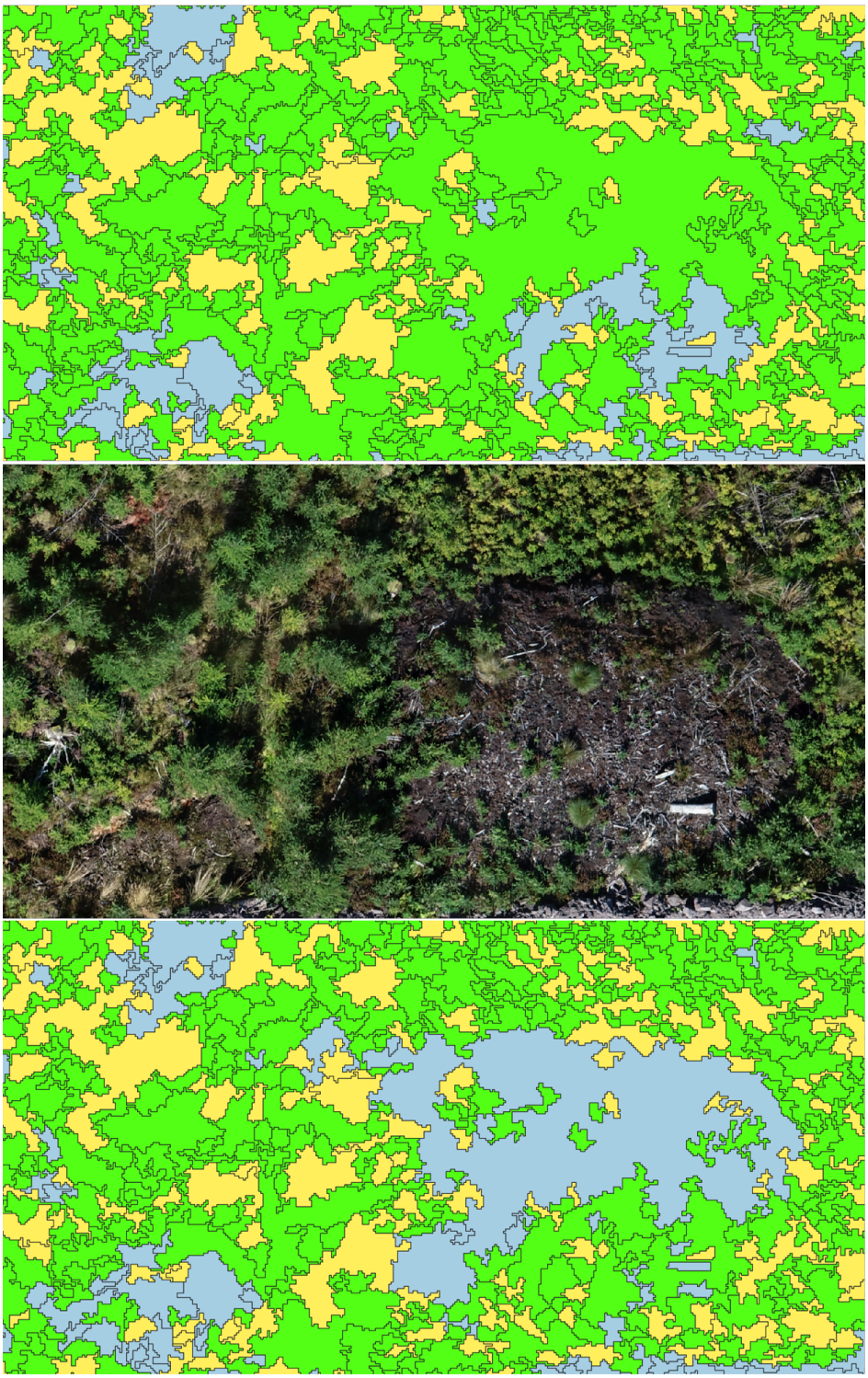

Figure 14: Predictions of the model 2 and model 4 for a small area of the site 2. From top to bottom predictions model 2, orthoimage, predictions model 4. The legend is the following: green for Sitka spruce predictions, yellow represents shade predictions and blue is the others features predictions.

The results for site 3 (figure 15) seem to indicate that the detection of the seedling class is high, although an underestimation in the global number pixels assigned to this class is observed (e.g. the individuals are correctly identified but the extension of their crowns is not fully captured). This could be due to cast and self shadow. Grass and bushes are easily recognisable from the seedlings.

