MACHINE LEARNING BASED CLASSIFICATION OF EARLY SERAL VEGETATION IN CUT-BLOCKS IN THE INTERIOR OF NORTHERN BRITISH COLUMBIA

by

Matt McLean

BSc, University of Northern British Columbia, 2017

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
NATURAL RESOURCES AND ENVIRONMENTAL STUDIES

UNIVERSITY OF NORTHERN BRITISH COLUMBIA

December 2024

Abstract

Globally forests provide a wide range of essential services such as lumber for construction, tourism value, and habitat for animals. In many regions forest management is performed to maximize the utilization of these services and to promote sustainable forest ecosystems. Effective management requires detailed information on the current state of forests, how the forest is projected to develop through time, and knowledge about the provisioning of desired forest services, such as forage for wildlife species. Historically this information has been acquired using traditional field surveys, which is both costly and limited in the extent of area that can be sampled. The use of Remotely Piloted Aircraft Systems (RPAS) combined with machine learning potentially allows for more scalable methods of gathering information on forest inventories. In this thesis, I evaluate and advance the use of multispectral imagery collected from RPAS for the classification of early seral vegetation. This specific type of vegetation is both a key indicator of forest regeneration and habitat suitability for ungulates. However, accurate identification and classification of early seral vegetation is particularly challenging due to its small size, the fact that individuals are highly variable, and the fact that individuals can overlap and not exhibit distinct boundaries.

The process of image classification is broken down into two major components: the segmentation of collected imagery into discrete units of vegetation and then the classification of those units into their specific species. These two components are presented as an overall framework for classification. I also provide operational recommendations to achieve successful results.

The algorithms used in the segmentation of images are highly configurable and can be tuned to the input data to yield high quality results; however, what is more challenging is

determining what a high-quality result is, and applying suitable metrics that allow the accuracy of the segmentation process to be evaluated. In this research I propose a method for scoring the quality of segmentation quality applied to forest imagery, in a format that can be easily integrated into a larger framework that will integrate with the classification of results.

In the second component of my thesis, I evaluate various common classification algorithms and assessed their accuracy. This analysis considered both overall accuracy of classification, as well as only the classification accuracy of species of interest. I also explore under what circumstances this type of classification be feasible and provide recommendations on what variables are most important to control during the collection of training data, and best practice for capture of new datasets for classification with already trained models.

My research demonstrates both the benefits and limitations of using RPAS imagery for segmentation and classification of early seral vegetation and suggests best practices that can be used when applying this framework.

Table of contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Globally forests provide a range of essential ecosystem services that support communities, provide a wide range of regulating, provisioning and support services (Baskent et al., 2020; Taye et al., 2021), and are the basis of many economic sectors (Costanza et al., 1997; Millennium Ecosystem Assessment, 2005). In western Canada forest ecosystems are the providers of the timber supply that forms the economic foundation for many communities. Forest also help maintain the quality of water (Pearce, 2001) and air (Nowak et al., 2014) necessary for our survival. As time progresses there may be more significance placed on a variety of factors, such as wildlife habitat (Oettel and Lapin, 2021).

In most regions of the world forests are explicitly managed by forest professionals with the aim of promoting sustainable forest ecosystems and often specific forest attributes and products, such as timber production (Boukherroub et al., 2017) or carbon storage (Lemprière et al., 2013). Forest management involves not only determining how forest should be harvested and regenerated, but also developing strategies that inform which forest or stands should be managed, when they should be managed, and what is the most appropriate management strategy (D'Amato et al., 2011). Throughout time there have been a variety of forest management goals (McGrath et al., 2015), however a constant has been a desire to maximize the utility of available forests, regardless of the desired utility at the time.

As forest managers seek to utilize better decision processes, having access to the most accurate and detailed information about the environment is a necessity (Tompalski et al., 2015;

van Leeuwen et al., 2011). Understanding how well forests are regenerating requires knowing how many trees are present of each species, and their regrowth. As forests provide a variety of ecosystem services, they also have a variety of information that can help with determining these values. Ranging from elements such has how many of which trees are present, to more complicated metrics where regrowth rate predictions can be determined based on density of trees and other factors. The more fine-grained knowledge of the forest inventory allows for more precise management plans to maximize forest values.

Historically, information on forest ecosystem condition and forest structure was obtained by having field crews conduct field surveys. These types of surveys can be very expensive to conduct, which thus limits the amount of knowledge available for making decisions. Historically the primary methods for obtaining tree species data includes ground survey-based identification of tree species, which is costly, time-consuming, and typically samples only a small percentage of the land base and extrapolates across the full area(British Columbia et al., 2007). Cost reduction has been achieved through manually interpreted air photos but at the cost of species accuracy ("Forest Health Aerial Survey Manual," 2012; Seely, 1934); however, this involves a technologically complicated and expensive process.

Traditional field surveys are limited in effectiveness due to both issues of scale and quality of data. The scale of data that can be collected is limited by the high costs of obtaining field data. Traditional plot sampling methods utilize relatively low sampling intensity combined with extrapolating the results across a disproportionately broad land base. Additionally, this extrapolation also introduces a degree of surveyor bias, which typically includes both subjective

personal interpretations of observations, as well as design bias resulting from avoiding difficult to access/remote areas vs an unbiased grid-based design over the entire population.

1.2 Technical Background

Remote sensing technologies provide opportunities to survey forest blocks in their entirety as opposed to averaging sample plots against the land base (British Columbia et al., 2007). Knowing the vegetation species and quantifying its abundance helps to inform various types of models enabling us to better understand how to manage our land base. Particular to this research there is a goal of developing data that can be used in wildlife models focused on food availability and wildlife cover for ungulates (Brown et al., 2007; Terry et al., 2000; Whitman et al., 2017). This research presents additional value in providing detailed information on the state of regrowth in the cut block (Pitt et al., 2010; Weisberg and Bugmann, 2003).

The ability to acquire higher resolution imagery in terms of both spatial resolution and spectral resolution opens new possibilities for observing the environment around us. Different methods of data collection allow for different questions to be answered based on what data is produced. As an example, one of the earliest forms of remote sensing is aerial photography, which can cover moderately large areas, with relatively high resolution. Air photos are most commonly available as panchromatic or true colour. Satellites can cover much larger areas, and often include multispectral data that is useful for environmental modeling, however this comes at the cost of spatial resolution making identification of smaller objects not possible. Recently we have seen a growth in the use of Remotely Piloted Aircraft Systems (RPAS) commonly known as drones. RPAS while limited in the amount of area that can be surveyed compared to a manned

aircraft are able to capture even higher spatial resolutions than traditional aerial photography, as well as some systems still providing a high spectral resolution.

LiDAR (Light Detection and Ranging) is a very useful tool in forest classification (Brandtberg, 2007; Coops et al., 2016), as it provides insight into the structure of the tree canopy as opposed to viewing only the surface reflectance of the vegetation present. However, this system may begin to have trouble where unique structures cannot be determined. This can be especially true of early seral vegetation, where the individual trees may be too small to be viewing any form of structure in the trees due to lack of point density. Additionally, by using multispectral imagery for classification, there is the potential for future work leveraging the same raw data for forest health assessments in addition to classification.

Using RPAS provides opportunities to examine the entire land base of interest allowing for a more thorough sampling. RPAS may carry a variety of sensors; some of these such as the MicaSense RedEdge-M used in this research are capable of sensing spectrums of light that are not visible to the human eye. The addition of nonvisible light spectrums adds more ability to discriminate species than with standard colour air photos intended primarily for human sight. Finally, due to the level of interpretation needed there is a potential for different surveyors to produce differences in their final report; adding a system of automation that can be executed across all sites can help to reduce operator bias.

Reducing survey costs would allow land managers to collect more data allowing more refined management strategies. Machine learning (ML) represents a potential avenue for increasing the completeness and repetition of such surveys. Developing robust methods for

colleting remote sensing data is essential to the reliable implementation of machine learning for classifications.

The first step in the process for using machine learning for classification is the collection of data. This can be split into two separate but related problems; what data to collect, and how to collect it. In terms of what data to collect we have a variety of methods at our disposal, ranging from photos, LiDAR, and Radar. Further expanded, each of these methods have multiple resolutions and modes available. Using the example of photos, it is important to select appropriate resolutions, and spectral ranges; then selecting the appropriate capture method such as satellite, manned aircraft or RPAS. Another crucial decision involves determining appropriate times to collect the data. Some of these factors might relate to weather, sunlight, seasons, or phenological cycle of vegetation imaged.

Once data is captured it is then processed into usable products. For RPAS data this generally involves the use of photogrammetry to produce orthomosaics covering the entire capture area in a single image. It is at this stage that we may also apply radiometric calibrations, for RPAS data this is generally done by using a combination of images of panels with known reflectance, and sunlight sensors mounted on the drone (as is the case for the MicaSense cameras used in this research). Satellite data generally would undergo atmospheric correction, and other types of data would have other standard processes to prepare data for analysis.

Following this, the machine learning pipeline will convert this standardized data into a suitable format for use within the algorithm. An example of this would be many of the algorithms in this research require that the images as a grid of pixels be converted into an

average reflectance for all pixels contained within an object. Once data is in the proper format, machine learning is then finally used to determine a class of the input data.

Combining all these processing steps from data capture to classified data comprise of the analysis framework. While each step of the framework can be relatively simple to understand, they all work together, to produce the final outputs, where the quality of the final product is dependant on how robust the framework is and how well the various stages work together. A correct classification is dependant upon proper data collection, which is processed in such a way that accurate data is going into the classification step.

Despite recent advancements in RPAS technology segmentation and classification still rely on surface reflectance as opposed to structural attributes. Leveraging differences in reflected light spectrums has been a staple of machine learning classifications in other fields; a trivial example of this is for self driving cars there is a different meaning to white and yellow lines on the road, even if they have a similar structural appearance. This can then be taken furth into using multispectral sensors to gain more accurate identifications in changing conditions (Takumi et al., 2017).

However, in the field of forestry and environmental management more generally, using surface reflectance to identify features can be problematic as reflectance can change through the phenological cycle of trees. This can be made more challenging when looking at deciduous vegetation which tends to have less defined boundaries than conifer species. Additionally young trees being smaller provide a greater challenge in identification given the resolution of the data. This is further complicated by the natural variance between trees of a given species; and that

their presentation will also be affected by presence of disease, and differences in environmental factors such as access to water and nutrients.

This research focuses specifically on early seral vegetation, which is and important food source for ungulates. Early seral is the first stage of vegetation in a forest lifecycle as it regenerates from cut to old growth. Characteristics of this vegetation are that it is small and may be less rigid or well defined than mature forest; studying this stage of forest growth presents some unique challenges to forests in general. While a more mature forest may have measurements of the structure from LiDAR (Holmgren et al., 2008), the small size of early seral vegetation can mean it has little difference in height between trees and the understory. Additionally, at this stage very high spatial resolutions are required to have opportunities to delineate edges of tree stems due to small sizes and less definition of edges than more mature forest. Furthermore, the species of particular interest in this research are deciduous where branches tend to overlap with neighboring trees further reducing the ability to cleanly delineate tree edges.

This research presents methods for the collection and processing of RPAS multispectral imagery and applying and evaluating machine learning based classification of image segments representing early seral vegetation. The goal is to present methods that are abstractable; as imaging technology improves both in terms of spatial and spectral resolution, and as new machine learning algorithms are developed, these methods could be repeated, and evaluations be usable as a comparison of the potential improvements of new technologies. The research presented shows results for a specific type of imagery and a specific set of algorithms. It should not be interpreted as the optimal solution for early seral vegetation classification, but rather a proof of concept, and a starting point for continuous evaluation of technological possibilities.

Image segmentation is the process of splitting a whole image into discrete units known as segments. These segments can then be passed into other algorithms for further processing. While segmentation is a generally well researched field, this research looks to apply the techniques to the specific case of early seral vegetation. There are a wide variety of algorithms already developed e.g. Random [decision] Forests (Tin Kam Ho, 1995) or Support Vector Machines [Support-vector Networks] (Cortes and Vapnik, 1995), however not all approaches are equal, and different approaches work better for different problems. This research seeks to explore how some of the common approaches can be used to segment early seral vegetation.

Quality analysis of segmentations is inherently difficult, as in almost all cases a decision about what is a correct segmentation needs to be determined. For the purposes of this research, human drawn polygons are taken to be correct. There are three primary errors that are faced with this form of truth; first humans may make mistakes in drawing the boundaries, this is true of any human derived dataset. second is the risk of oversimplification, as there is a tendency for humans to produce over simplified polygons that either over or under encapsulate the vegetation to produce polygons with fewer vertices. The final source of error faced is the impure pixels, in the context of this research project the multispectral data is made of pixels that represent a 3cm-6cm square on the ground; a problem may arise at the edges of the trees where a pixel will cover a percentage of the tree, and this poses another question of minutia; should these impure pixels be included, excluded, or a percentage of them included?

1.3 Project Roadmap

At a high level this research seeks to provide a process for taking data collected from RPAS, along with ground truth data and produce models that show the amount of coverage of each species of vegetation captured in the data; along with an analysis of the accuracy of the data

produced. This is accomplished using a modular approach to the process of identifying early seral vegetation, such that any piece should be able to be replaced without altering other pieces of the process; thus, providing a path for natural evolution and continual growth. While this research was conducted specifically with a MicaSense RedEdge-M camera, the processing pipeline presented in this paper could work just as well with another camera system. Likewise, more machine learning algorithms could be added, or different scoring metrics used. As the processing pipeline is in place to take images with truth segments and output trained models with their achieved accuracies; determining if a change to the process is beneficial becomes trivial.

There are two options for running the framework for early seral vegetation classification demonstrated in this research (Figure 1); the first is the Training Process where inputs of Imagery and Ground Truth Data are provided. This option can be used any time new training data is collected allowing for accuracy to be improved as training sets grow; or alternatively additional algorithms could be added to the test. An important point here is that thanks to the automatic storing this can be collected as a single monolithic script that while it would take days to even weeks to run; minimal human effort could allow for new optimization. Once a model has been trained it can be applied very quickly needing only segmentation and classification steps to be performed. Again, there would be opportunities here to automate this process such that when a new segmentation or classification algorithm is identified existing imagery could be re-processed ideally providing enhanced results.

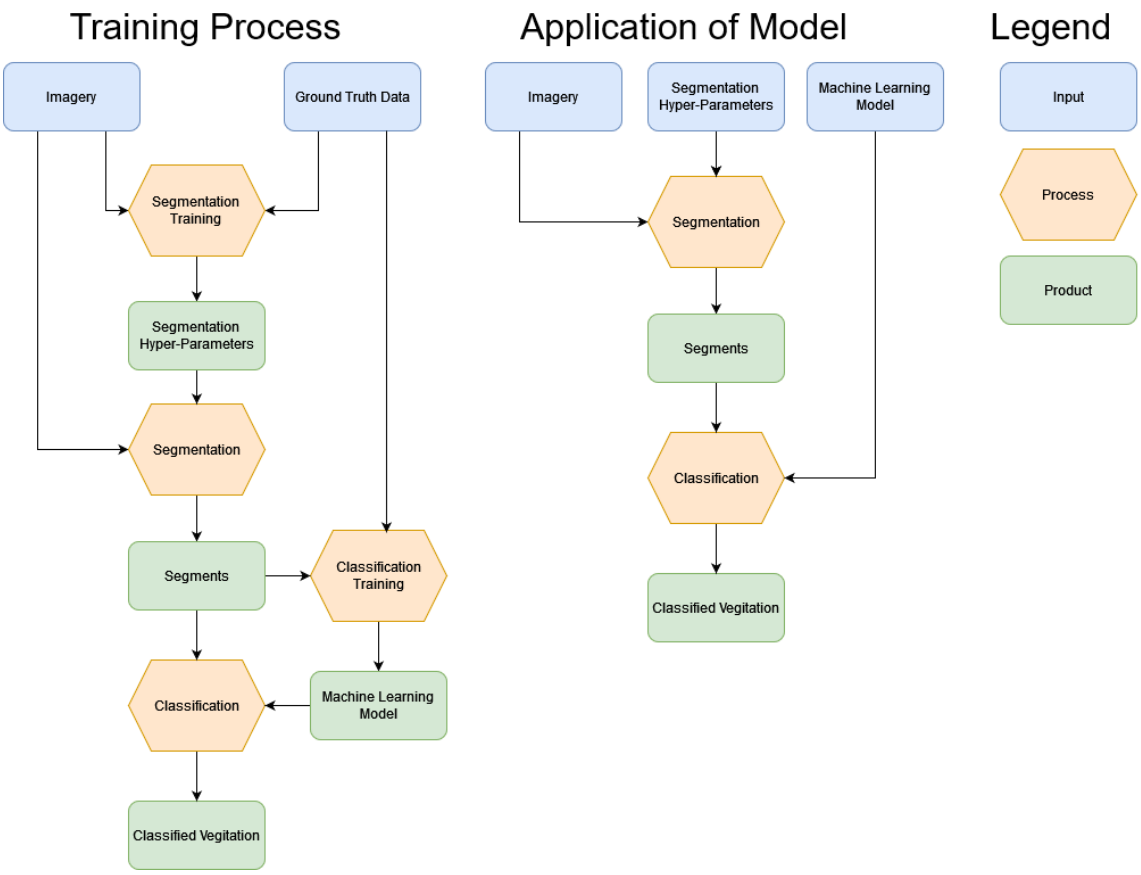


Figure 1 | ER Diagram of processing pipeline for both initial training, as well as applying trained models to novel data.

Chapter 2 defines a framework for segmentation of early seral vegetation from RPAS Multispectral imagery. This will be done by establishing a method for evaluating the effectiveness of segmentations, and then segmenting the imagery for all the sample sites using standard segmentation algorithms. Then using the scoring techniques identify the most effective algorithm along with the parameters that can be used to produce the most effective results.

Classification of imagery with RPAS based systems provides the flexibility to capture data as needed at moderate scales as well as providing various options regarding the type of imagery collected. In this work we examine specifically the ability to train models to identify tree species using the spectral information collected by a MicaSense RedEdge-M camera system.

This camera provides not just standard colour images, but also add Near-Infrared and RedEdge bands; as these bands are known to be useful in the analysis of vegetation (Schuster et al., 2012).

Chapter 3 will build upon the results of Chapter 2, using the best segments produced as the basis for classification. In this chapter several common algorithms will be optimized, scored for accuracy, and compared for accuracy. A proposed framework for classification will be presented based on the combination of algorithms from chapter 2 and 3, along with discussion of the usability of these results.

Chapter 2

Image Segmentation for Early-Stage Vegetation in RPAS Imagery

2.1 Introduction

Effective forest management requires widespread, yet detailed information of forest composition, to select the management techniques that will achieve the wide-ranging objectives of resource stewardship. This balancing of competing objectives is often addressed at the landscape level, where regions of forest are assigned, a class based upon species composition (Baleshta et al., 2015; Dhar, 2013), age of trees in the stands (Zheng et al., 2007), and the structural attributes of the forest (Bugmann et al., 1996).

Forest composition can be segmented at a variety of scales ranging from landscape scale determining surface cover types (Walsh, 1980; Wulder, 2003), to surveys of individual sites, down to segmenting individual trees to build more detailed forest models. The use of RPAS systems combined with modern image analysis techniques allows for capturing data fine scale enough to separate individual trees from the forest; while also providing an efficiency required to survey large areas that would be cost probative with traditional survey techniques. The ability to segment all trees as individuals can then be used to produce highly accurate models of ecosystem dynamics(Seidl et al., 2012).

The objective of image segmentation is to inspect an image and return boundaries delineating individual features from an image. An analogue to the research presented in this paper is Optical Character Recognition (OCR) (Gupta and Nair, 2005); a process where each letter is first segmented from the image for identification, as well as words are segmented from groups of symbols. Segmentation is also used in a variety of computer vision tasks such as recognizing licence plates, or barcodes.

Effective segmentation has a variety of challenges that must be addressed to ensure optimal accuracy. Ideally a system for segmentation will be able to identify features based upon similar objects, yet not identical; this requires the collection of very large datasets of tagged training data. This process can be made more challenging when tagging of training requires expert intervention; in comparison to a problem such as OCR training data which could be collected from anyone with the ability to read, an example of this is reCAPTCHA used on websites (Pettis, 2023). Additionally, segmentation relies upon the ability to separate objects from the background. Based upon the size and types of objects to be segmented along with the backgrounds upon which they are placed there is a variety of algorithms that may present different levels of success based on a given dataset.

This research seeks to delineate individual stems (trees) within early seral vegetation. The identification of this vegetation is useful for managing forest regrowth; after cutting, forest fires or other disturbances displacing mature stands. Early seral vegetation is important as it provides sources of food and camouflage for animals within the stand and is a meaningful metric for habitat suitability. Further this vegetation will grow, and an accurate inventory can help to provide the base information needed for predictive forest models.

Early seral vegetation also provides some distinct challenges for segmentation. Segmentation is generally less complex when uniform boundaries are present, however in the case of vegetation the structure of branches and leaves leads to a broken silhouette without clearly defined boundaries. As tree crowns grow and increase in size, the crowns of individual trees will begin to overlap, their boundaries become obfuscated or ambiguous. The other primary challenge faced is due to the scale; more mature stands will have taller trees with more defined shapes, and gaps between them, allowing for segmentation using spatial information in addition

to spectral information (Yancho et al., 2019). The small size of early seral vegetation necessitates the use of higher resolution data, but now the spatial component is less useable due to the granular blending of crown edges and the understory or forest floor.

In this chapter I demonstrate scoring metrics that can be used to evaluate the quality of returned segments. The scoring metrics presented will be used to evaluate the effectiveness of four different image segmentation algorithms. These results are used to determine which algorithm is the most effective for the given dataset; in this case early seral vegetation captured using RPAS. While there are many more than four segmentation algorithms available, this provides a starting point for developing the methodology. The results for this section will include results for both the accuracy of segmentation; as well as how stable this accuracy is from site to site, an important factor for a generalized solution that could be implemented at scale.

2.2 Study sites

The data used in this research was captured from a selection of sites both in the Ungulate Winter Range north-west of Mackenzie, as well as some sites closer to Prince George that provide easy access. The sites have all been logged approximately five years prior to data collection and are being monitored for their quality of Ungulate habitat as they regenerate.

The sites contain a variety of young vegetation; of particular interest to this research are the deciduous species that will act as a food source and provide cover for ungulates in the area. Five target vegetation species have been identified as valuable for moose browse by the British Columbia Ministry of Forests: Trembling Aspen (*Populus tremuloides*), Red-osier Dogwood (*Cornus stolonifera*), Paper Birch (*Betula papyrifera*), Highbush-Cranberry (*Viburnum edule*), and Willow (*Salix spp.*).

Eleven case study sites were analyzed; these sites were situated in the central interior of B.C. and were in the Sub-boreal Spruce (SBS) biogeoclimatic zone (Beaudry et al., 1999).

Table 1 | Study site data collection statistics with tree counts and capture date.

Site \ Attribute	Total Samples	Target Samples	Target %	Capture Date
200rd	11529	2031	18%	August 24th, 2020
700rd	3774	1977	52%	September 29th, 2020
Alezza	5649	2064	37%	September 25th, 2020
Bend05km	1521	636	42%	September 21st, 2020
ChiefLake	12861	1491	12%	August 24th, 2020
ConifexH47	1398	261	19%	September 2nd, 2020
ConifexK14	1077	513	48%	September 3rd, 2020
NorthFraser11	8331	3453	42%	September 22nd, 2020
NorthFraser41	1677	270	16%	September 14th, 2020
NorthFraser50	1107	468	42%	September 21st, 2020
Olson5km	2862	567	20%	September 22nd, 2020

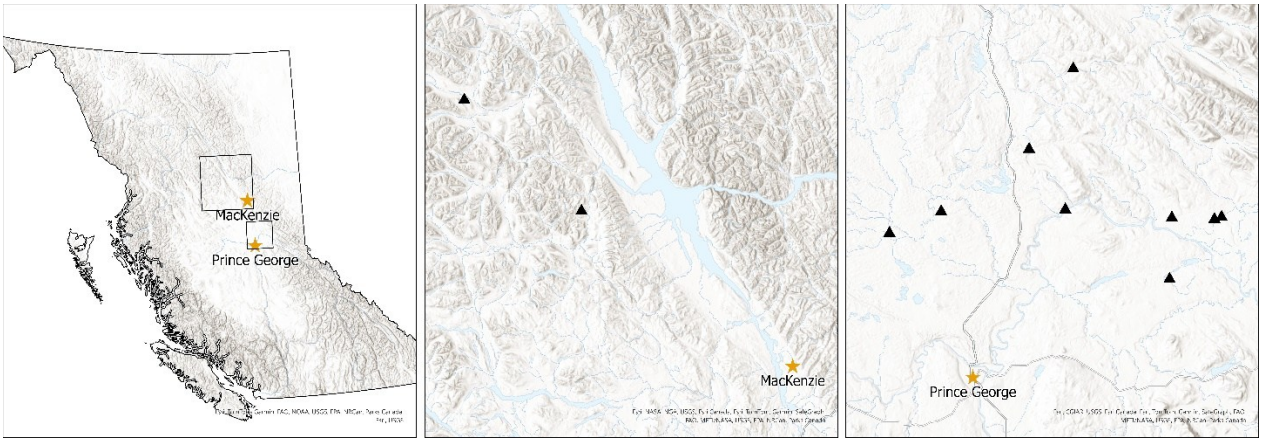


Figure 2 | Map of locations where data was collected.

Table 2 | Extended site attributes; with information from British Columbia Vegetation Resource Index (Government of British Columbia, n.d.)

Site \ Attribute	Leading Species	Planted	Brushed	Previous Dominant	BEC Zone	BEC Subzone
200rd	Pl (33%)	May 2016	NA	PLI	SBS	mk
700rd	Cs (32%)	July 2017	August 2014	SW	SBS	wk
Alezza	Sx (42%)	June 2019	NA	SX	SBS	wk
Bend05km	Sx (50%)	July 2013	August 2015	BL	SBS	vk
ChiefLake	Pl (23%)	June 2013	August 2018	--	SBS	mk
ConifexH47	Li (23%)	July 2017	NA[1]	SX	ESSF	mv
ConifexK14	Bl (42%)	August 2017	NA[2]	PLI	BWBS	dk
NorthFraser11	Sx (42%)	June 2008	August 2010	PLI	SBS	mk
NorthFraser41	Ri (21%)	July 2017	August 2018	SX	SBS	wk
NorthFraser50	Sx (44%)	July 2016	August 2015	BL	SBS	wk
Olson5km	Bl (32%)	June 2015	NA	PLI	SBS	mk

2.3 Data Collection

Aerial image data were collected by flying survey missions with a DJI Matrice 210 RPAS, and a payload of a MicaSense RedEdge-M multispectral camera. The collected imagery was then processed using Agisoft Metashape 1.5 processed on high quality for all stages. First calibrating the sensor against its Downwelling Light Sensor (DLS), matching tie points, then creating a dense point cloud. This dense cloud then has a ground filter applied, and a Digital Terrain Model DTM is produced using only ground points. Finally, an orthomosaic is produced by ortho-correcting and mosaicking the images collected and is then exported as a GeoTIFF for the segmentation process.

[1] Records indicate most recent brushing completed after imagery collected, unknown previous brushing date.
[2] Records indicate most recent brushing completed after imagery collected, unknown previous brushing date.

In addition to the imagery a corresponding set of ground truth data was also collected. This dataset was generated by field crews marking in-person species assessment on an orthophoto. Once back in the office these spatial delineations were converted to a digital format using GIS software to draw polygons around each feature and adding a species code as an attribute to the polygon. It is recognized that this process of digitizing is not a pixel perfect representation as it occasionally excludes the tip of a branch or includes a small piece of ground. Later, subsequent use of this data will be evaluated on a per-pixel basis thereby adding an element of error to the results. However, classifying at an object level, and analyzing at a pixel level is believed to have minimal impacts, and represents errors that would also occur with existing fully manual methods of tree segmentation.

To prepare the data for machine learning processing truth segments needed to be created, to represent the trees to be classified. After the drone imagery was collected, and processed; these orthomosaics were then imported into GIS software, where the crowns were manually delineated as polygons, including an attribute for species code (Figure 3).

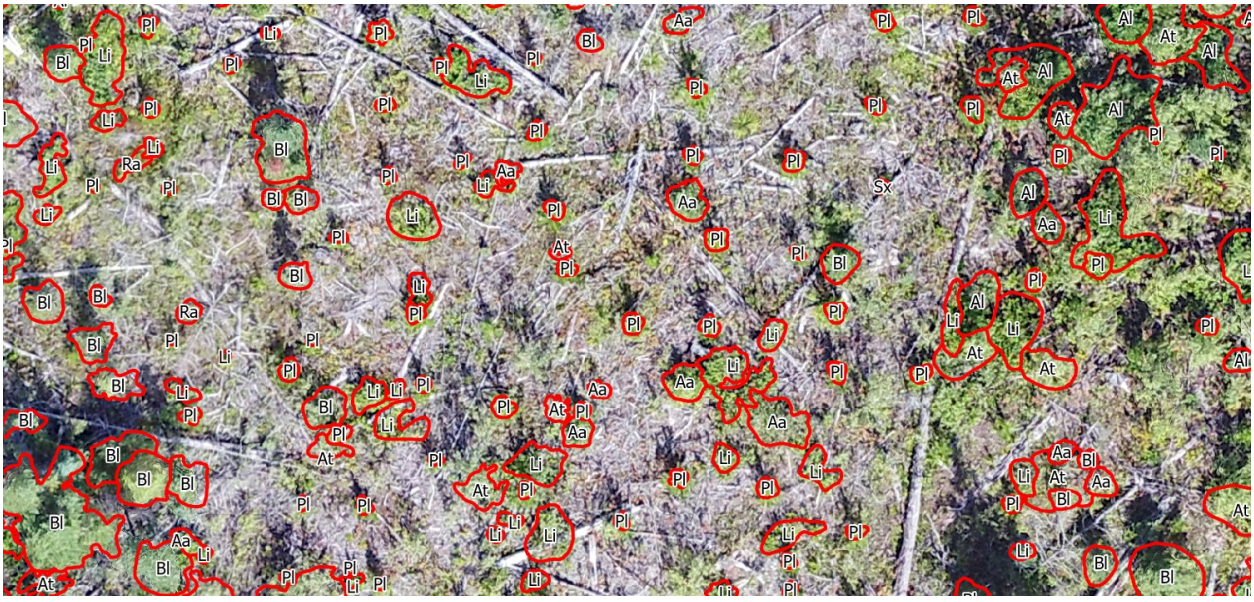


Figure 3 | Example of manually delineated crowns in red and corresponding species code labeling.

2.4 Segmentation Algorithms

The algorithms used for image segmentation come from SciKit-learn and SciKit-Image. The motivation for this choice is based upon several factors including the relative popularity of SciKit, and the extensive documentation and community that comes with that widespread use. The open-source nature of the code makes it an ideal foundation to build upon, allowing for others to continue future research without needing to worry about licencing costs. Finally, SciKit provides many algorithms that utilize standardized API's which allows an easy path to develop modular code where algorithms can be easily compared for the specific datasets being used. Finally, SciKit has support for PyTorch-CUDA, an industry standard library for GPU accelerated Machine Learning. All of the algorithms selected focus on spectral reflectance as opposed to those that look for structure such as the watershed algorithm on DEM's as the trees examined were too small to be properly represented in surface models.

2.4.1 SLIC

Simple Linear Iterative Clustering (SLIC) (Achanta et al., 2010) is a method for creating Super pixels, a computer vision term which is closely analogous to clustering in remote sensing. Super pixels can be thought of simply as a group of pixels sharing similar characteristics; these super pixels then have a border drawn around them, and this border is the segment. This method is based upon K-Means clustering (Pollard, 1982), and is used with a number of starting seeds, randomly placed though somewhat uniformly placed pixels, that are then grown by including pixels in the neighboring cluster that has the most similar properties to the given pixel. In the implementation of this algorithm in SciKit-Image the two primary parameters we can pragmatically test across are n_segments, and compactness.

N_segments is the estimated number of segments we expect the algorithm to output, this variable can be informed by knowing how large objects are in relation to the size of image. And compactness refers to how smooth we want individual objects to be, where maximum compactness would be squares (not circles as each segment must touch other segments on all sides and this algorithm is seeding from a grid). Lowering the compactness allows segments to contour to the objects being detected; however, this can also lead to more ambiguity as less significant features may be considered as edges.

2.4.2 Quick Shift

Quick Shift (Vedaldi and Soatto, 2008) is a clustering algorithm (much like SLIC above) that works by calculating the average of clusters. However, due to the computational time required Vedaldi and Soatto proposed the addition of a Quick Shift that reduces the computational complexity of the algorithm.