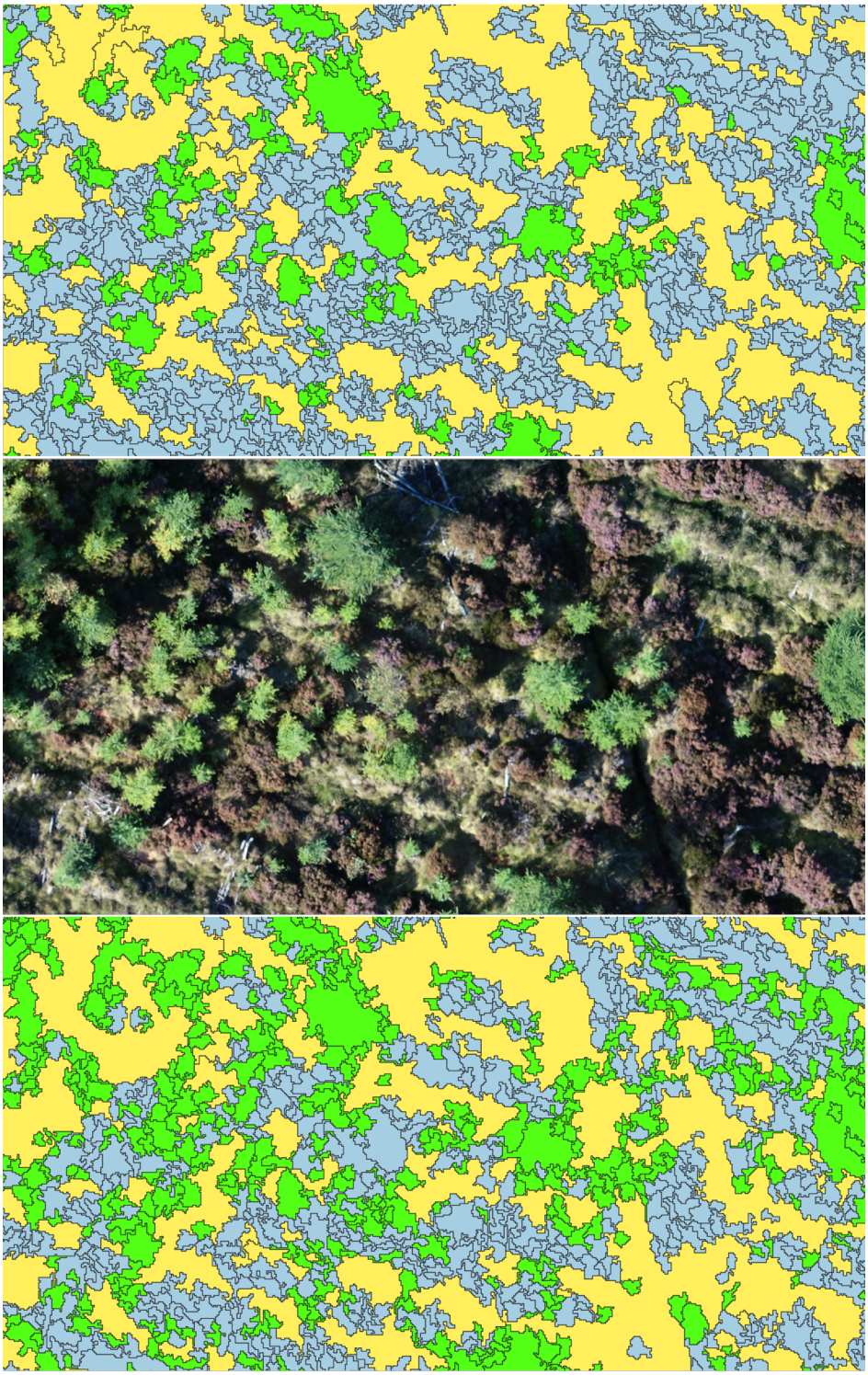

Figure 15: Predictions of the model 3 and model 4 for a small area of the site 3. From top to bottom predictions model 3, orthoimage, predictions model 4. The legend is the following: green for Sitka spruce predictions, yellow represents shade predictions and blue is the others features predictions.

The next figure shows the results of classification for the site 4 using model 3 (figure 16). The classification seems in general accurate for the specific model. On the contrary, the model 4 applied over this site present a high number of misclassfication, the majority of the segments are misclassified as seedling.

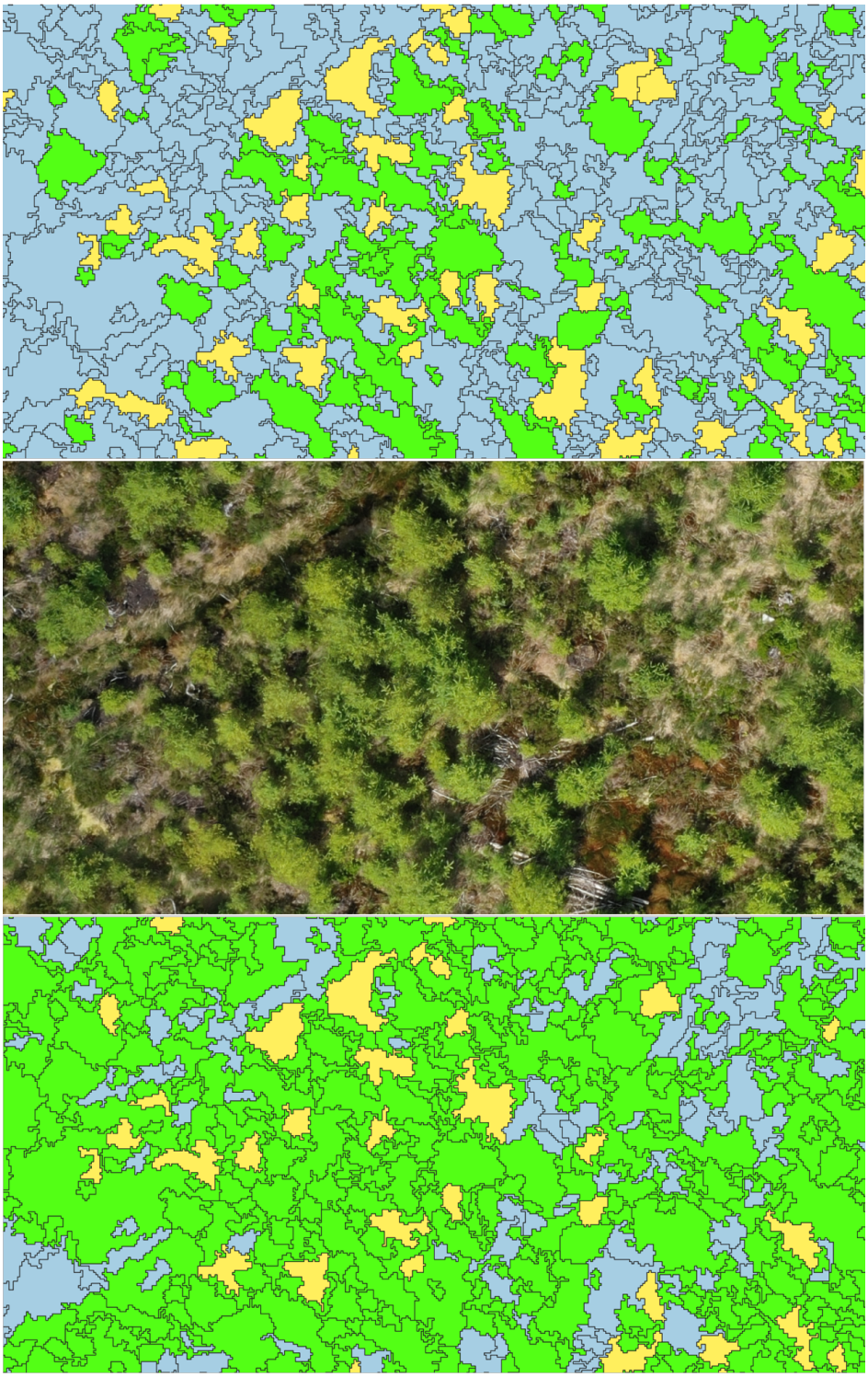

Figure 16: Predictions of the model 3 and model 4 for a small area of the site 4. From top to bottom predictions model 3, Orthoimage, predictions model 4. The legend is the following: green for Sitka spruce predictions, yellow represents shade predictions and blue is the others features predictions.

5.3 Model evaluation and validation

The confusion matrix (table 8) shows that, although most segments were correctly classified, some confusion does exist when applying the specific models (1,2,3) to the pseudo-independent validation datasets. Class 1 (Sitka spruce seedlings) is commonly confused with class 3 (any other features). Conversely, class 3 is mainly confused with the class 1, except for the specific model 3. The shades for the different models are confused with the two others classes except for the model 3 where the shades are only misclassified as the third class. This one is rarely confused with the second class (shades). The global model yields similar results across all sites, except for site 3 where the seedlings were slightly worse classified and all other features were slightly better classified.

Table 8: Confusion matrix for the site 1,2 and 3 with the corresponding model for each sites (model specific) and the model 4 in addition for all the sites (model global). Land cover classes are denoted as follows: 1 = Sitka spruce; 2 = shades; 3 = others features. The Rows correspond to produced labels and the columns correspond to reference labels.