Quick Shift uses slightly different hyperparameters than before, where sigma, sometimes referred to as kernel size has a similar impact as N_segments had in SLIC. And the ratio hyperparameter has an effect like compactness, adjusting the ratio of importance for changes in colors vs changes in position within the image.

2.4.3 Felzenszwalb's Efficient Graph (F-Graph)

Felzenszwalb's Efficient Graph (Felzenszwalb and Huttenlocher, 2004), works by looking at the image both in terms of regions and neighboring pixels. By examining the difference in intensities of regions we can find an estimate of where segments should exist in the image. The next step is to look at neighboring pixels to find where the intensity changes to determine the precise edge of the segments.

2.4.4 Mean Shift

Mean shift (Comaniciu and Meer, 2002) is a clustering algorithm that like SLIC is a mean based clustering scheme. Where SLIC uses parameters to constrain the clusters and thus the computational time. Mean Shift has a single hyperparameter for bandwidth and was calculated at runtime by the SciKit library. This algorithm has a much higher computation time than the other algorithms, however, it may have benefits where prior testing for hyper parameters is not possible.

2.4.5 YOLO

The final algorithm that was examined but was ultimately determined to be unsuitable at this stage is YOLOv3 (Redmon and Farhadi, 2018). YOLO works on a process of object identification as opposed to segmentation, and as a result the boundaries of detected objects are unclear as they are defined in terms of a bounding box as opposed to tracing the object.

2.5 Method of Evaluation

To evaluate the effectiveness of the algorithms a training mask is used to compare their outputs. The output of the segmentation algorithms is a bitmap (image) where each pixel's value is linked to a unique id for a given segment. The training mask is derived from the ground truth data collected and has identical resolution to the output of the segmentation algorithms. The training mask can later be used as a template for scoring algorithms. An important note here is that the classification and training mask must be of identical position and number of pixels. These images are then stacked, and scoring is based upon the relative similarity of the images. By relative similarity it is meant that pixels will not have identical values, however it should be possible to map them. That is if a pixel in the segmentation has value α, and if the corresponding pixel in the training mask has value β, then for every pixel in the segmentation with value α, the corresponding pixel should also be β. For cases where only some pixels with value α map to β we know an error has been made; likewise, when multiple values from the segmentation mask map to the same value in the training mask an error has been made.

Figure 4 | Example of perfect over-segmentation; each value in right table is mapped to only one value in left table.

The segmentation methods are applied using the Python frameworks Sci-Kit and ImageAI. These frameworks provide the base implementation of the algorithms used and are designed to accept hyper-parameters that can be used to fine tune the algorithms. The advantage of using such frameworks is that they provide for rapid development, unified API's making it easy to consistently change between algorithms and are optimized code allowing for computation to be completed in a reasonable amount of time.

2.5.1 Base-Metrics

The metrics listed below represent various methods of comparing the segmentation to training mask, each placing value on different types of errors. Used in combination these metrics put more weight towards errors that would be more detrimental to the results.

A False merge (FM; Figure 4) is the case where a segment encompasses more than a single object. For example, a segmentation that includes multiple trees, or the ground surrounding the tree. False mergers are a measure of entropy associated with segmentation and are a representation of over-segmentation (many objects per segment). For this metric a lower score is better, with 0 implying that there is no over-segmentation present which is to say there is never a segment including more than one tree. (Meilă, 2007)

False splits (FS; Figure 4) occur when a single object is represented by multiple segments. False merges are a measure of entropy associated with the segmentation and are a representation of under segmentation (many segments per object). For this metric a lower score is better, with 0 implying that there is no under-segmentation present. (Meilă, 2007)

As a note if both False Merge and False Splits were 0 that would suggest that the output of the segmentation was a pixel perfect representation of the training mask (Figure 5).

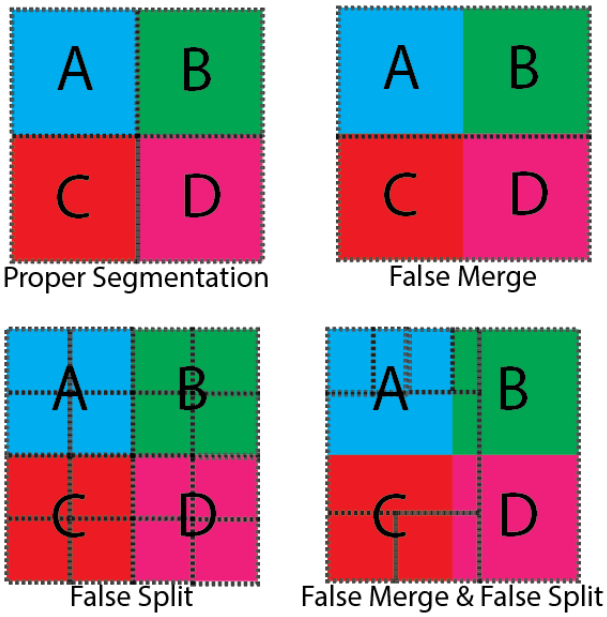


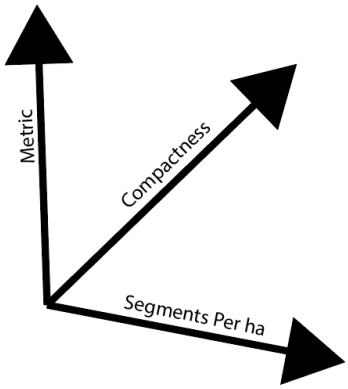
Figure 5 | Examples of Segmenations on data; colors represent distinct data, dashed lines determined segments

Adapted Random Precision (ARP) is the probability that a pixel of a given class in the results is the same class in the truthing data, and normalized over the total number of pixels in the classified data, in simple terms this can be thought of as the percentage of pixels correctly segmented. (Arganda-Carreras et al., 2015) Higher values are better for this base-metric. This can be thought of as like false merges, in that if a segment is too large it will lower the ARP, by a proportion of the area of over segmentation as opposed to the quantity of segments.

Adapted Random Recall (ARR) is the probability that a pixel of a given class in the results is the same class in the truthing data, and normalized over the number of pixels in the truthing data. (Arganda-Carreras et al., 2015) Higher values are better for this base-metric. As ARP was to FM, ARR is to False Splits, again looking at the number of pixels that have been placed into alternate segments as opposed to the number of additional segments produced.

ARP and ARR are closely related statistics and thus it may be useful to take their average. Adapted random error is defined by the equation. $ARE = 1 - \frac{2(ARP*ARR)}{ARP+ARR}$, (Arganda-Carreras et al., 2015) and in this case, we seek a lower value, with 0 being the perfect classification. In the case of 0 all pixels can be directly mapped from the classified data to the training mask.

Table 3 | Figures in table below each demonstrate the scoring of a base metric as the SLIC hyper-parameters are changed.



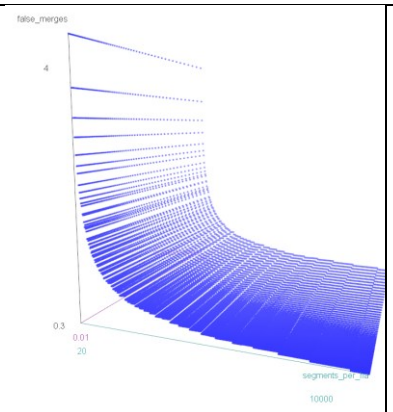


Figure 6 - SLIC False Merges, as segments per hectare increases false merges decreases

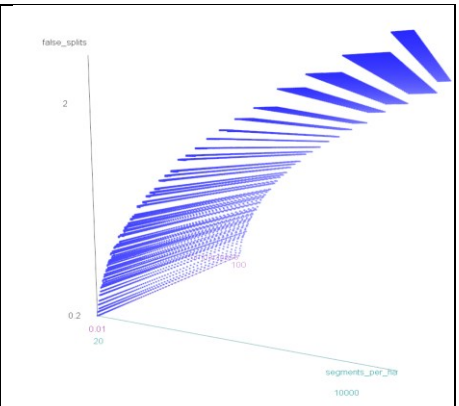


Figure 7 - SLIC False Splits, as segments per hectare increases False Splits increase

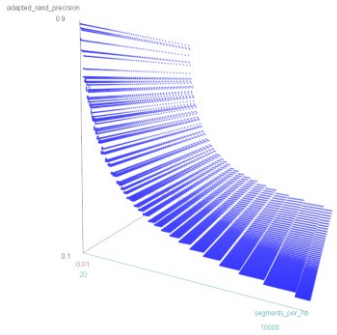


Figure 8 - SLIC Adapted Random Precision decreases with segments per hectare

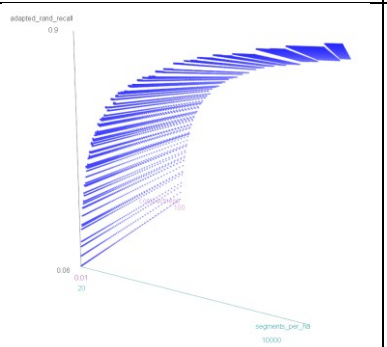


Figure 9 - SLIC Adapted Random Recall increases with segments per hectare

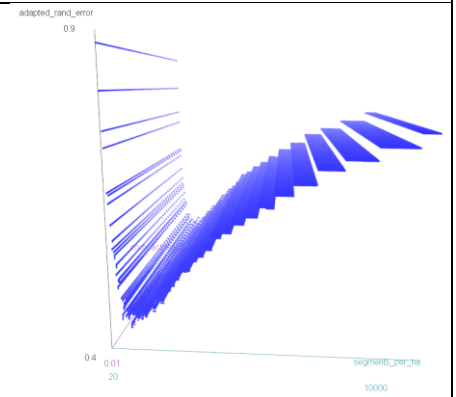


Figure 10 - Adapted Random Error, local minimum is achieved with segments per hectare. This is the only base metric that the optimal solution does not extend to limit

In the table above we see some graphs which show the reaction of the metrics against the hyper parameters used in the SLIC algorithm (discussed later). These graphs are useful for visualizing how False Merges and False Splits compete against each other, as do Adapted Random Precision and Recall. Adapted Random Error is the only metric that does not extend towards a limit as it is made up of two competing metrics. It is for this reason that these metrics must be defined to produce useful results.

2.5.2 Evaluation-Metrics

For this research, scoring methods that help to moderate the base-metrics were needed. At first glance it is easy to think that the goal is to min/max the metrics used, however these metrics taken to the extreme (perhaps except for Adapted Random Error); will not actually provide useful results. To illustrate this, consider false merges, if we simply make every pixel its own segment there will be no false merges; likewise making the entire image a single segment will present as no false splits. ARP and ARR also suffer from this however to a lesser extent as we recognize that the training data has multiple classes, as such simply saying the whole image is one segment means that some of the classes must be wrong; however, even here caution must be used in relation to unbalanced data sets. If our training data is for example 75% class α, then a single segment for the entire image could still represent a 0.75 ARR.

Five base-metrics were utilized to evaluate the quality of the segmentation algorithms and the segments they produced: False Splits (FS), False Merges (FM), Adapted Random Error (ARE), Adapted Random Precision (ARP), and Adapted Random Recall (ARR) (Arganda-Carreras et al., 2015). While each of these metrics provide insight into the general accuracy of the segmentation algorithms, they represent different and sometimes contrasting accuracy components. For example, a segment that covers the entire image contains no false splits; while a

segmentation where every pixel is a segment would contain no false merges. Given this quandary it becomes necessary to find a formula where a balance is provided between these metrics.

Subtracting the False Merges reduces the favourability of parameter sets that have segments which are too large. The theory here being that it would be better to have multiple segments per tree, than multiple trees per segment or worse yet, including ground in the segments. The logic behind this assumption is that a segment representing half of species γ should still present most of the characteristics of a segment representing an entire species γ, though potentially with lower deviation. In contrast, a segment that includes both species γ and species ε will have characteristics that are an average of the two species. This would make classification extremely difficult as this segment would not be like any of the training samples.

Four equations are proposed to merge the metrics to a form where a maximum can be found as the desirable solution. The one exception to this is that Equation 1 converges towards a limit, for this equation the most desirable state is the first occurrence of this limit, making this equation a little bit harder to work with than the other three.

2.5.2.1 Metric 1: "Small Segments"

MAX (ARP - FM)

This metric is chosen for having a strong emphasis on reducing the amount of over-segmentation in the image, this is based on the theory that a tree spilt into multiple segments could still have all segments classified as that species. However, multiple trees in the same segment could never be classified correctly).

2.5.2.2 Metric 2: "Average of Metrics"

MAX (-ARE - FM - FS)

This metric represents an evenly weighted combination of all metrics.

2.5.2.3 Metric 3: "Weighted Small Segments"

MAX (-FM - 0.5 * FS))

This metric puts a strong emphasis on avoiding false merges however does still consider false splits, just to a lesser degree. This represents that while less impactful false splits may still be damaging to the classification.

2.5.2.4 Metric 4: "Weighted Average of Metrics"

MAX (-ARP - 0.5 * ARR - FM - 0.5 * FS))

This final metric can be thought of as a combination of Eq2 and Eq3, where all metrics are considered but those with higher weighting giving to those metrics that avoid false merges.

2.5.2.5 Stability:

For the purposes of this paper stability is the absolute value of standard deviation divided by score; this is meant to provide a way of comparing scores from metrics that may not produce results within the same range.

2.5.2.6 Sharpness:

Sharpness is calculated as the mean of edge intensity using Sobel's filter and is used as an indication of how distinctly edges are presented within the image. This is a method that has been used for simple camera autofocus; and in this case is attempting to evaluate the amount of motion blur present in the image. However, it should be noted that the content of the image affects the sharpness and thus is only a proxy for amount of motion blur.

2.5.3 Compilation of Results

Experiments are run on all study sites, with all 4 algorithms, where the hyperparameters were selected by distributing tests across the entire range of possible settings. The result of this experimentation is over 140,000 tests run producing a very large data set; results are picked from this set to distill a more usable set of results. For each pair of metric and algorithm the results tests are extracted using an SQL query that picks the highest score for each site, as well as the hyper parameters that produce the highest average score when those parameters are applied to all sites. With 4 algorithms, 4 metrics, and 11 sites yields 4 * 4 * (11 + 1) results providing a much more manageable dataset to draw conclusions from.

2.6 Results of Image Segmentation Algorithms

Results are based upon the highest scoring hyperparameters for these sites, with complete tables located in Appendix C through I.

2.6.1 SLIC

To make effective use of SLIC it is helpful to have an estimate of the number of segments that should be expected in the image. This was calculated by demining the average size of a feature from the ground truth data, then determining how many times that area could be placed in each hectare. This parameter was corrected in terms of hectares, as opposed to over the whole image to help provide transferability to the results. Analysis of the training data shows that the average canopy size of trees at the study site is 0.519m^2 or 9576 clusters per ha.

Table 4 | Average segmentation score for SLIC across all four metrics

SLIC				
	Metric 1	**Metric 2**	**Metric 3**	**Metric 4**
Average	-0.1137	-0.4829	-0.3045	-0.2907
SD	0.0601	0.2835	0.1797	0.1872
Stability	0.5283	0.5871	0.5902	0.6439

Table 4 above shows the scores attained with SLIC, along with the deviation in those scores and a third value representing Stability. A full list of site scores is available in Appendix C.

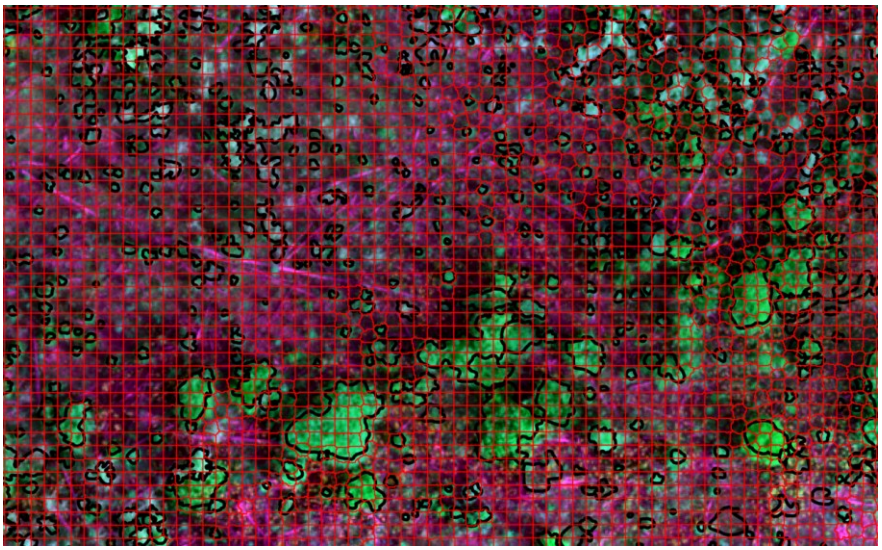


Figure 11 | Results of SLIC segmentation on 200rd site optimized for metric 1 in red, training segments in black.