Site 1					Site 2					Site 3				
Model 1					Model 2					Model 3				
Classes	1	2	3	Total	Classes	1	2	3	Total	Classes	1	2	3	Total
1	52	26	71	149	1	46	18	115	179	1	93	0	9	102
2	1	74	2	77	2	0	45	3	48	2	4	95	3	102
3	8	18	294	320	3	7	4	315	326	3	41	44	193	278
Total	61	118	367	546	Total	53	67	433	553	Total	138	139	205	482
Model 4					Model 4					Model 4				
Classes	1	2	3	Total	Classes	1	2	3	Total	Classes	1	2	3	Total
1	57	35	104	196	1	43	16	110	169	1	126	44	51	221
2	0	71	1	72	2	1	47	5	53	2	1	87	3	91
3	4	12	262	278	3	9	4	318	331	3	11	8	151	170
Total	61	118	367	546	Total	53	67	433	553	Total	138	139	205	482

Table 9 presents the results for the different metrics derived from the confusion matrix (UA, PA, F-score) for all the classes and Kappa for the different models. The UA of the Sitka spruce seedlings is highest with the model 3 (91.2). For the same model and class, PA is the lowest, with 67.4. It therefore presents a lower detection rate of seedlings but a higher specificity. The value of F-score obtained for the seedlings for the model 3 in site 3 is sightly better than that of the other sites. The second class presents similar results across the different models.

The Kappa index corresponding to the global accuracy of each model moderately varies across the different sites but it remains noticeably constant within a single site when the specific model is compared to the global model. The lowest value is for the site 2 and the better accuracy is for the site 3 with 66.9 (table 9). According to the table of Landis and Koch, the classification of the site 1 and 2 is moderate and the classification of the site 3 is substantial.

Table 9: UA (User accuracy), PA (Producer accuracy), F (F-score) and kappa produced with the confusion matrix for the site 1,2 and 3 with specifics models and global model.

	Site 1			Site 2			Site 3		
	Model 1			Model 2			Model 3		
	UA	PA	F	UA	PA	F	UA	PA	F
Sitka spruce	34.9	85.2	49.5	25.7	86.8	39.7	91.2	67.4	77.5
Shades	96.1	62.7	75.9	93.8	67.2	78.3	93.1	68.3	78.8
Others	91.9	80.1	85.6	96.7	72.7	83.0	69.4	94.1	79.9
$Kappa$		57.7			46.5			66.9	
	Model 4			Model 4			Model 4		
	UA	PA	F	UA	PA	F	UA	PA	F
Sitka spruce	29.1	93.4	44.4	25.4	81.1	38.7	57.0	91.3	70.2
Shades	88.7	70.1	78.3	93.8	67.2	78.3	95.6	62.6	75.7
Others	94.2	71.4	81.2	96.0	73.4	83.2	88.8	73.7	80.5
$Kappa$		51.5			46.5			63.1	

The site 4 and the validation points edited for this area were used to test the robustness of the best specific model in terms of accuracy (model 3) and the global model (model 4), which somehow gathers information from all sites. The confusion matrix for the two models is displayed to show the distribution of the different validation points through the classes (table 10). The seedlings class tended to be correctly classified with both models, particularly with model 4. The shades were on the contrary mostly wrongly classified by both models. Model 3 confuses the shades with class 3, whereas model 4 tends to assign classify shades as class 1. Class 3 is correctly classified by model 3. Model 4 presents a lot of misclassification points in class 3, which assigned class 1.

Table 10: Confusion matrix produce for the site 4 with the model 3 (model specific) and the model 4 (model global). Land cover classes are denoted as follows: 1 = Sitka spruce; 2 = shades; 3 = others features. The Rows correspond to produced labels and the columns correspond to reference labels.

Site 4									
	Model 3					Model 4			
Classes	1	2	3	Total	Classes	1	2	3	Total
1	90	0	11	101	1	106	38	295	439
2	1	30	3	34	2	1	18	1	20
3	17	32	511	560	3	1	6	229	236
Total	108	62	525	695	Total	108	62	525	695

The result presented in the table 11 shows that the F-score and global accuracy represented by Kappa index decreases when the global model is applied to site 4, in contrast to model 3, which performs even better than in site 3 (global accuracy 74.7). The shades are in general more poorly classify than on the other sites but the seedling reaches a f-score of 86.1 which is the highest value for the validating. According to the table of Landis and Koch, the supervised classification is substantial for the site 4 when using model 3.

Table 11: UA (User accuracy), PA (Producer accuracy), F (F-score) and kappa produced with the confusion matrix for the site 4 with the model 3 and the model 4.

	Site 4					
	Model 3			Model 4		
	UA	PA	F	UA	PA	F
Sitka spruce	89.1	83.3	86.1	24.1	98.1	38.8
Shades	88.2	48.4	62.5	90.0	29.0	43.9
Others	91.3	97.3	94.2	97.0	43.6	60.2
$Kappa$		74.7			23.4	

6 Discussion

In order to meet the objective of this dissertation, a number of intermediate methodological steps have been taken within a OBIA framework: (i) automatic segmentation of orthoimages, (ii) preparation of training, pseudo-independent and completely independent validation datasets, (ii) training of an RF algorithm to produce a classification model, and (iv) validate such model. As a result of this effort, a preliminary model to detect natural regeneration of Sitka spruce in Scotland from orthoimages has been created for the first time and, more importantly, a methodological protocol has been set for future work in this area, which will be of great interest to practitioners when making decisions around natural regeneration of Sitka spruce.

6.1 Segmentation

The first part of the workflow was the segmentation, which, based on visual observation, seemed to be accurate enough to carry out a supervised classification. Some isolate seedlings were grouped together with the ground but the majority of the image looked well segmented. Those isolate stems of small area were probably missed due to the segment minimum size parameter that was set at the beginning of the segmentation process. This parameters could still be reduced to take into account the smallest and isolated seedlings.

Despite this sort of errors and the fact that it was a very time consuming process compared to the next stages, the advantages of the segmentation largely exceed its disadvantages. Conducted on orthoimages with a 5 cm resolution and with a few bands (RGB bands, Brightness, GRVI), the approach massively contributed to save time with the classification part, which would have otherwise been extremely slow due to the huge number of pixels that form these images. Crucially, once the segmentation has been programmed, it can safely be applied to all future work of model refinement, which will be more efficient and rigorous than in the absence of segmentation. A conceptual drawback of this approach is the fact that the segmentation itself cannot be validated with the methodology developed in this study. Nonetheless, the evidence presented suggests that the any inaccuracy in the models is mostly related to causes other than errors in the segmentation, as it will be described in the next section.

6.2 Classification model, evaluation and validation

Ideally, one would expect that the order of variable importance is the same for all models trained through RF because same classes are classified using the same predictors. However, the mean decrease Gini plots showed that the order of the variables is somehow different in this case. This may be an indication that the images are different in terms of spectral signature. Another reason for this relative lack of consistency in the importance order may be that the number of training points varies between the different models, especially with the global model, which gathers training points from all three sites. In spite of this, it is reassuring that the four main variables used by all models in the classification did remain the same.

The results achieved in terms of UA and PA in this study were lower than those in (Feduck et al., 2018), who worked with a similar approach. However, their studied area contained no vegetation other than the seedlings object of the study. Unfortunately, the authors did not calculate the Kappa index leading, which makes it impossible to compare this classification metric. Nonetheless, it is important to notice that global accuracy obtained in the present study for model 3 was substantial according to the classification scale in (Landis and Koch, 1977).