Overall, the results show that the cluster scale has a heavy impact on the final segmentation producing what is nearly a grid, with segments larger than the smallest trees, yet still requiring many for the larger trees.

This method could theoretically be advanced further by looking at cluster merging algorithms to merge the larger trees into single segments, however at this time as we are currently more concerned with coverage of species than actual counts; this will be left to future work should tree counts become a desirable attribute. Additionally, it may be easier to merge the trees post species classification as it can be assumed that only segments of the same species should be merged.

2.6.2 Quick Shift

Quick Shift shows a much a more complicated pattern of results however it is not visually obvious that the segments are following tree crowns and many of the segments cover both crown and not crown. Compared to SLIC above there is generally higher stability but at the cost of lower scores (except for Metric 1). A complete list of site specific hyperparameters can be found in Appendix D.

Table 5 | Average segmentation score for Quick Shift across all four metrics

Quick Shift				
	Metric 1	**Metric 2**	**Metric 3**	**Metric 4**
Average	-0.0755	-0.5020	-0.3464	-0.3315
St. Dev	0.0295	0.2610	0.1195	0.1255
Stability	2.5569	1.9230	2.8991	2.6417

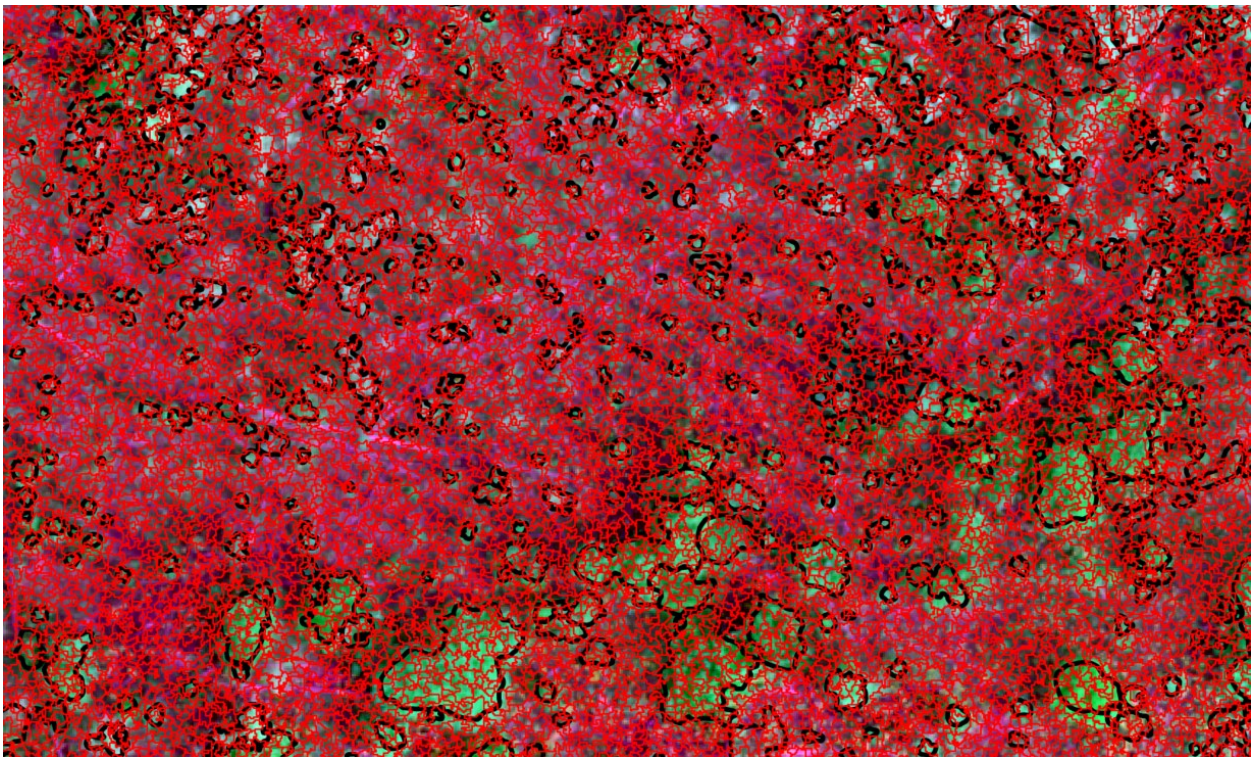


Figure 12 | Results of Quick Shift segmentation on 200rd site optimized for metric 1 in red, training segments in black.

2.6.3 F-Graph

Felzenszwalb's Efficient Graph produced interesting results; looking at Figure 13 it can be seen while many smaller trees were segmented, it simultaneously missed segmenting some trees completely including some of the larger ones. A complete list of site specific hyperparameters can be found in Appendix E.

Table 6 | Average segmentation score for Felzenszwalb's Efficient Graph across all four metrics

F-Graph				
	Metric 1	**Metric 2**	**Metric 3**	**Metric 4**
Average	-0.0823	-4.7949	-2.3340	-2.8789
St. Dev	0.0496	2.8222	1.2867	1.1941
Stability	1.6588	1.6990	1.8139	2.4109

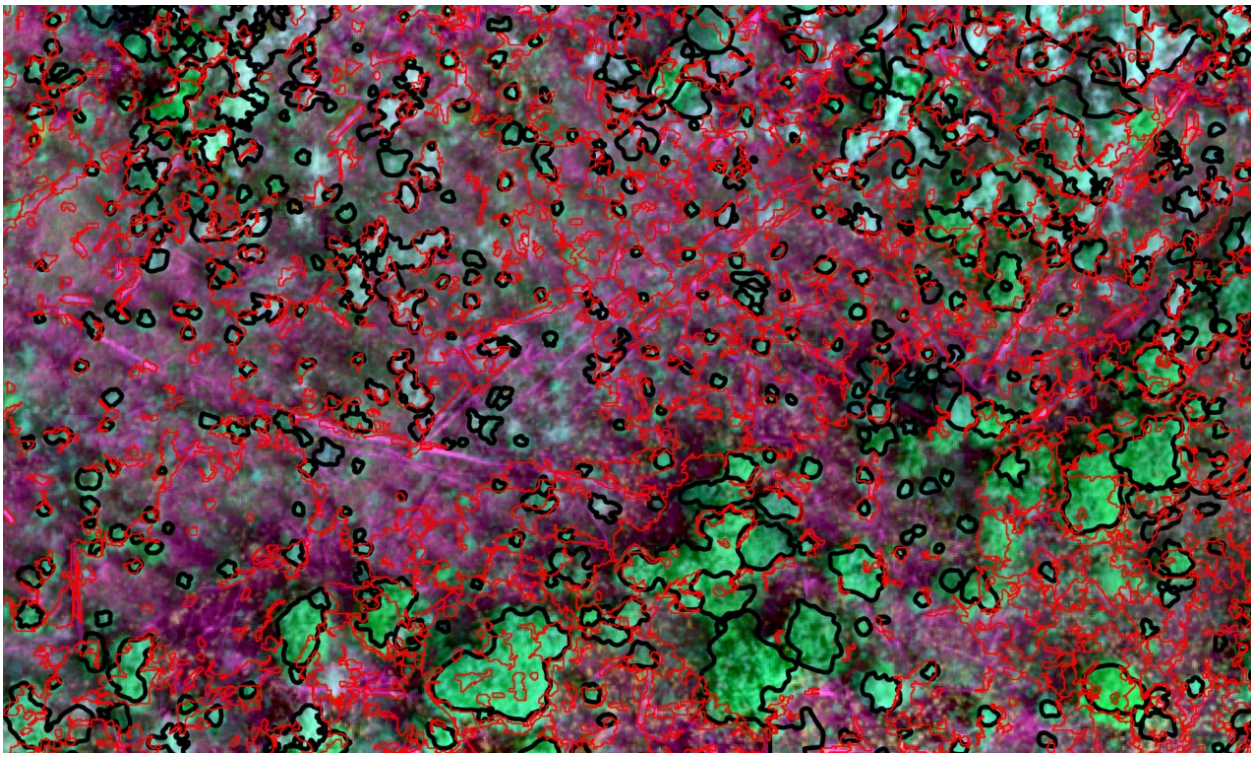


Figure 13 | Results of Felzenszwalb's Efficient Graph segmentation on 200rd site optimized for metric 1 in red, training segments in black.

2.6.4 Mean Shift

Mean shift was completed using SKLearn's estimate bandwidth function, however given how poor the results of the algorithm were, there was additional tests done manually setting bandwidth at various points ranging from 1 to 1000, however this did not produce any improvement in results, only a reduction in computation time as it bypassed the estimate bandwidth function. Mean Shift missed segmenting nearly all trees. The individual site scores are available in Appendix F.

Table 7 | Average segmentation score for Mean Shift across all four metrics

Mean Shift				
	Metric 1	**Metric 2**	**Metric 3**	**Metric 4**
Average	-1.8970	-3.3371	-2.9115	-4.1571
St. Dev	3.3994	3.8410	3.3919	3.1491
Stability	0.5580	0.8688	0.8584	1.3201

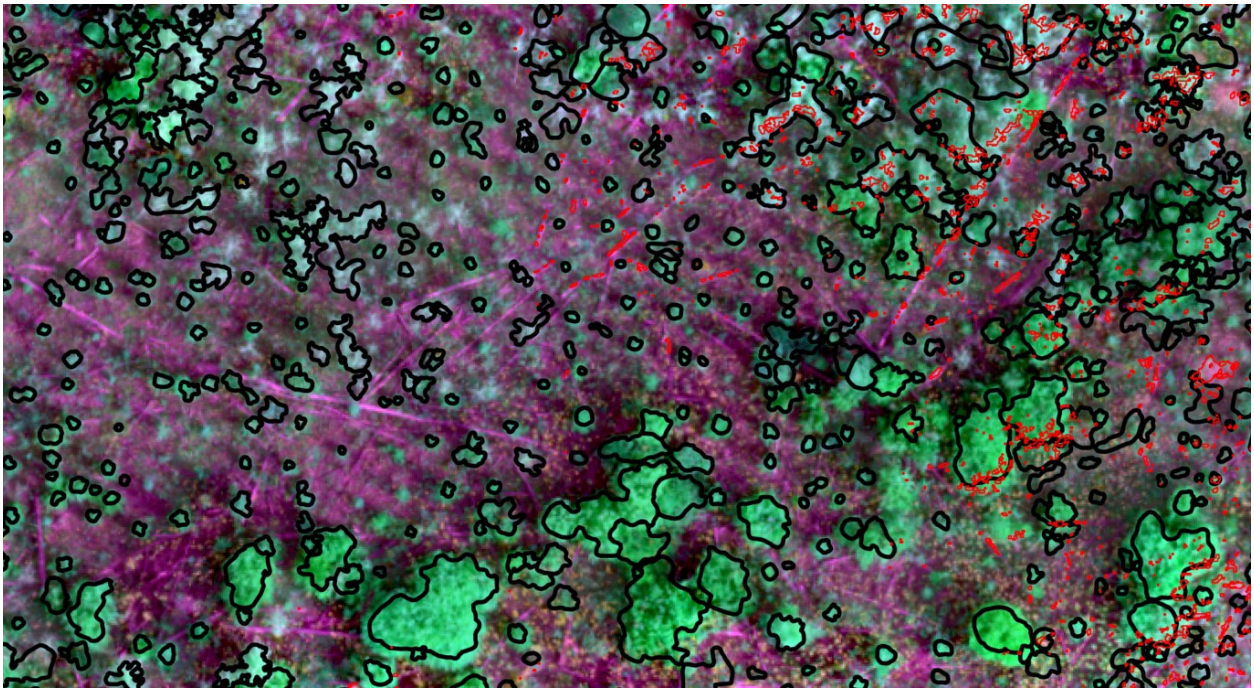


Figure 14 | Results of Mean Shift segmentation on 200rd site optimized for metric 1 in red, training segments in black.

2.6.5 YOLO

Figure 15 below shows the results; each object is given a bounding box, each box belongs to a class (maps to species of training data), followed by the likelihood that the object is of that class. One interesting result that was captured here however, is that attempting to train the algorithm to detect trees in general (tree vs no tree), was essentially a complete failure. By providing the training data classified as species, it was able to provide a higher success rate, though with a great deal of variance between species.

detection_model-ex-006--loss-0033.073.h5
Evaluation samples: 300
Using IoU: 0.5
Using Object Threshold: 0.3
Using Non-Maximum Suppression: 0.5
0: 0.0550 - Cs / Red-osier Dogwood (*cornus stolonifera*)
1: 0.0000 – Ac / Cottonwood (*populus balsamifera*)
3: 0.0000 – Bl / Subalpine Fir (*abies lasiocarpa*)
4: 0.2057 – Ep / Paper Birch (*betula papyrifera*)
5: 0.0000 – Fd / Douglas-Fir (*pseudotsuga menziesii*)
7: 0.1273 – Sx / Hybrid White Spruce (*picea glauca x engelmannii*)
mAP: 0.0647.

Figure 15 | Example of YOLO Segmentation and Classification, bounding boxes are labeled with species code: confidence percentage.

The detection rate of YOLOv3 was overall very low, however it does show a strong contrast between the ability to classify different species of trees. Due to its distinctly different format compared to the other algorithms, and the low accuracy of results this algorithm was not explored further for this paper.

2.6.6 Summary

Table 8 below provides a summary to make an easier comparison between methods. In Table 8 each metric has an average accuracy for all sites where each site used optimal parameters, as well as an aggregate value where the same hyperparameters were used for all sites. It is expected that the Aggregate scores will be lower than Average score, as it is less tuned to specific sites.

Table 8 | Summary of Algorithm performance across all four metrics. Average is the score of each site segmented independently and scores averaged, Aggregate is the score when the same hyper-parameters are used on all sites then averaged.

Average Scores				
Metric 1				
	SLIC	**QuickShift**	**F_Graph**	**MeanShift**
Average	-0.11367	-0.07550	-0.08227	-1.89704
Aggregate	-0.1161	-0.0763	-0.0897	-1.897
Metric 2				
	SLIC	**QuickShift**	**F_Graph**	**MeanShift**
Average	-0.48293	-0.50196	-4.79493	-3.33711
Aggregate	-5.0441	-5.6223	-5.3433	-3.3371
Metric 3				
	SLIC	**QuickShift**	**F_Graph**	**MeanShift**
Average	-0.30448	-0.34644	-2.334	-2.91146
Aggregate	-2.5232	-2.765	-2.7298	-2.9115
Metric 4				
	SLIC	**QuickShift**	**F_Graph**	**MeanShift**
Average	-0.29074	-0.33147	-2.87894	-4.15706
Aggregate	-3.1395	-3.2964	-3.2869	-4.1571

Based upon the scores above we can see that SLIC had the highest scores for metrics 3 and 4 and was a strong contender on all metrics; Metric 2 does show Mean Shift with the advantage for average score. Finally, Quick Shift as the leader for Metric 1 and showing up in 2nd place for average score on all metrics. To further examine this is it helpful to understand how the stability of the metrics between sites, as the Aggregate functions are what would be applied to novel sites without training data and having stable results provides expectations of the data quality.

Table 9 | Stability Matrix for segmentation algorithms and scoring metrics, higher stability indicates more consistent results between sites.

Stability	**Metric 1**	**Metric 2**	**Metric 3**	**Metric 4**
SLIC	1.892777	1.703242	1.69425	1.553036
Quick Shift	2.55691	1.923048	2.899101	2.641684
F Graph	1.658837	1.698979	1.813909	2.410898
Mean Shift	0.558048	0.868819	0.858369	1.320089

Table 9 shows that Quick Shift using Metric 3 provided the most stable results, with Metrics 4 and 1 claiming respective 2nd and 3rd place and was also the top performer on Metric 2. Conversely Mean Shift took last place in every category for stability.

2.6.7 Inter-site results

The results were also calculated with the case that every site is scored separately, however all using the same hyper-parameters. By using the same parameters for all sites, it provides an indication of expected results if segmenting a novel image without training. Calculated hyperparameters are in Appendix G, Appendix H, and Appendix I.

Figures were created for each of the four-scoring metrics. Metric 1: Figure 16, Metric 2: Figure 17, Metric 3: Figure 18, and Metric 4: Figure 19. These graphs show the accuracy of the site specific hyperparameters as blue dots, and the aggregate hyperparameters as red dots. The vertical distance between the blue and red dots shows the loss in accuracy from using aggregate hyperparameters.

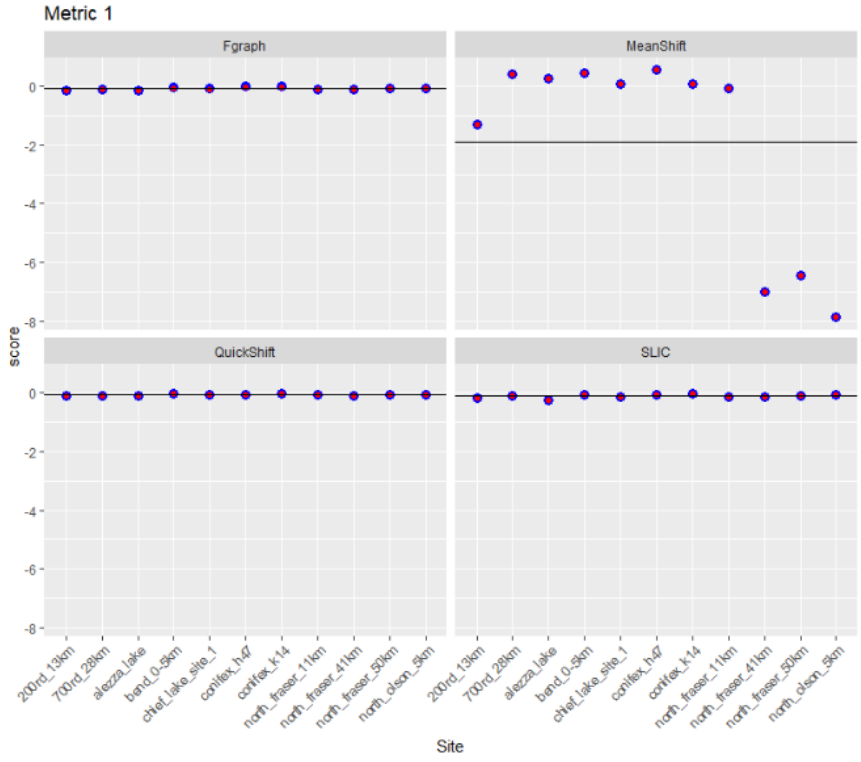


Figure 16 | *Metric 1 scores across sites*, horizontal line represents sore of all sites combined.

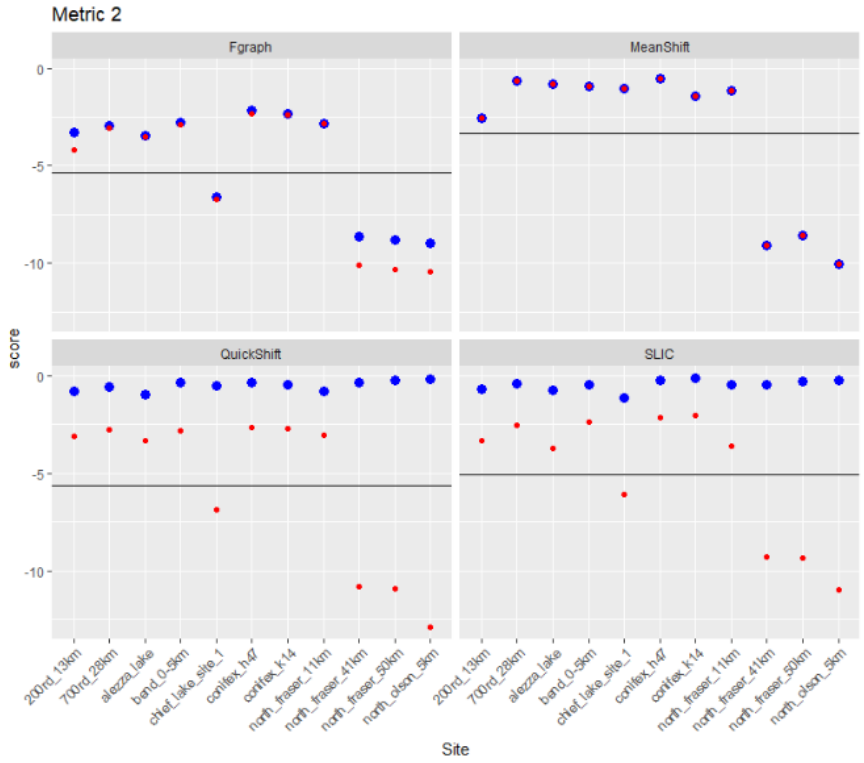


Figure 17 | Metric 2 scores across sites, horizontal line represents sore of all sites combined.

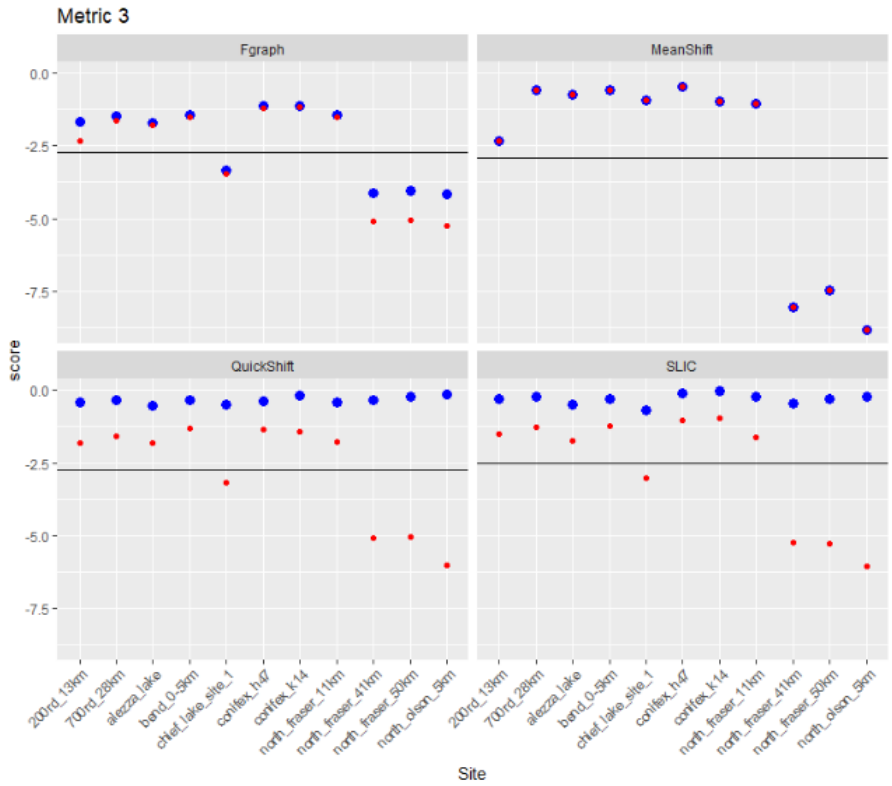


Figure 18 | Metric 3 scores across sites, horizontal line represents sore of all sites combined.

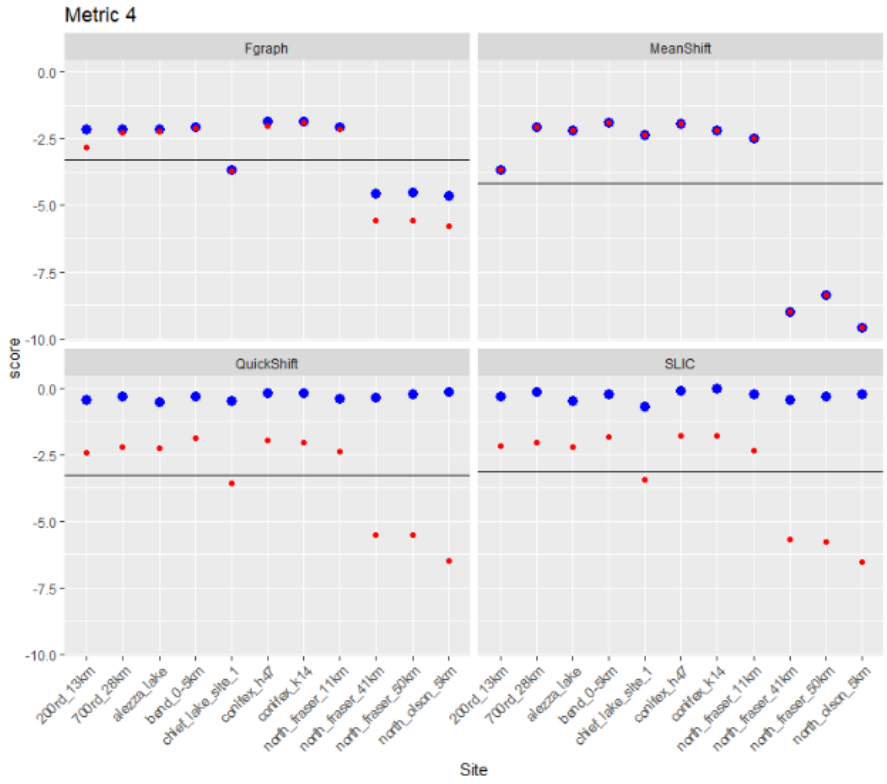


Figure 19 | Metric 4 scores across sites, horizontal line represents sore of all sites combined.

Figure 16 shows that there is almost no difference on score whether using site specific or aggregate hyperparameters, while all others show better scores when using site specific, except for Mean Shift which does not have hyperparameters and has identical results for all four metrics.

To help examine why some sites have larger differences in single site as opposed to aggregate hyperparameters, several additional factors were compared for correlations with aggregate scores in Table 10. The various sites represent a variety of conditions, variations in species composition, as well as age distribution (here represented as tree size).

Table 10 |Correlations between Metric 1 score and site attributes

Algorithm	SLIC	Fgraph	Quick Shift	Mean Shift
Conifer Coverage %	0.0184	0.0888	0.353	-0.0489
Deciduous Coverage %	0.3397	0.4988	0.5964	0.4531
Plant Coverage %	-0.157	-0.2813	-0.5394	-0.149
Leading spp. Coverage %	-0.1655	-0.1387	0.1435	0.0952
2nd spp. Coverage %	0.5085	0.4722	0.6989	-0.0685
3rd spp. Coverage %	0.5895	0.6098	0.4173	0.3034
Tree Size	-0.2081	-0.2692	-0.4849	-0.5384
Tree Size Variation	-0.1079	-0.102	-0.0487	-0.3578
Tree Count	-0.5408	-0.51	-0.4105	0.3193
Sharpness	-0.3125	-0.3462	-0.4116	0.0624

Some interesting notes on the correlations are how high third leading species seems to correlate with segmentation score. Tree Size and Variation have low correlation with segmentation scores.

2.7 Discussion

With the analysis completed, it is interesting to note that based on Metric 1 quick shift produces the best results. However, when comparing the results between SLIC and Quick Shift (available in appendix), SLIC's segments give the appearance of segments closer in size to the trees than with Quick Shift, Quick Shift's lack of over segmentation places it higher in the scoring system. Looking at this in a wholistic sense moving forward it will be essential to take a few of these metrics forward into future classification work to see which works best in the final classification.

Moving forward to classification of segments, SLIC and Quick Shift, optimized to Metrics 1 and 3 seem the most beneficial to proceed with. Between these two metrics, Metric 3 was chosen for its higher stability especially when used with the Quick Shift segmentation

algorithm. SLIC consistently produced the highest scores, additionally its uniform distribution may be beneficial as there are generally few crowns per segment due to the small size of each segment.

While there is likely an optimal combination of the base-metrics for the sake of producing an initial dataset that can be used to start training models, the naïve approach and equal rating of all 5 parameters was the initial attempt at finding a balance to these extremes. However, upon further testing, based on a visual examination using the ARP and subtracting the FM, seemed to produce the most usable results. The visual examination was focused mainly by looking at segmentations where the lines most closely followed the edges of the trees and were small enough to capture the smaller trees. However, at this stage it is impossible to definitively say this is the ideal metric, as it is unclear whether the visual intuition will translate directly to classification performance, and it was only practical to review a subset of the 135,000 experiments that were completed. A variety of metrics for determining the best segmentation were used to select those experiments from the database that would receive a manual review.

The parameters that maximize the metric ARP – FM produce segments that most closely align with the original training data. However, it should be noted that this formula was chosen by visual qualitative analysis of various simple potential metrics. This formula is likely not the ideal metric for optimization of segmentation, however there is a bit of a "which comes first: chicken or egg" problem at this stage; although the goal is to determine which segments produce the best classification; segments are concurrently used to develop the classification algorithms, which then are used to quantitatively score the segmentation properly. As such there is room for future work once the classification algorithm is developed to then run various segmentations against this classification to find an optimal solution.

It is notable that none of the algorithms produced cleanly delineated tree crowns. A possible explanation for this is the lack of sharpness between the crowns and surrounding vegetation, as these algorithms rely upon edge detection to place the segment boundaries. As an example, (Figure 20 below) is a Sobel Edge filter applied to the 200rd_13km site, displayed as intensity of edge. This image covers roughly the same area as figures above showing segments, however it is very hard to identify crowns at all in this image.

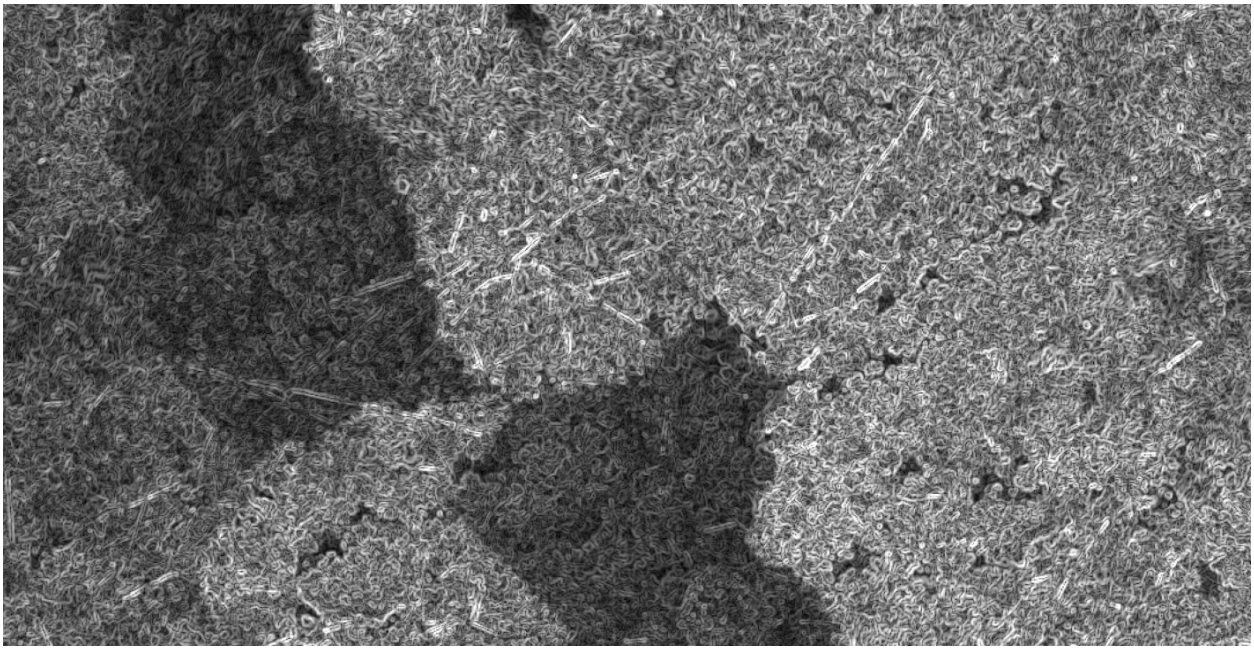


Figure 20 | Sobel Edge Detection from 200rd_13km site.

Another point of interest in this research is that SLIC performed better than anticipated, as it is the simplest of the algorithms it was selected to provide a control of sorts, to show the naive solution. However, it has shown itself as the strongest contender on three of the metrics, and 2nd on the fourth metric.

There is also future work looking at segmentation that works at multiple scales as there are large discrepancies in size between the smallest and largest trees, presenting a challenge to optimizing for segment size. Alternatively, if a particular size of tree is of interest it could be

useful to filter the training data to train base only upon trees of that size; giving up accuracy for the set of all trees to gain accuracy for a subset that is the target tree size.