The misclassification between classes observed within the different models are mostly found between class 1 (seedlings) and class 3 (any other feature). This is surely due to the presence of photosynthetically active vegetation such as grass or heather along with Sitka spruce seedlings, which may present similar reflectances. Images acquired during other periods would allow to discriminate more easily the seedlings from the other vegetation features, especially annual plants but also heather, which dramatically changes colour over the seasons. In this respect, April and May seems the ideal months to obtain images for the purposes of assessing regeneration because herbaceous vegetation would not have yet emerged and heather would not have significant photosynthetic activity due to the low temperatures. Earlier than that, assessments are not recommended given the possible presence of snow.

The training point number is lowest for the model 3 but the global accuracy is the highest. This is attributable to the characteristics of site 3, where seedlings are significantly larger than in the other two sites. Moreover, the orthoimage for this site exhibits a lot of shades, which are relatively well classified. In addition, the image presents less grass in comparison with site 1 and 2 and the color of any other species is easily discriminated from Sitka spruce seedlings. On top of this, the seedling area is proportionally larger than in the other sites. These findings suggest that regeneration assessments using photogrammetry would be optimised if conducted at least six years after felling. The site 4 shows a situation similar to that of site 3, leading to a high degree of global accuracy (74.7) when applied to these independent data.

The global model works as well as all other models individually in sites 1,2 and 3, although it presents a slightly lower accuracy. Although an improvement in the accuracy due to the higher number of training points can nevertheless be expected, this seems to indicate that it is possible to generalise the predictions by using one single model that has been fitted using many different situations. The fact that model 3 outperformed the global model in the validation (site 4) only stresses the need for a wider collection of images to properly train a model for the purposes of this study. Due to the COVID-19 pandemic, it was unfortunately difficult to obtain more images to train the model from more sites at the right ages and at the right time of the year.

Regardless of the number of training images, a mean to improve the models is to increase the number of training segments. Caution needs to be observed when taking this approach, as it can easily lead to overtraining, which would make the models non-operational at a larger scale. Other way to improve the models could be to add other indices derived from the RGB bands. Moreover, a texture index may be added under the hypothesis that the texture of Sitka spruce seedlings is different from that of other species and that it varies little over the growing season. Although beyond the scope of this study and probably adding costs to the surveys, the use of near infrared and LiDAR point clouds could be an interesting alternative to increase the quality of these models by, respectively, providing more spectral data and adding geometric information such as the height of the seedlings. LiDAR data must be acquired with a high number of points by m^2 to perform a canopy high model accurate enough, given the small crown radii of the seedlings. A digital surface model could be generated through photogrammetric analysis but the dense foliage in forest would make it considerably difficult to generate digital terrain model (Gu et al., 2020; Lisein et al., 2013).

Field validation was not possible, also due to the COVID-19 pandemic, which led to a validation through photo interpretation with Qgis. Field validation could consist in new randomly chosen points distributed within the stands that are then classified on the field (note that this would not be possible for the shades). Despite the lack of field validation, image interpretation seems a reasonable option. The obvious downside of this approach is that the quality of the outputs ultimately depends on the operator/analyst and could become complicated especially when it comes to shade areas or if the seedlings are too young and are hardly distinguishable from the other vegetation.

The classification was performed on four specific sites/images. Due to the changes in spectral information due to the acquisition time, date and weather conditions, using several aerial photographs of the site that consider those variations site at different hours could enhance the robustness of the model.

The proportion of shade area is about 15 %, 8 % , 37 %, 6 % for sites 1,2,3 and 4, respectively. The proportion of shades is especially high for site 3 and site 1. Photos taken at the zenith would reduce the size of shadows and increase the amount of information that is hidden to the user. Nonetheless shades will probably remain an issue in Scotland due to the high latitude. One possible assumption, should further analysis be undertaken, could be that the shade area contains the same proportion of natural regeneration than the non-shade area (Manso and McLean, 2020).

6.3 Implications for management

In spite of the aforementioned limitations, this study sets the methodological bases to produce accurate regeneration maps for Sitka spruce that, ultimately, could assist forest managers to make optimal decisions. Recall that classic inventories of natural regeneration at the operational scale are unaffordable and informed forestry choices can hardly be done at the moment. A perfected version of this model will help with these decisions.

Managers who choose to rely on natural regeneration are confronted with two problems: high stocking in densely regenerated areas and poor stocking in regeneration gaps across the regenerated landscape. If no respacing carried out in the high stocking areas, the excessive competition will lead to good timber properties but poor timber assortments. On the contrary, if no restocking is conducted poor stocking areas would cause timber products to not be suitable for structural uses (Price, 2016). Both cases are detrimental from the business point of view, although the increasing demand for biomass may still provide some income. Managers can, on the contrary, opt for respacing high stocking areas and restocking gaps, in which case the associated cost and labour may or may not compensate the increase in product value. The third alternative, the current scenario, is to get rid of the natural regeneration and to replant, which is more straightforward but also requires an initial investment.

In order to help managers to make this decision, the predictions from a model like that presented in this dissertation will need to be further analysed to properly understand the spatial patterns of the established seedlings – or, conversely, they regeneration gaps. This would serve to better quantify the cost and effort of the silviculture operations mentioned before. To this aim, a number of dispersal indices could be calculated from the classification maps. Some examples are the proximity index (Gustafson and Parker, 1994), which takes into account the patch size and the proximity between the patches; the Moran's index, which deals with the spatial correlation (Moran, 1950); or the Ripley's K and L functions, which provides a measure of spatial homogeneity (Ripley, 1977). The values of these indices in natural regeneration sites could be compared with the ideal situation of planted Sitka spruce seedlings.

From a more ecological point of view, the model could be useful to determine whether natural regeneration has successfully been achieved. Although Sitka spruce is not eligible to obtain financial support in Scotland given its non native status, natural regeneration adds a degree of diversity that is environmentally beneficial. As a guideline, the Scottish government provides values to evaluate whether regeneration of native species has been successful (at least a minimum of 20 % canopy cover over a minimum of 80 % of the area, Forest Enterprise Scotland, 2017) and these could be used to assess the regeneration of Sitka spruce as well. On this note, the method proposed here can actually be applied to other native species of interest, such as *Pinus sylvestris*, with their higher value in terms of biodiversity.

7 Conclusion

In this study, we evaluated a supervised classification aimed at mapping Sitka spruce seedlings in Scotland based on orthophotographs acquired through a UAV mounted RGB camera. The followed method included a preliminary segmentation process of the different scenes acquired by the drones. Based on this segmentation, a supervised learning with a RF algorithm was computed to create three site specific models and a global model. The training and validation points of these models was carried out with reference data obtain by photointerpretation over the studied sites.

Due to the seasonal leaf-on condition, grasses, bushes and seedlings are not spectrally distinct, which led to some confusion in the classification of these features, particularly in model 1 and 2. Global accuracy varied for the different models. Model 3 yielded better results with a global accuracy, UA and PA of 66.9, 91.2 and 67.4, respectively. On a leaf off situation, the work carried out in this study would be expected to achieve a higher efficiency for the reasons stated above.

The robustness of model 3 was tested on one independent site and seems promising for the use on a larger scale, at least when large seedlings (i.e. over 6 year old) are present. This remains to be proven through validation over more sites and varying seasonal and timing conditions. On top of robustness, this model woulds still benefit from some refining to improve its PA and UA.

The outputs from any classification model are not readily usable or interpretable by forest managers, whose decisions would strongly be conditioned by the spatial distribution of the natural regeneration. This information can be obtained from this model in the form of spatial indices. The height of the seedlings could also prove helpful. This could be calculated with LiDAR data in the future at the cost of more complex processing and more expensive surveys.

The quantification and assessment of any vegetation type at this level of detail with RGB cameras is a promising and unexpensive development for forest management, which can be applied over a large range of species that holds a massive potential, not only for commercial forestry, but also in the context of ecological restoration or plant health.

References

V. Arevalo, J. González-Jiménez, J. Valdes, and G. Ambrosio. Detecting shadows in quickbird satellite images. 05 2006.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 01 2001. doi: 10.1023/A:1018054314350.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–, 04 1960.

C. Feduck, G. Mcdermid, and G. Castilla. Detection of coniferous seedlings in uav imagery. *Forests*, 9:432, 07 2018. doi: 10.3390/f9070432.

Forest Enterprise Scotland. *Restocking Strategy for the National Forest Estate*, 2017.

Forestry Commission. National forest inventory-proportion of sitka spruce in the upper canopy of each nfi sample plot. 2018. URL `https://www.forestresearch.gov.uk/documents/5387/FR_NFI_SITKA_SPRUCE_SPCS_SCOTLAND.pdf`.

Forestry Commission. *Forestry Statistics 2020 – A compendium of statistics about woodland, forestry and primary wood processing in the United Kingdom*. Forestry Commission, Edinburgh, 2020.

M. Fromm, M. Schubert, G. Castilla, J. Linke, and G. Mcdermid. Automated detection of conifer seedlings in drone imagery using convolutional neural networks. *Remote Sensing*, 11:2585, 11 2019. doi: 10.3390/rs11212585.

GISgeography. *OBIA – Object-Based Image Analysis (GEOBIA)*, 2021. URL `https://gisgeography.com/obia-object-based-image-analysis-geobia/`.

T. Goodbody, N. Coops, T. Hermosilla, P. Tompalski, and P. Crawford. Assessing the status of forest regeneration using digital aerial photogrammetry and unmanned aerial systems. *International Journal of Remote Sensing*, 39:1–19, 11 2017. doi: 10.1080/01431161.2017.1402387.

M. Grizonnet, J. Michel, V. Poughon, J. Inglada, M. Savinaud, and R. Cresson. Orfeo toolbox: open source processing of remote sensing images. *Open Geospatial Data, Software and Standards*, 2, 06 2017. doi: 10.1186/s40965-017-0031-6.

J. Gu, H. Grybas, and R. Congalton. A comparison of forest tree crown delineation from unmanned aerial imagery using canopy height models vs. spectral lightness. *Forests*, 11:605, 05 2020. doi: 10.3390/f11060605.

E. Gustafson and G. Parker. Using an index of habitat patch proximity for landscape design. *Landscape and Urban Planning*, 29:117–130, 08 1994. doi: 10.1016/0169-2046(94)90022-1.

S. Havryliuk, M. Korol, Tokar, V. Olena, and K. Lubov. Using the random forest classification for land cover interpretation of landsat images in the prykarpattya region of ukraine. 09 2018. doi: 10.1109/STC-CSIT.2018.8526646.

R. Jannoura, K. Brinkmann, D. Uteau, C. Bruns, and R. Joergensen. Monitoring of crop biomass using true colour aerial photographs taken from a remote controlled hexacopter. *Biosystems Engineering*, 129, 11 2014. doi: 10.1016/j.biosystemseng.2014.11.007.

M. Kelly, S. Blanchard, E. Kersten, and K. Koy. Terrestrial remotely sensed imagery in support of public health: New avenues of research using object-based image analysis. *Remote Sensing*, 3: 2321–2345, 12 2011. doi: 10.3390/rs3112321.

C. Kirby. A camera and interpretation system for assessment of forest regeneration. 1980.

S. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31, 2007.

F. Kressler, M. Franzen, and S. Klaus. Segmentation based classification of aerial images and its potential to support the update of existing land use data bases. 01 2005.

J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33: 159–74, 04 1977. doi: 10.2307/2529310.

Y. li, P. Gong, and T. Sasagawa. Integrated shadow removal based on photogrammetry and image analysis. *International Journal of Remote Sensing - INT J REMOTE SENS*, 26:3911–3929, 09 2005. doi: 10.1080/01431160500159347.

T. Lillesand, R. Keifer, and J. Chipman. *Remote Sensing and Image Interpretation*. Seventh edition edition, 2015.

J. Lisein. *Application des techniques de photogrammétrie par drone à la caractérisation des ressources forestières*. PhD thesis, Université de Liège - Gembloux Agro-Bio Tech, Université de Paris Est, 2016.

J. Lisein, M. Deseilligny, S. Bonnet, and P. Lejeune. A photogrammetric workflow for the creation of a forest canopy height model from small unmanned aerial system imagery. *Forests*, 4:922–944, 12 2013. doi: 10.3390/f4040922.

R. Manso and P. McLean. *UAV survey to assess natural regeneration in Sitka spruce*. Forest Research, 2020.

F. Martinez-Taboada and J. I. Redondo. Variable importance plot (mean decrease accuracy and mean decrease gini)., Apr 2020. URL `https://plos.figshare.com/articles/figure/Variable_importance_plot_mean_decrease_accuracy_and_mean_decrease_Gini_/12060105/1`.

F. Mesas-Carrascosa, I. Rumbao, J. Berrocal, and A. García-Ferrer. Positional quality assessment of orthophotos obtained from sensors onboard multi-rotor uav platforms. *Sensors*, 14:22394–22407, 12 2014. doi: 10.3390/s141222394.

J. Moore. Wood properties and uses of sitka spruce in britain. *Forestry Commission Research Report.Edinburgh*, pages 1–48, 2011.

P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37:17–23, 07 1950. doi: 10.2307/2332142.

T. Motohka, K. Nasahara, O. Hiroyuki, and T. Satoshi. Applicability of green-red vegetation index for remote sensing of vegetation phenology. *Remote Sensing*, 2, 10 2010. doi: 10.3390/rs2102369.

P. Nygaard and B.-H. Øyen. Spread of the introduced sitka spruce (picea sitchensis) in coastal norway. *Forests*, 8:24, 01 2017. doi: 10.3390/f8010024.

L. Oddi, E. Cremonese, L. Ascari, G. Filippa, M. Galvagno, D. Serafino, and U. Morra di Cella. Using uav imagery to detect and map woody species encroachment in a subalpine grassland: Advantages and limits. *Remote Sensing*, 13:1239, 03 2021. doi: 10.3390/rs13071239.

OTB. *Cookbook:A Guide for OTB-Applications and Monteverdi Dedicated for Non-developers*, 2021. URL `https://www.orfeo-toolbox.org/CookBook/`.

C. Pelletier, S. Valero, J. Inglada, N. Champion, and G. Dedieu. Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*, 187:156–168, 12 2016. doi: 10.1016/j.rse.2016.10.010.

A. Price. *Effects of early release of Picea sitchensis natural regeneration on the mechanical properties of the juvenile and mature wood*. PhD thesis, University of Aberdeen, 2016.