Finally, it should be emphasised that many training sites must be used in the development of a set of generic aggregate hyperparameters. Initial testing for this paper was conducted using a single site (North Fraser 41km), and led to early results that were dramatically different than looking at the set of sites, and even at this stage eleven site is still a small sample size, and while it may be tempting to say that the sites [bend_0-5km, conifex_k14, 700rd_28km, conifex_h47, 200rd_13km, alezza_lake north_fraser_11km] are the most common presentation of sites, it cannot be ignored that four of the sites performed atypically, and three of these relatively similar.

Chapter 3

Image Classification for Early-Stage Vegetation in RPAS Imagery

3.1 Introduction

Classification is the process where the computer evaluates information and places that information into a class with the objective of providing meaning or context to that information. Classification is the fundamental problem that machine learning was pioneered upon. The origins of machine learning has been proposed as Frank Rosenblatt who developed the 'perceptron' in 1957 for the detection of letters of the alphabet (Fradkov, 2020).

As research advances machine learning is being used to classify an ever growing variety of objects, such as automated counting of cars (Biswas et al., 2017), monitoring sea ice (Dumitru et al., 2019) and forest fire detection (Seydi et al., 2022). The same way we as humans use our perception of the world to identify our surroundings, a computers ability to classify information allows it to solve problems within the real world.

The classification of information is an extensive discipline able to use a variety of techniques to work with a range of different data types. In this thesis a specific vein of research is pursued; specifically, the techniques that look at regions of images that can be processed as tabular data for classification.

The motivation to perform these classifications automatically is twofold. The first and more obvious motivation is that the identification of thousands or (if done at scale) millions of samples is a very mundane job which would be cost prohibitive and likely unfulfilling to have completed manually by human operators. Secondly the types of data produced are at a

dimensionality that is difficult to express to human operators; allowing the computer to have a more precise view between samples in a much faster time.

No classification system will yield perfect results; in practice a human would perform tree identification with much greater accuracy. However, a human would simply not undertake the task of individual tree identification at the landscape level. In this respect the advancement of research looks to make possible access to information at scales effectively not currently possible; despite the seemingly simple problem presented. An example of why this distinction matters is the development of Optical Character Recognition (OCR) that allowed for the digitization of entire libraries; while true reading and typing books is not a particularly challenging task, completing this process for entire libraries would have been effectively impossible without the aid of machine learning.

The ability to take imagery and classify it by vegetation type will provide land managers with new opportunities for decision making about their land base. An understanding of what vegetation is present, and to watch the development over time will allow for better understanding of how healthy the environment is; and make choices about how to best enhance ecosystems for a variety of factors. Some of these factors may include keeping it suitable for wildlife, ensuring economic value in the timber supply, or considering vegetation type impacts on forest fire susceptibility.

This chapter aims to answer the following questions:

- How should data be segmented?
- Should algorithms be used to balance the quantity of training data?
- Which algorithm should be used as input data for use with classification?

- Are trained models re-usable?
- What are the key areas for improvement?

3.2 Definitions

The definitions presented below are intended to help provide clarity to readers as to the specific definitions being applied to terms in this thesis. These are especially important as the terms used can carry domain specific definitions that may not be consistent across the domains of study involved: Remote Sensing, Machine Learning, Ecology, Computer Vision, etc. The details within the definitions also help to provide clarity to the methods presented.

Feature: an instance of what is being modelled in the real world, in the case of this work a feature is an individual tree.

Segment: a region of imagery; these may be hand drawn (**Truth Segments** as seen in Figure 3); or automatically produced by segmentation algorithms. In an ideal circumstance there would be a 1:1 mapping between features and segments, however in practice this is rarely the case.

Sample: the information about the intersection of imagery and segments; the machine learning algorithms used here do not work directly on imagery, but rather numeric information. Samples are produced by calculating raster statistics for each segment. Additionally, *Truth Segments* contain information about species classes of the feature they represent; algorithm generated samples contain information about overlap between the given sample and intersecting *Truth Segments*.

Coverage: For data that has been automatically segmented the boundaries will very rarely be fully contained within ground truth polygons, as such it is beneficial to calculate how much of the area in the segment being trained on is represented by the labeled species. Note that coverage

is a measure of sample purity only and does not convey how much of the original feature is contained by the segment; see the first and last examples in Figure 21, both have the same coverage, yet the first example covers more of the original feature.

Limitations of assigning a single species to each segment are that it does not account for the case where segments overlap two trees of the same species, coverage will be calculated only as the single tree with the greatest coverage, potentially underrepresenting coverage.

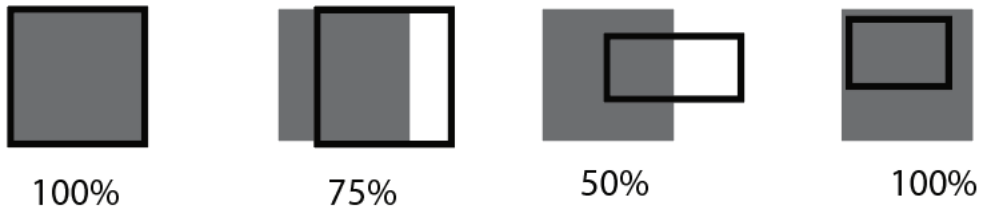


Figure 21 | Examples of how coverage is measured based on segment overlap.

Dimensionality: Refers to the number of attributes each sample contains, for example five bands each with a mean and standard deviation would provide ten dimensions to the training process. High dimensionality is a relative term generally comparing the ratio of dimensions to training samples, this is useful in conveying comparisons of the general behavior of various algorithms.

Model: the result of using training algorithms against the training dataset; the resultant model can then be used to classify novel data.

Overfitting: an occurrence in models where the trained model can easily identify the training data but fails to maintain accuracy on novel datasets. Conceptually this could be thought of as a fit model being able to identify people in photographs, where an overfit model may only identify specific persons as people. In simpler terms a fit model would be able to identify aspen trees,

while an overfit model would only identify a specific subset of aspen trees; likely those contained in the training data or possessing an extremely similar representation.

3.3 Data Preparation

The data used in this Chapter is the same dataset that was used in Chapter 2, with the results of Chapter 2 preparing the data for further analysis. More generally before the data can undergo object-oriented classification, it must be provided to these algorithms as segments of the image, where each segment will be treated as a unit to be segmented. This also begins to show the modular approach of the overall process from data collection to classification in that any process which segments the image could be used in place of the results of Chapter 2 as the input data for this section.

3.3.1 Zonal Statistics

To prepare the previously segmented imagery for classification it must be converted from raster into tabular data. This is done by overlaying the segments on the orthomosaic, calculating statistics for each band, and saving the results into a database table. The statistics calculated here are minimum, maximum, mean, number of pixels, standard deviation, median, and range for all pixels that fall within the segment.

In addition to the zonal statistics the table includs site, date of capture and how many bands are present. This information will be required as the geometry of segments will not be included in the algorithm training process.

3.3.2 Species Coverage

It is also important that we know what species are represented by the recorded statistics; this is achieved by overlaying the segments with the ground truth data. Each segment is assigned

a species code based upon which shape in the ground truth data has the most overlap; percent coverage is calculated as what percentage of the new segment is contained within the truth segment with the greatest overlap.

Table 11 | Table of sample counts by site, both total samples as well as counts of only target species

Samples in each Site											
	200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
All	11529	3774	5649	1521	12861	1398	1077	8331	1677	1107	2862
Target	2031	1977	2064	636	1491	261	513	3453	270	468	567

Table 11 shows how many training samples are present in each site of both all samples as well as samples of the target species. This will be used later to see if there is correlation between quantity of training data and accuracy. As well as how the ratio of target species affect the target accuracy specifically.

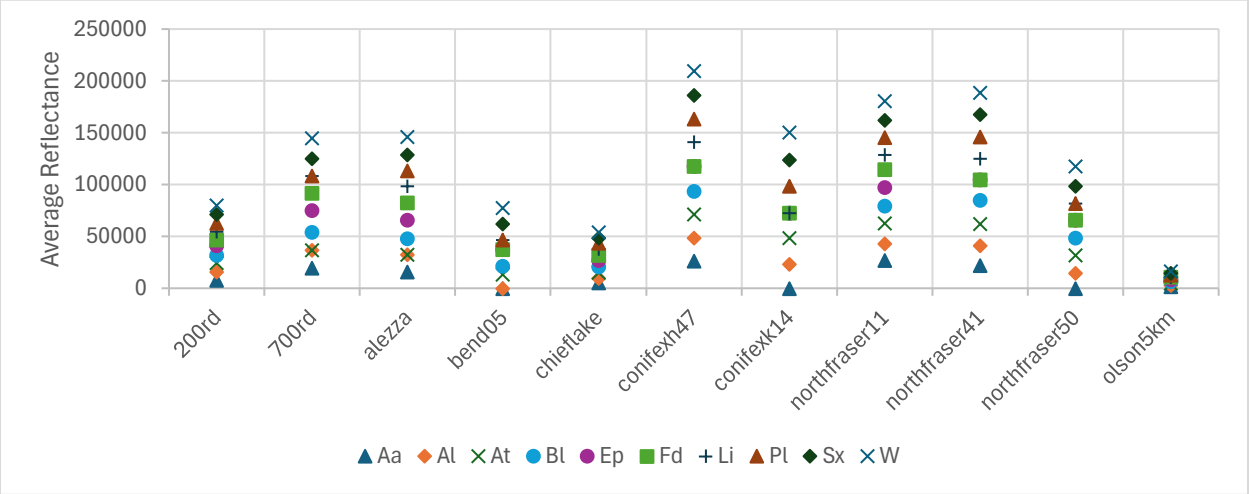


Figure 22 | Average reflectance of species across all five bands of 10 most common species per site.

The sites varied considerably in terms of quantity and quality of data. Even after calibration, 200rd, Chief Lake, and Olson5km have significantly lower reflectance than the other sites (Figure 22). This is believed to be due to issues with the calibrated reflectance panel with the RedEdge-M camera. Through other ongoing research projects, it has been observed that newer versions of the MicaSense cameras utilizing version 2 of the DLS sensor provide more consistent and reliable calibration results. This lack of consistent calibration is likely to be the cause of negative impacts on accuracy of models trained on multiple sites.

3.4 Classification Algorithms

In this research a variety of common classification algorithms were tested and compared to each other for effectiveness at classifying early seral vegetation. While machine learning is a powerful tool for solving a variety of problems, not all methods of learning work for all types of data. Different algorithms may perform better or worse depending upon the amount of noise in data, number of variables present, correlation between input variables amongst others.

3.4.1 Support Vector Machines (SVM)

Support Vector Machines (Cristianini and Shawe-Taylor, 2000) are a binary non-probabilistic classifier building upon the work at UC Berkley and Bell Laboratories (Boser et al., 1992) building classifiers that optimize the margins between classes. SVM classifiers are particularly suited to problems with high dimensionality (Erfani et al., 2016). However SVM classifiers also have the drawback of being a binary classification such that if there is a goal of detecting multiple classes, multiple classifiers must be used with a method of combining results (Duan and Keerthi, 2005). SVM could represent an advantage in this study due to the high number of input variables, with five image bands, each containing multiple statistics (mean, median, SD, etc.) and avoiding the 'curse of dimensionality' (Friedman, 1997) is valuable; conversely it is also hypothesized that if the representation of a given species were to change due to lighting conditions or time in the phenological cycle, the classifier could have a challenge differentiating these different representations.

Implementation Used: https://scikit-learn.org/stable/modules/svm.html. With gamma parameter set to auto.

3.4.2 Random Forest (RF)

The Random Forest Classifier (Breiman, 2001) is one of the more common machine learning algorithms used today. Random Forest classifiers have a relatively high resistance to overfitting as well as resistance to outliers compared to some other methods; however random forests can be heavily impacted by the curse of dimensionality especially when using imbalanced quantities of training data (Evangelista et al., 2006).

Random Forest is a type of ensemble learning building upon a traditional decision tree. To build the random forest multiple random subsets of data, a random subset of variables is used to build a decision tree; this collection of trees comprises the forest. When classifying new data each sample is run through all the decision trees and the most common result of all trees is the identified class. This method of building prevents overfitting as the complexity of each individual tree is contained and can provide resistance to outliers as even if a tree is providing invalid results, it can still be overpowered by the rest of the forest. Conversely highly imbalanced data may produce poor results as the trees are built by forming even distribution at each decision point which may force the algorithm to split inside of classes as well as provide an over abundance of leaf nodes for the overrepresented class. Additionally, the curse of dimensionality can be particularly challenging here as redundant variables will cause more decisions to be made on the redundant information potentially reducing the impact of other important variables.

For the purposes of this study, it is expected that Random Forest will have a good ability to handle the large number of input classes, as well as handle outliers such as a tree that is unhealthy or in shadows on the imagery. It is also susceptible to the unbalanced quantities of training data, and that all five bands of imagery will be highly correlated due to relative closeness in spectrum especially in the visible bands.

Implementation Used: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. With number of estimators increased from 10 to 1000.

3.4.3 K Nearest Neighbors (KNN)

K Nearest Neighbor classification (Taunk et al., 2019) is a common algorithm used in the field of remote sensing, and is often utilized due to its ease of implementation and computation; yet is still able to produce useful results in many cases. This algorithm works by placing all samples into n-dimensional space where n is the number of input variables provided, along with K randomly selected seeds. Each iteration every sample is placed into the cluster it is nearest to based on an n-dimensional vector; then each seed is updated to be the center or average of all points in the given cluster; this process is repeated either a fixed number of times or until stability is reached in which cluster each sample belongs to. After training new data is simply assigned to the cluster to which it is nearest.

KNN is utilized in this research as a baseline of a relatively naive approach to classification, making it a form of base line for the lowest effort approach to classification.

Implementation Used: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.

3.4.4 Multinomial Naive Bayes Probabilistic Classifier (MNB)

The Multinomial Naive Bayes Probabilistic Classifier works by calculating the normal distribution of values for each class and then places each new sample into the distribution where it has the highest probability of matching. Naïve Bayes classifier is well suited to handling the curse of dimensionality, however, it also assumes that all input variables are independent. For the

purposes of this research this has the potential to be beneficial as the data has relatively high dimensionality, however at the same time may suffer in the ability to tell a more reflective species from an over-exposed photo as the relative weights of variables are not considered.

Implementation Used: https://scikit-learn.org/stable/modules/naive_bayes.html.

3.5 Minimum data requirements

At the scale of the current research the ground truth serves as a baseline for classification effectiveness with the given dataset. At the scale of the broader project the accuracy of ground truth opposed to algorithmically generated segments provides some indication as to feasibility of fully automated systems.

During training for the classification algorithms, species with too few samples must be filtered out; the minimum requirement is that there be at least as many samples as folds during training. In the case where oversampling is to be used there must be the number of unique classes times the number of folds samples of each species. In the case that a species does not have enough samples from a given set of sites it is removed from the training.

As the data used in this research has large variations in the quantities of training samples available per species per site; there is a potential for the classification algorithms to over classify those samples which have a higher representation. To attempt to mitigate this effect all classification algorithms were tested with Synthetic Minority Over-sampling Technique (Chawla et al., 2002). SMOTE works by oversampling underrepresented classes while simultaneously under sampling overrepresented classes. In SMOTE the synthetic oversampling is done by producing new features within the sample space of underrepresented classes as opposed to simply repeating real samples.

3.6 Framework for Evaluation

To evaluate the comparable effectiveness of different ML Classification Algorithms, each algorithm was tested for accuracy using a 4-fold cross validation, for a variety of configurations, the choices made for each configuration are based on the following factors.

1. Algorithm used: Random Forest, Support Vector Machines, K Nearest Neighbor, or Multinominal Naive Bayes Classifier.
2. Samples: ground truth, SLIC segmentation or Quick Shift segmentation.
3. How many sites are included: 1, 3, 7, 8, or 11.
4. Data may be either over-sampled or not using the Synthetic Minority Over-sampling Technique (SMOTE) function (Chawla et al., 2002).

The above options produce 120 scenarios for configuration of the processing pipeline including options for classification algorithm, segmentation algorithm, how many sites to use and whether or not to oversample the data. Within each scenario the accuracy is calculated for a range of minimum coverage thresholds. The objective is to determine which of the scenarios has the highest accuracy in classification; as well as looking for trends in specific options (i.e., does SLIC produces consistently better, consistently worse, or mixed performance compared to Quick Shift when used as the input segments; regardless of the other options chosen).