QGIS Development Team. *QGIS Geographic Information System*. QGIS Association, 2021. URL `https://www.qgis.org`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `R-project.org`.

B. Ripley. Modeling spatial patterns (with discussion). *Journal of the Royal Statistical Society Series B*, 39:172–212, 01 1977.

G. Schaepman-Strub, M. Schaepman, T. Painter, S. Dangel, and J. Martonchik. Reflectance quantities in optical remote sensing-definitions and cases studies. *Remote sensing of Environment*, (103):27–42, 2006.

A. Shahtahmassebi, N. Yang, K. Wang, N. Moore, and Z. Shen. Review of shadow detection and de-shadowing methods in remote sensing. *Chinese Geographical Science*, 23:403–420, 08 2013. doi: 10.1007/s11769-013-0613-x.

T. Thanh Ngo. *Shadow/Vegetation and building detection from singleoptical remote sensing image*. PhD thesis, Université de Strasbourg, 2015.

G. Willhauck. Comparison of object oriented classification techniques and standard image analysis for the use of change detection between spot multispectral satellite images and aerial photos. *International Archives of Photogrammetry and Remote Sensing*, 33, 01 2000.

B.-H. Øyen and P. Nygaard. Impact of sitka spruce on biodiversity in nw europe with a special focus on norway – evidence, perceptions and regulations. *Scandinavian Journal of Forest Research*, 35: 1–17, 04 2020. doi: 10.1080/02827581.2020.1748704.

Appendix

This appendix shows a part of the code applied for this study. It presents the code for model 1 as well as the generation of model 4, the robustness evaluation of model 3 and 4 and finally the generation of the mean decrease Gini plot. To get the entire code, you can contact the following mail address: yann.lahaise@gmail.com

```
#code TFE
#Yann Lahaise
#-----------------------------------------------

install.packages("RStoolbox")
library(raster)
library(future)
library(gdalUtils)
library(sf)
library(dplyr)
library(RStoolbox)
library(lwgeom)
library(lidR)
library(dplyr)

# Path OTB
# These paths are used in command line

path_otb = "C:/OTB-7.2.0-Win64/bin/"
path_gdal = "C:/OSGeo4W64/bin/"
path_ogr =  "C:/OSGeo4W64/OSGeo4W.bat ; C:/OSGeo4W64/bin/"
path_gdal_py = "C:/OSGeo4W64/OSGeo4W.bat ; python C:/OSGeo4W64/bin/"

# checking path exists
dir.exists(path_otb)
dir.exists(path_gdal)
#-----------------------------------------------------------------------
#Model 1
#-----------------------------------------------------------------------
#preprocessing and segmentation(1)
#-------------------------------
#output
path0="D:/yannl/Documents/cours_gbx_master_2/TFE/image"
path_out=paste0(path0,"/output_tfe")
if(!dir.exists(path_out)){
  dir.create(path_out)
}

#Ortho
path_in="D:/yannl/Documents/cours_gbx_master_2/TFE/image"
file_site1=paste0(path_in,"/A7166_Site_1.tif")
site1=raster(file_site1)

#extent site 1
ext=extent(site1)
ext=round(ext,digits=0)
ext

#extent for gdal
bbox_gdal=paste(ext[1],ext[4],ext[2],ext[3])
bbox_gdal=round(bbox_gdal,digits=0)
```

```
#resample ortho
file_in=file_site1
file_out=paste0(path_out,"/site1_5cm.tif")

cmdline = paste0(path_gdal,"gdal_translate ",
                 " -tr 0.05 0.05 ",
                 " -projwin ", bbox_gdal,
                 " -r bilinear ",
                 file_in," ",
                 file_out)
system(cmdline)

#brigthness computing
inputs=paste0(path_out,"/site1_5cm.tif")
output=paste0(path_out,"/brightness5cm.tif")
expr='"(im1b1+im1b2+im1b3)/3"'
cmdline= paste0(path_otb,"otbcli_bandMathX",
                " -il ", inputs,
                " -out ", output,
                " uint16 -ram 4000 ",
                " -exp ", expr)
system(cmdline)

#GRVI computing
inputs=paste0(path_out,"/site1_5cm.tif")
output=paste0(path_out,"/greenred5cm.tif")
expr = '"im1b2>0?im1b1>0?(im1b2-im1b1)/(im1b2+im1b1)*1000+1000:0:0"'
cmdline = paste0(path_otb,"otbcli_BandMathX " ,
                 " -il ", inputs,
                 " -out ", output,
                 " uint16 -ram 4000 ",
                 " -exp ", expr)
system(cmdline)

#rescale GRVI
file_in=paste0(path_out,"/greenred5cm.tif")

file_out=paste0(path_out,"/green_red_512_5cm.tif")
cmdline = paste0(path_gdal,"gdal_translate ",
                 " -scale 0 1714 1 512",
                 " -ot UInt16 ",
                 file_in," ",
                 file_out)
system(cmdline)

#vrt computing with 3 bandes of the ortho (RGB) + GRVI + Brightness

file_site1=paste0(path_out,"/site1_5cm.tif")
list_band=list()
for (kband in 1:3){
  print(kband)
  file_vrt=paste0(path_out,"/",kband,".vrt")
  gdalbuildvrt(gdalfile=file_site1 , b=kband,
               output.vrt= file_vrt ,overwrite=TRUE)
  list_band=rbind(list_band,file_vrt)
}

file_brightness=paste0(path_out,"/brightness5cm.tif")
```

```
list_band=rbind(list_band,file_brightness)
file.exists(file_brightness)

file_greenred=paste0(path_out,"/green_red_512_5cm.tif")
file.exists(file_greenred)
list_band=rbind(list_band,file_greenred)

list_band=list_band[,1]
list_band
file_vrt=paste0(path_out,"/site1_stack_5_5cmbands.vrt")
gdalbuildvrt(gdalfile=list_band, separate=TRUE,
             output.vrt= file_vrt,overwrite=TRUE,
             allow_projection_difference=TRUE)

#VRT sample for segmentation parameters test
bbox =" 296020 589340 296040 589310" # xul, yul, xlr, ylr
vrt_in=paste0(path_out,"/site1_stack_5_5cmbands.vrt")
vrt_out = paste0(path_out,"/stack_5bands_sample.vrt")
cmdline = paste0(path_gdal,"gdal_translate ",
                 " -ot UInt16 ",
                 "-of VRT ",
                 " -projwin ", bbox, " ",
                 vrt_in," ",vrt_out)
system(cmdline)

#Test parameters spatial radius and range radius
file_in=paste0(path_out,"/stack_5bands_sample.vrt")
path_out_file=paste0(path_out,"/parameters_segm")
if(!dir.exists(path_out_file)){
  dir.create(path_out_file)
}

result=data.frame(spatial = numeric(), range= numeric(),
                  dt=numeric(), nbpol=numeric())
for (k_spatial in c(5,10,15,20,40)){
  print(k_spatial)
  for (k_range in c(15,20,30,40,50)){
    print(k_range)
    file_out = paste0(path_out_file,"/segm_",k_spatial,
                      "_",k_range,".sqlite")
    if(file.exists(file_out)){
      file.remove(file_out)
    }
    cmdline = paste0(path_otb, "otbcli_Segmentation ",
                     ' -in ' , file_in ,
                     ' -filter "meanshift" ' ,
                     ' -filter.meanshift.spatialr ',k_spatial,' ' ,
                     ' -filter.meanshift.ranger ',k_range,' ' ,
                     ' -filter.meanshift.maxiter 10' ,
                     ' -filter.meanshift.minsize 10' ,
                     ' -mode.vector.out ', file_out ,
                     ' -mode.vector.outmode "ovw" ' ,
                     ' -mode.vector.tilesize 4000 ',
                     ' uint32 ',
                     ' -mode.vector.simplify 0 ')

    t1=Sys.time()
    system(cmdline)
```

```
    t2=Sys.time()
    (t2-t1)
    sf=st_read(file_out)
    sf=st_make_valid(sf)

    k=nrow(result)+1
    result[k,1] = k_spatial
    result[k,2] = k_range
    result[k,3] = t2-t1
    result[k,4] = nrow(sf)
  }
}
result

#segmentation of the orthophoto
file_in=paste0(path_out,"/site1_stack_5_5cmbands.vrt")
file_out = paste0(path_out,"/segm5bands_5cm.sqlite")
file_out
if(file.exists(file_out)){
  file.remove(file_out)
}
cmdline = paste0(path_otb, "otbcli_Segmentation ",
                 '-in ' , file_in ,
                 '-filter "meanshift" ' ,
                 '-filter.meanshift.spatialr 10 ' ,
                 '-filter.meanshift.ranger 30 ' ,
                 '-filter.meanshift.maxiter 10' ,
                 '-filter.meanshift.minsize 32' ,
                 '-mode.vector.out ', file_out ,
                 '-mode.vector.outmode "ovw" ' ,
                 '-mode.vector.tilesize 4000 ',
                 'uint32',
                 '-mode.vector.simplify 0 ')

t1=Sys.time()
system(cmdline)
t2=Sys.time()
(t2-t1)

# sqlite to shp
file_in = paste0(path_out,"/segm5bands_5cm.sqlite")
file_out = paste0(path_out,"/segm_5band_5cm.shp")
segm=st_read(file_in)
segm=st_make_valid(segm)
segm$area=st_area(segm)
st_write(segm,file_out,delete_layer = TRUE)


#spectral statistics computed for each segments
file_vrt=paste0(path_out,"/site1_stack_5_5cmbands.vrt")
file_in = paste0(path_out,"/segm_5band_5cm.shp")

cmdline = paste0(path_otb,'otbcli_ObjectsRadiometricStatistics ',
                 '-in ', file_in ,
                 '-im ', file_vrt ,
                 '-field "dn" ')
t1=Sys.time()
system(cmdline)
```

```
t2=Sys.time()
(t2-t1)

# Lire les segments dans 1 objet sf
file_segm = paste0(path_out,"/segm_5band_5cm.shp")
segm=st_read(file_segm)
names(segm)

# CRS -> 27700
st_crs(segm)=27700

#Training data
file_train=paste0(path_in,"/training_site1.shp")
train=st_read(file_train)
st_crs(train)=27700
nrow(train)
table(train$classe)

# Select segments with one training point
# Class labels attribution to the segments
segm_train=st_intersection(segm,train)
segm_train=st_drop_geometry(segm_train)
segm_train=segm_train[,c("dn","classe")]
head(segm_train)
nrow(train)
nrow(segm_train)
names(segm_train)

# If more than one points in a segment -> deleting duplicate
segm_train2=segm_train %>% group_by(dn) %>%  summarize(classe=min(classe))
segm_train2=as.data.frame(segm_train2)
nrow(segm_train2)
table(segm_train2$classe)

# join the field "class"
segm2=inner_join(segm,segm_train2,by= c("dn" = "dn"))
segm2$classe=as.integer(segm2$classe)
names(segm2)
file_out = paste0(path_out,"/segm_train_site1.shp")
st_write(segm2,file_out,delete_layer = TRUE)
names(segm2)
names(segm1)

#editing validation data
file_site1=paste0(path_out,"/site1_5cm.tif")
site1_5cm<-raster(file_site1)
ext=extent(site1_5cm)
#random coordinate with the extent of the site
x <- runif(1000, ext[1]+15,ext[2]-30)
y <- runif(1000, ext[3]+35,ext[4]-30)
coordinates <- cbind(x,y)