3.6 Results

The sites can be broken down into the following sets, based on the results of single site segmentation accuracy (Figure 16 to Figure 19):

Alpha (α) 1 site: Chief Lake

Beta (β) 3 sites: North Fraser 41, North Fraser 50, North Olson 5km

Gamma (γ) 7 sites: 200rd, 700rd, Alezza, Bend05km, ConifexH47, ConifexK14, NorthFraser11

Delta (δ) 8 sites: 200rd, 700rd, Alezza, Bend05km, ChiefLake, ConifexH47, ConifexK14, NorthFraser11

Epsilon (ε) 11 sites: all sites

These groups were formed based on the results of single site segmentation accuracy found in Figure 18. Group α contains only a single (Chief Lake) based on having largest quantity of training samples available. Group β is composed of three sites that represented the lowest segmentation scores (Figure 18); when looking at only the target species for Quick Shift and SLIC segmentation. Group γ represents the seven sites with the highest segmentation scores and group δ is a combination of groups α and γ, making it the set of the 8 highest segmentation accuracies. Finally, group ε represents the set of all sites with training data.

These sets are used for separating the data for training and testing models and are intended to help determine the stability of models moving from site to site.

Training was performed using all available species, the accuracy of the classification was computed in two ways; using all species in data set (Appendix A), as well as a subset of species. The target species are a subset that have been identified as valuable for moose browse; these species are Trembling Aspen (*Populus tremuloides*), Red-osier Dogwood (*Cornus stolonifera*), Paper Birch (*Betula papyrifera*), Highbush-Cranberry (*Viburnum edule*), and Willow (*Salix spp.*).

3.6.1 Ground Truth Segments

This section begins by analyzing what accuracy can be achieved by training ML algorithms on the ground truth segments. Answering this question has two motivations, first it shows if the trees can even be identified by spectral reflectance at all. Second this provides a baseline accuracy that segmentation can be compared against; as the true effectiveness of segmentation can not be directly observed, but rather how good the final classification will be.

3.6.1.1 Classification of individual Sites

First the accuracies for each site are determined using unaltered data in Table 12, then the accuracies with SMOTE are shown in Table 13, finally Table 14 demonstrates how SMOTE changed the accuracy of classification achieved. The ideal classification algorithm will provide accuracies that are consistently high across sites.

Table 12 | Classification accuracy of training segments for all species at each site with unbalanced data

Accuracy **No oversample**		Site Classification Accuracy										
		200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
Algorithm	RF	63%	81%	73%	78%	63%	68%	79%	81%	56%	75%	61%
	SVM	65%	83%	74%	78%	65%	68%	80%	82%	52%	79%	63%
	MNB	34%	57%	47%	60%	33%	54%	54%	42%	34%	49%	28%
	KNN	52%	79%	66%	77%	48%	62%	69%	75%	46%	72%	40%

Table 13 | Classification accuracy of training segments for all species at each site with SMOTE balanced data

Accuracy **SMOTE**		Site Classification Accuracy										
		200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
Algorithm	RF	63%	85%	73%	78%	62%	74%	87%	82%	NA[3]	83%	57%
	SVM	65%	85%	71%	77%	64%	74%	88%	78%	NA	83%	59%
	MNB	41%	74%	53%	67%	40%	66%	72%	61%	NA	70%	33%
	KNN	43%	77%	69%	68%	44%	62%	76%	66%	NA	75%	37%

Table 14 | Change in classification accuracy of training segments for all species resulting from using SMOTE.

Accuracy **Change**		Effect of SMOTE on Site Classification Accuracy										
		200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
Algorithm	RF	0%	4%	0%	0%	-1%	6%	8%	1%	NA	8%	-4%
	SVM	0%	2%	-3%	-1%	-1%	6%	8%	-4%	NA	4%	-4%
	MNB	7%	17%	6%	7%	7%	12%	18%	19%	NA	21%	5%
	KNN	-9%	-2%	3%	-9%	-4%	0%	7%	-9%	NA	3%	-3%

Table 14 shows the overall net effect of SMOTE is very small, with the exception being improvements across the board with the MNB classifier; these results must however be tempered with the understanding that even with these improvements the MNB classifier still produces consistently lower accuracies than the other classifiers. Further NorthFraser41 had too few samples to use SMOTE, resulting in a failure to train; highlighting that while SMOTE may be able to help balance the data, there does still need to be sufficient data captured.

Another question that one might ask is what accuracy can be obtained if looking only for a few target species, as opposed to identifying everything in the land base. Table 15 presents the accuracy of models, tuned to provide the best possible accuracy of the five target species

[3] Insufficient training samples were present to complete classification.

presented in this paper. Again, SMOTE was used to oversample and balance the number of samples in the dataset shown in Table 16, with the difference in results in Table 17.

Table 15 | | Classification accuracy of training segments for target species at each site with unbalanced data

Accuracy **No oversample**		**Site Classification Accuracy (Target Species Only)**										
		200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
Algorithm	RF	59%	90%	56%	62%	46%	38%	82%	78%	35%	67%	43%
	SVM	61%	91%	69%	64%	44%	37%	81%	79%	18%	71%	43%
	MNB	35%	57%	43%	49%	26%	34%	60%	34%	45%	52%	28%
	KNN	48%	88%	56%	63%	30%	28%	70%	70%	27%	64%	29%

Table 16 | Classification accuracy of training segments for target species at each site with SMOTE balanced data

Accuracy **SMOTE**		**Site Classification Accuracy (SMOTE & Target Species Only)**										
		200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
Algorithm	RF	60%	95%	66%	65%	53%	35%	83%	78%	NA	67%	52%
	SVM	59%	96%	68%	68%	57%	43%	83%	77%	NA	71%	49%
	MNB	42%	83%	50%	56%	30%	30%	68%	63%	NA	59%	48%
	KNN	42%	89%	51%	56%	43%	32%	69%	65%	NA	61%	45%

Table 17 | Change in classification accuracy of training segments for target species resulting from using SMOTE.

Accuracy **Change**		**Effect of SMOTE on Site Classification Accuracy (Target Species Only)**										
		200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
Algorithm	RF	1%	5%	10%	3%	7%	-3%	1%	0%	NA	0%	9%
	SVM	-2%	5%	-1%	4%	13%	6%	2%	-2%	NA	0%	6%
	MNB	7%	26%	7%	7%	4%	-4%	8%	29%	NA	7%	20%
	KNN	-6%	1%	-5%	-7%	13%	4%	-1%	-5%	NA	-3%	16%

Looking at Table 17, it is shown that when examining only the target species SMOTE is generally positive for Random Forest, and SVM in addition to the MNB as seen when using all sites.

Table 16 shows that most sites saw a reduction in classification accuracy when looking at only a subset of species that are targeted. This could be from a variety of reasons, such as how hard the target species are to classify relative to the overall set, or how well represented the target species are in the training data. This demonstrates that there are different ways to assess accuracy

based upon the question being asked. If a forest manager wants to know about all species vs only target species; in the case of ConifexH47, this could be over 30% difference which may lead to different conclusions for the viability of machine learning for classification.

3.6.1.2 Classification done with Groups of sites.

The next set of results look at training the classification models on groups of sites; such that the models are more generalized than those using only a single site for input. By training on multiple sites models are more likely to be representative of results that might be expected when classifying novel data. Just as with the individual sites these tables are presented as Table 18 showing the accuracy with the original unbalanced data; Table 19 showing the results after oversampling with SMOTE, and Table 20 showing the effects of SMOTE on classification accuracy.

Table 18 | Classification accuracy of training segments for all species by site group with unbalanced data

Accuracy No oversample		Site Group				
		α	β	γ	δ	ε
Algorithm	RF	63%	64%	75%	73%	71%
	SVM	65%	65%	76%	74%	72%
	MNB	33%	37%	50%	48%	45%
	KNN	48%	53%	69%	66%	62%

Table 19 | Classification accuracy of training segments for all species by site group with SMOTE balanced data

Accuracy SMOTE		Site Group				
		α	β	γ	δ	ε
Algorithm	RF	62%	70%	77%	76%	74%
	SVM	64%	71%	77%	75%	74%
	MNB	40%	52%	62%	59%	58%
	KNN	44%	56%	66%	63%	62%

Table 20 | Change in classification accuracy of training segments for all species resulting from using SMOTE with site groups.

Accuracy Change		Site Group				
		α	β	γ	δ	ε
Algorithm	RF	-1%	6%	3%	2%	4%
	SVM	-1%	6%	1%	1%	3%
	MNB	7%	15%	12%	12%	13%
	KNN	-4%	3%	-3%	-3%	-1%

With the site groupings, we again see that MNB sees the biggest improvement in accuracy from SMOTE, but is also still the lowest accuracy algorithm tested, even with SMOTE applied. In terms of both Random Forest and SVM, there is on average a small increase in accuracy that can be achieved by using SMOTE on the data before classification.

The groups were then again tested for classification on the target species only in Table 21 and Table 22, just as were the individual sites.

Table 21 | Classification accuracy of training segments for target species by site group with unbalanced data

Accuracy No oversample		Site Group (Target Species Only)				
		α	β	γ	δ	ε
Algorithm	RF	46%	48%	66%	64%	60%
	SVM	44%	44%	69%	66%	60%
	MNB	26%	42%	45%	42%	42%
	KNN	30%	40%	60%	57%	52%

Table 22 | Classification accuracy of training segments for target species by site group with SMOTE balanced data

Accuracy SMOTE		Site Group (Target Species Only)				
		α	β	γ	δ	ε
Algorithm	RF	53%	60%	69%	67%	65%
	SVM	57%	60%	71%	69%	67%
	MNB	30%	54%	56%	53%	53%
	KNN	43%	53%	58%	56%	55%

Table 23 | Change in classification accuracy of training segments for target species resulting from using SMOTE with site groups.

Accuracy Change	Site Group (Target Species Only)				
	α	β	γ	δ	ε

Algorithm						
	RF	7%	11%	2%	3%	6%
	SVM	13%	16%	2%	3%	7%
	MNB	4%	12%	11%	11%	11%
	KNN	13%	13%	-3%	-1%	3%

Looking at Table 23, we do however see a stronger positive result to using SMOTE to oversample data, when training on and testing using groups of sites as opposed to individual sites. While it is hard to determine why a bigger increase is seen here, one potential cause could be that the sample sizes of the most underrepresented classes have more data, potentially causing less error to be induced by SMOTE.

3.6.2 Effects of Segmentation and Coverage on Classification Accuracy

The accuracy of classification using the automatically generated segments is presented in graphical form. The minimum coverage of a segment to be considered identified as a species has two primary effects making selecting a coverage threshold non-trivial. First, as the minimum coverage required decreases the purity of samples is reduced presenting more noise to the training algorithms. Conversely as the minimum coverage required increases the number of samples available for training decreases; this provides the need to determine an optimal ratio of quality and quantity samples. It is also at this stage where the differences between the human created and computer-generated segments may become most apparent, as a difference of even a single pixel would show as error at this stage; in order to prevent training sized from becoming vanishingly small some level of error must likely be tolerated at this stage.

Due to the large amount of computation required to train the models at various coverage levels only the two best performing algorithms Quick Shift and SLIC were continued to this stage of analysis.

The graphs presented below show the classification accuracies for the case with SMOTE applied to the data, tables without SMOTE can be found in Appendix J and Appendix K.

3.6.2.1 Quick Shift Segments

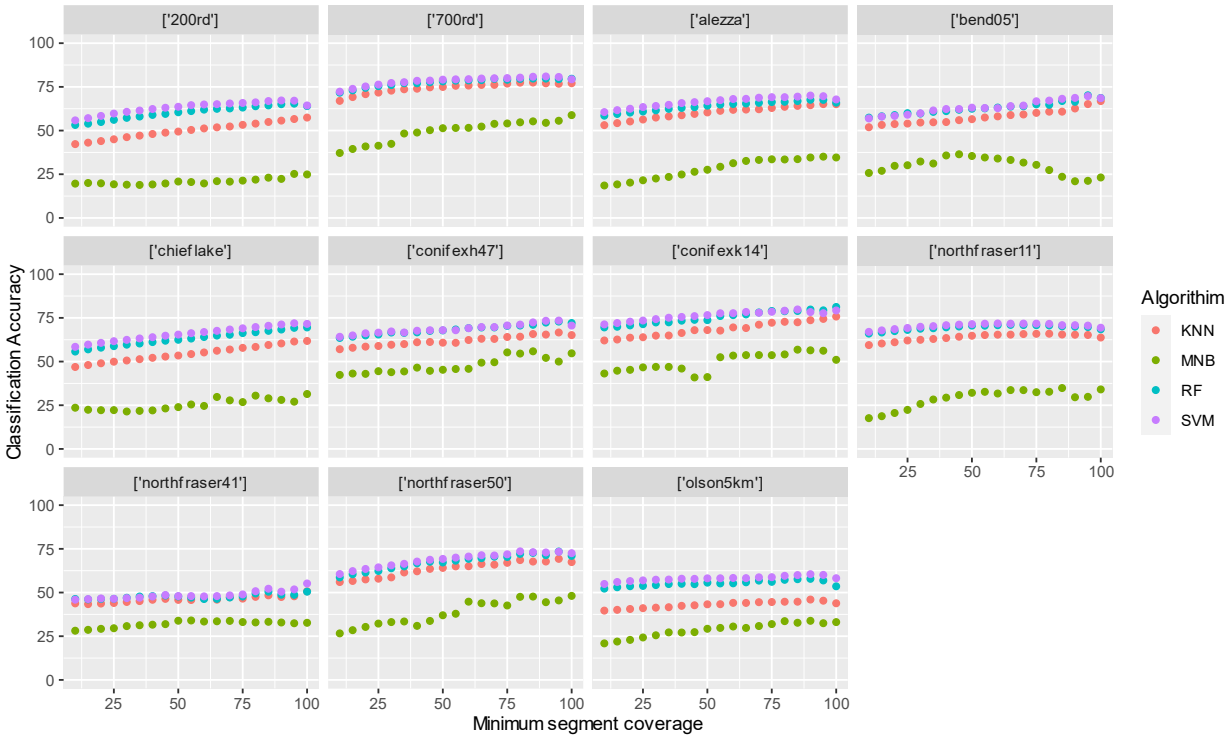


Figure 23 | Classification Accuracy of Quick Shift Segments on individual sites.

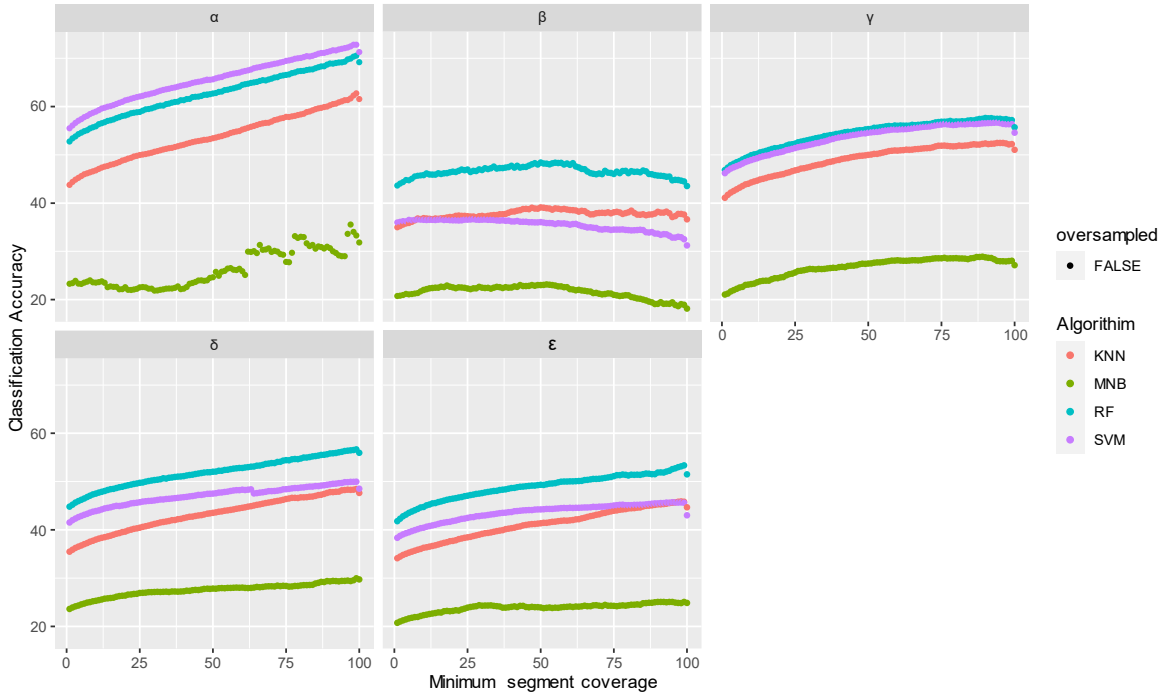


Figure 24 | Classification Accuracy of Quick Shift Segments on groups of sites.