coordinates<-as.data.frame(coordinates)
names(coordinates)
point_layer_site1<-st_as_sf(coordinates,coords=c("x","y"),crs=27700)
#random coordinate shp writing for hand classification
file_out = paste0(path_out,"/point_test_site1.shp")
st_write(point_layer_site1,file_out,delete_layer = TRUE)
```

```
#reading validation data
file_validating=paste0(path_out,"/point_test_site1.shp")
validating=st_read(file_validating)
st_crs(validating)=27700
nrow(validating)
table(validating$classe)

# select segment with one at least one point
# Assign class label to the segment
segm_test=st_intersection(segm,validating)
segm_test=st_drop_geometry(segm_test)
segm_test=segm_test[,c("dn","classe")]
head(segm_test)
nrow(segm_test)
nrow(segm_test)
names(segm_test)

# If more than one points in a segment -> deleting duplicate
segm_test2=segm_test %>% group_by(dn) %>%  summarize(classe=min(classe))
segm_test2=as.data.frame(segm_test2)
nrow(segm_test2)
table(segm_test2$classe)

#validation dataset
segm2=inner_join(segm,segm_test2,by= c("dn" = "dn"))
segm2$classe=as.integer(segm2$classe)
names(segm2)
segm2
file_out = paste0(path_out,"/segm_test_site1.shp")
polygon<-st_cast(segm2,"POLYGON")
st_write(polygon,file_out,delete_layer = TRUE)
names(segm2)

#-----------------------------------------------
#Supervised learning (2) and Validation (4)
#-----------------------------------------------
#train model (randomforest)
file_segm = paste0(path_out,"/segm_train_site1.shp")
file_model= paste0(path_out,"/model120.xml")
file_validate= paste0(path_out,"/segm_test_site1.shp")

features = "meanB1 meanB2 meanB3 meanB4 meanB5 stdB1 stdB2 stdB3 stdB4 stdB5"
cmdline = paste0(path_otb,'otbcli_TrainVectorClassifier ',
                 ' -io.vd ', file_segm,
                 ' -valid.vd ', file_validate,
                 ' -io.out ', file_model,
                 ' -feat ', features,
                 ' -cfield classe ',
                 ' -classifier rf ',
                 ' -classifier.rf.nbtrees 120',
                 ' -classifier.rf.max 10')
system(cmdline)
#-----------------------------------
#Mapping (3)
#-----------------------------------
file_segm = paste0(path_out,"/segm_5band_5cm.shp")
file_out = paste0(path_out,"/classif_objet_site1_model1.shp")
file_model= paste0(path_out,"/model120.xml")
```

```
cmdline = paste0(path_otb,'otbcli_VectorClassifier ',
                 ' -in ', file_segm,
                 ' -model ', file_model,
                 ' -cfield predicted ',
                 ' -feat ', features,
                 ' -out ', file_out,
                 ' -confmap 1')
system(cmdline)

#end model 1
#-------------------------------------------------------------------------
#model 4 training model 1,2,3 gathered

path_folder=paste0(path0,"/output_tfe","/model4")

if(!dir.exists(path_folder)){
  dir.create(path_folder)
}
segm1=paste0(path_out,"/segm_train_site1.shp")
segm2=paste0(path_out,"/site2_segm","/segm_train_site2.shp")
segm3=paste0(path_out,"/site3_segm","/segm_train_site3.shp")
seg1=st_read(segm1)
seg2=st_read(segm2)
seg3=st_read(segm3)

model4_train=rbind(seg1,seg2,seg3)
file_out=paste0(path_folder,"/model4_train.shp")
st_write(model4_train,file_out,delete_layer = TRUE)

#training and validation model 4
file_segm = paste0(path_folder,"/model4_train.shp")
file_model= paste0(path_folder,"/model4_120trees.xml")
file_validate= paste0(path_out,"/segm_test_site1.shp")
file_validate= paste0(path_out,"/site2_segm","/segm_test_site2.shp")
file_validate= paste0(path_out,"/site3_segm","/segm_test_site3.shp")

#robustness model 4
file_validate= paste0(path_out,"/site4_segm","/segm_test_site4.shp")

features = "meanB1 meanB2 meanB3 meanB4 meanB5 stdB1 stdB2 stdB3 stdB4 stdB5"
cmdline = paste0(path_otb,'otbcli_TrainVectorClassifier ',
                 ' -io.vd ', file_segm,
                 ' -valid.vd ', file_validate,
                 ' -io.out ', file_model,
                 ' -feat ', features,
                 ' -cfield classe ',
                 ' -classifier rf ',
                 ' -classifier.rf.nbtrees 120',
                 ' -classifier.rf.max 10')
system(cmdline)

file_model=paste0(path0,path_folder,"/model4_120trees.xml")

features = "meanB1 meanB2 meanB3 meanB4 meanB5 stdB1 stdB2 stdB3 stdB4 stdB5"
file_segm = paste0(path_out,"/site4_segm","/site4_segm_5band_5cm.shp")
file_segm= paste0(path_out,"/site3_segm","/site3_segm_5band_5cm.shp")
file_segm= paste0(path_out,"/site2_segm","/site2_segm_5band_5cm.shp")
file_segm= paste0(path_out,"/segm_5band_5cm.shp")
```

```
file_out = paste0(path_folder,"/site4_classif_objet_model4.shp")
file_out = paste0(path_folder,"/site3_classif_objet_model4.shp")
file_out = paste0(path_folder,"/site2_classif_objet_model4.shp")
file_out = paste0(path_folder,"/site1_classif_objet_model4.shp")

cmdline = paste0(path_otb,'otbcli_VectorClassifier ',
                 ' -in ', file_segm,
                 ' -model ', file_model,
                 ' -cfield predicted ',
                 ' -feat ', features,
                 ' -out ', file_out,
                 ' -confmap 1')
system(cmdline)

#-------------------------------------------------------------------------
#Robustness assessment model 3

path_model=paste0(path_out,"site3_segm")
file_segm = paste0(path_model,"/segm_train_site3.shp")
file_model= paste0(path_out,"/site3_model120trees.xml")
file_validate= paste0(path_out,"/segm_test_site4.shp")

features = "meanB1 meanB2 meanB3 meanB4 meanB5 stdB1 stdB2 stdB3 stdB4 stdB5"
cmdline = paste0(path_otb,'otbcli_TrainVectorClassifier ',
                 ' -io.vd ', file_segm,
                 ' -valid.vd ', file_validate,
                 ' -io.out ', file_model,
                 ' -feat ', features,
                 ' -cfield classe ',
                 ' -classifier rf ',
                 ' -classifier.rf.nbtrees 120',
                 ' -classifier.rf.max 10')
system(cmdline)

path_model=paste0(path0,"/output_tfe","/site3_segm")
file_model= paste0(path_model,"/site3_model120trees.xml")
file.exists(file_model)
features = "meanB1 meanB2 meanB3 meanB4 meanB5 stdB1 stdB2 stdB3 stdB4 stdB5"
file_segm = paste0(path_out,"/site4_segm_5band_5cm.shp")
file_out = paste0(path_out,"/site4_classif_objet_model3.shp")

cmdline = paste0(path_otb,'otbcli_VectorClassifier ',
                 ' -in ', file_segm,
                 ' -model ', file_model,
                 ' -cfield predicted ',
                 ' -feat ', features,
                 ' -out ', file_out,
                 ' -confmap 1')
system(cmdline)
#-------------------------------------------------------------------------
#Mean decrease Gini
library(randomForest)
library(caret)
segm1=paste0(path_out,"/segm_train_site1.shp")
segm2=paste0(path_out,"/site2_segm","/segm_train_site2.shp")
segm3=paste0(path_out,"/site3_segm","/segm_train_site3.shp")
seg1=st_read(segm1)
```

```
seg1=st_drop_geometry(seg1)
seg1<-as.data.frame(seg1)
site1 = dplyr::select(seg1,c(meanB1,meanB2,meanB3,meanB4,meanB5,
                             stdB1,stdB2,stdB3,stdB4,stdB5,classe))
site1$classe=as.factor(site1$classe)

seg2=st_read(segm2)
seg2=st_drop_geometry(seg2)
seg2<-as.data.frame(seg2)
site2 = dplyr::select(seg2,c(meanB1,meanB2,meanB3,meanB4,meanB5,
                             stdB1,stdB2,stdB3,stdB4,stdB5,classe))
site2$classe=as.factor(site2$classe)

seg3=st_read(segm3)
seg3=st_drop_geometry(seg3)
seg3<-as.data.frame(seg3)
site3 = dplyr::select(seg3,c(meanB1,meanB2,meanB3,meanB4,meanB5,
                             stdB1,stdB2,stdB3,stdB4,stdB5,classe))
site3$classe=as.factor(site3$classe)

file_out=paste0(path_folder,"/model4_train.shp")
seg4=st_read(file_out)
seg4=st_drop_geometry(seg4)
seg4<-as.data.frame(seg4)
sum_site= dplyr::select(seg4,c(meanB1,meanB2,meanB3,meanB4,meanB5,
                               stdB1,stdB2,stdB3,stdB4,stdB5,classe))
sum_site$classe=as.factor(sum_site$classe)


set.seed(2967)
#training model
model1<-randomForest(formula=classe ~ ., maxnodes=10,nodesize=10 ,
                     data=site1, ntree=120)
model2<-randomForest(formula=classe ~ ., maxnodes=10,nodesize=10 ,
                     data=site2, ntree=120)
model3<-randomForest(formula=classe ~ ., maxnodes=10,nodesize=10 ,
                     data=site3, ntree=120)
model4<-randomForest(formula=classe ~ ., maxnodes=10,nodesize=10 ,
                     data=sum_site, ntree=120)

#Gini Plot
varImpPlot(model1)
varImpPlot(model2)
varImpPlot(model3)
varImpPlot(model4)
```