For the case of the Quick Shift algorithm, Figure 23 and Figure 24 show that in general requiring a very high level of coverage is very beneficial, with the exception that at or very near 100% coverage, less training data is available due to rejecting segments for differences as small as 1px which could certainly be explained as human error in the delineation process. I would suggest that a 75% minimum coverage may be a good starting point when working with groups of sites as it is around this range that diminishing returns are starting be observed; while for a single site 90% coverage may make more sense.

3.6.2.2 SLIC Segments

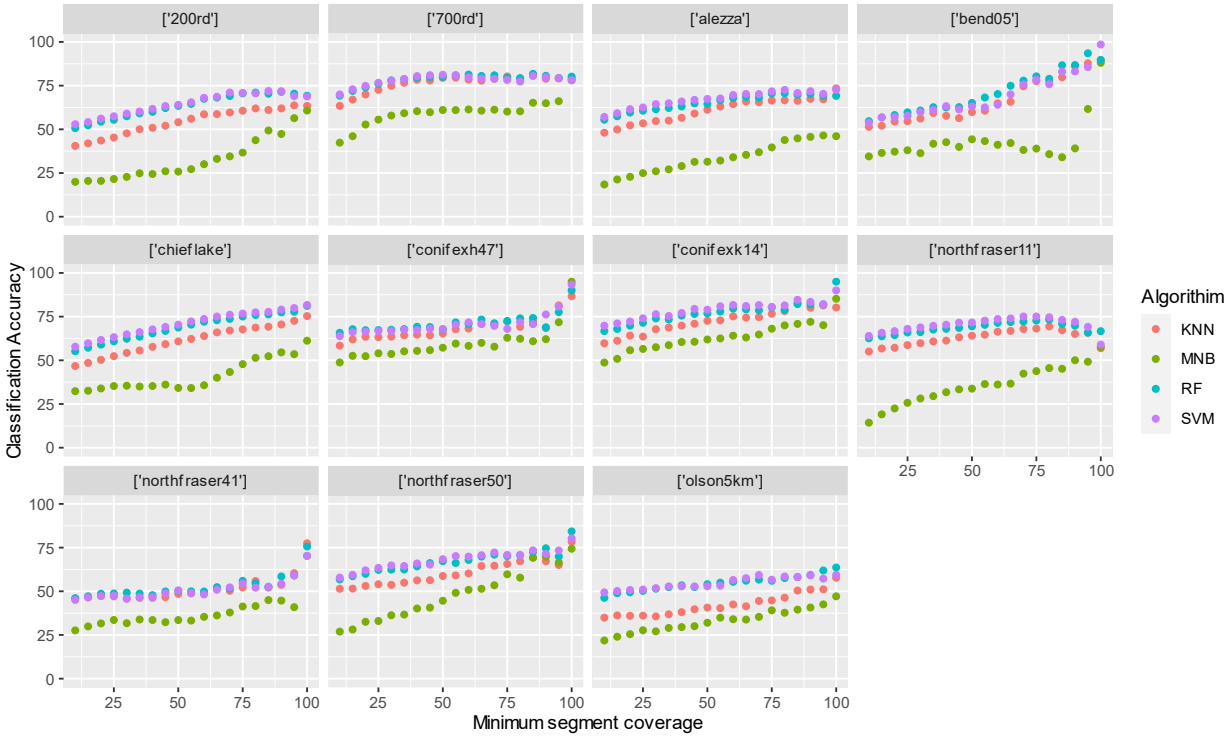


Figure 25 | Classification Accuracy of SLIC Segments on individual sites.

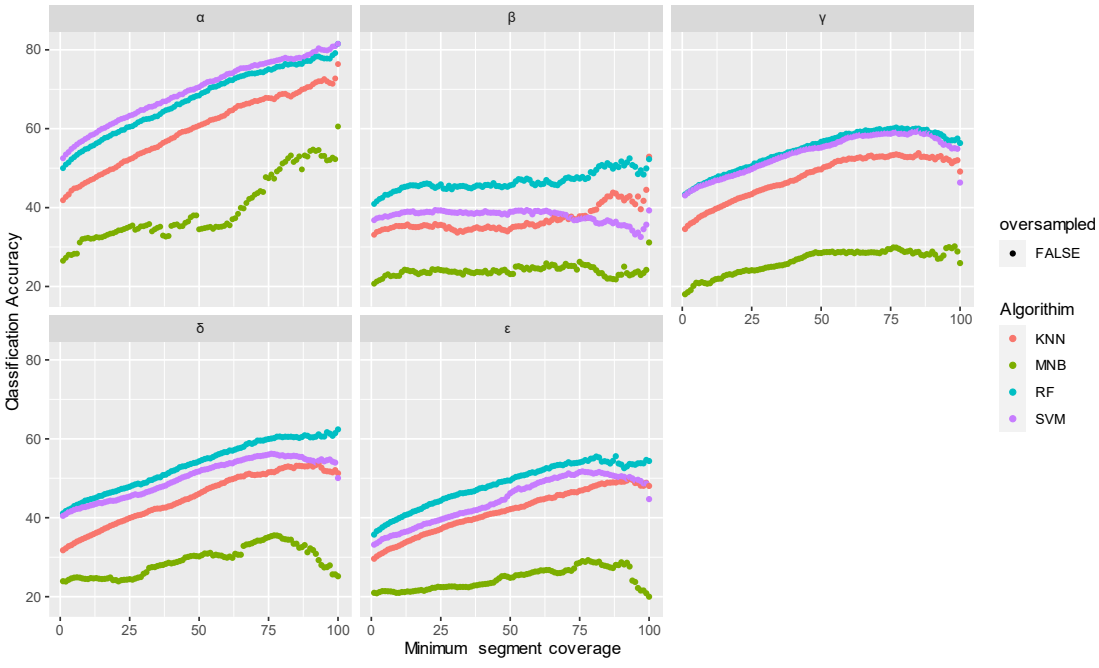


Figure 26 | Classification Accuracy of SLIC Segments on groups of sites.

The SLIC segments provide more interesting results in terms of ideal coverage levels with most cases preferring high coverage with some exceptions. The individual sites 200rd and NorthFraser11 along with groups γ and ε prefer a coverage in the range of 75%-80% before accuracy begins to decrease.

3.6.2.3 Segmentation Accuracy Comparison

The below graphs show a clearer comparison of the difference in accuracies achieved with SLIC and QuickShift.

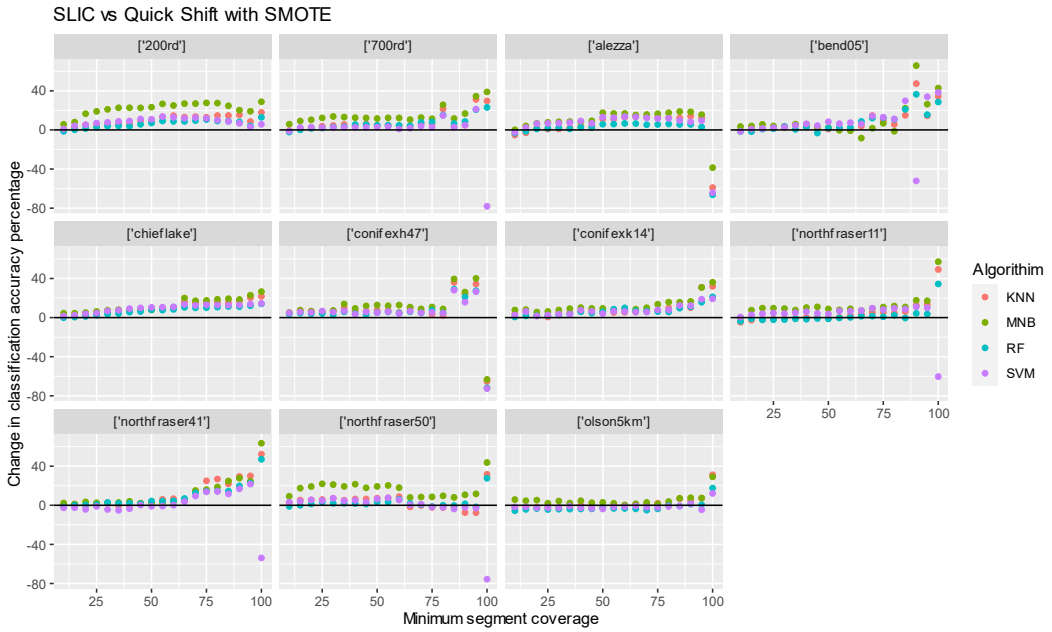


Figure 27 | Relative Accuracy of SLIC over QuickShift on sites.

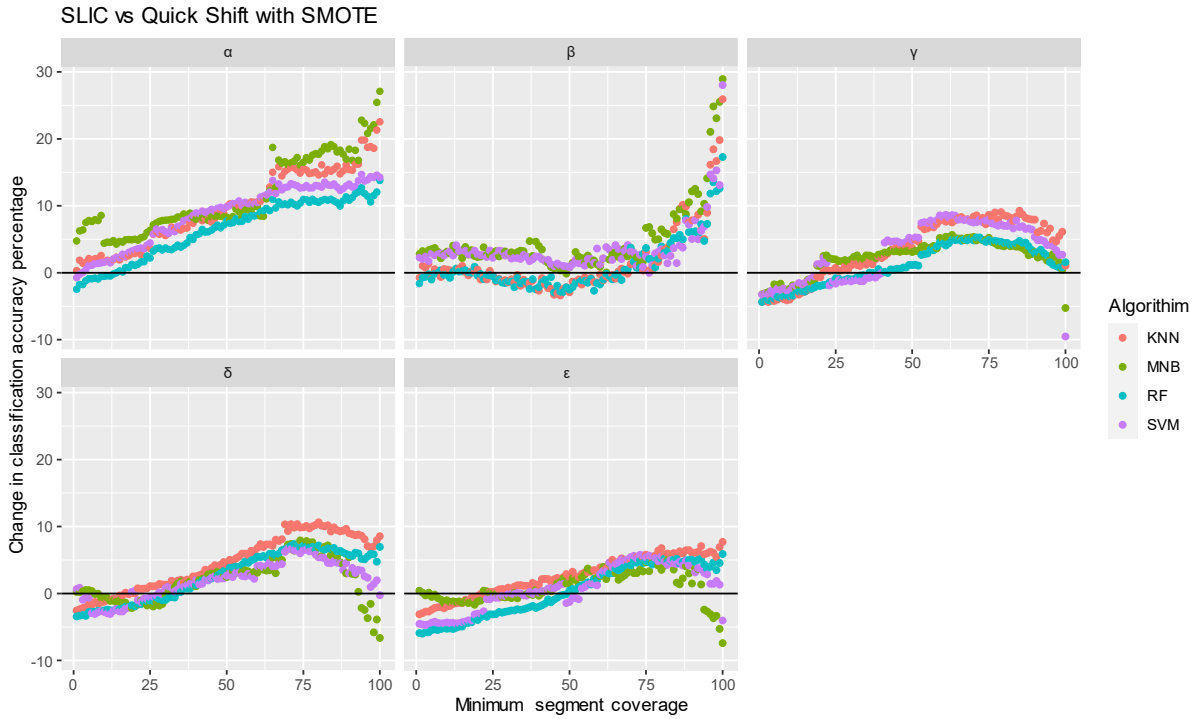


Figure 28 | Relative Accuracy of SLIC over Quick Shift on site groupings.

3.6.2.4 SMOTE

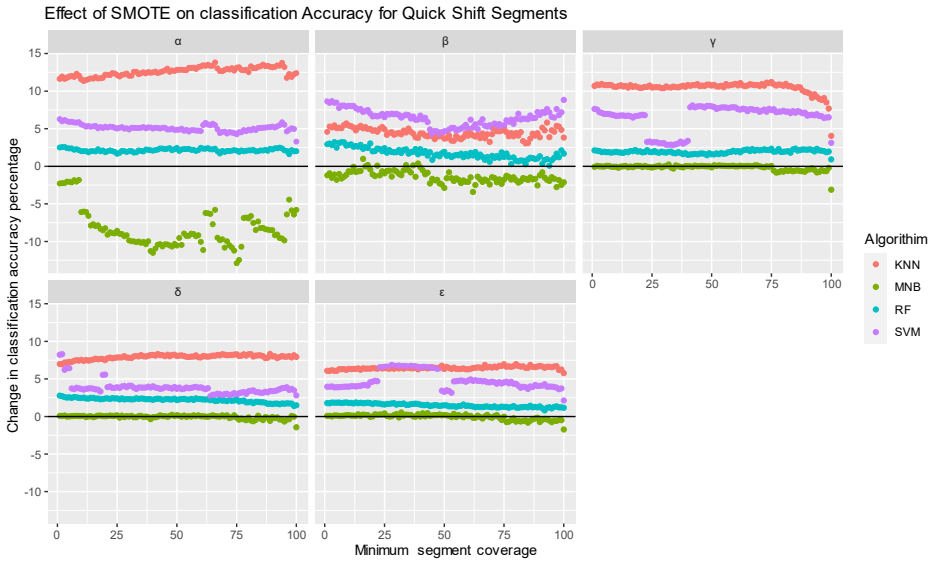


Figure 29 | Quick Shift Oversampling Change in Accuracy

Figure 30 | SLIC Oversampling Change in Accuracy

The use of SMOTE oversampling the generated segments increased accuracy in most cases (Figure 29, Figure 30), with the exception when using the MNB classifier. This is an interesting result, considering that SMOTE produced the biggest improvements for the MNB classifier when using the ground truth data as seen in Figure 27.

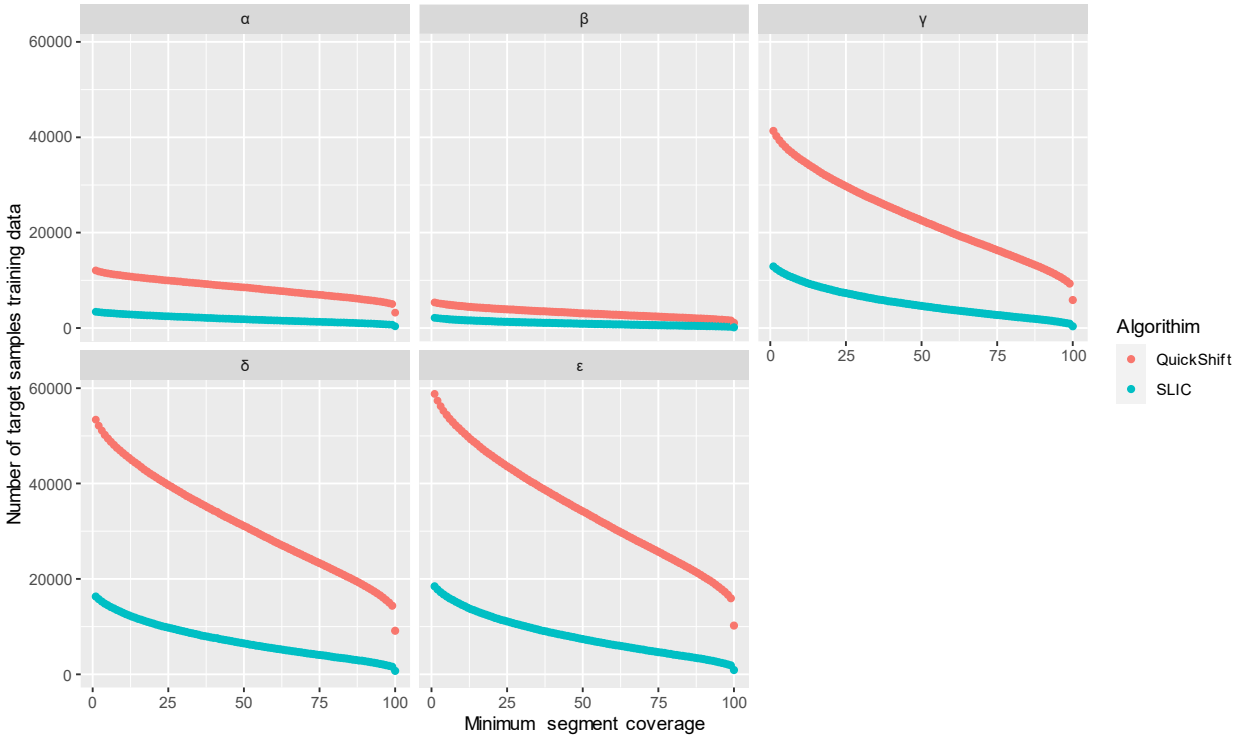


Figure 31 | Relationship between the number of samples available for training, and the minimum overlap required.

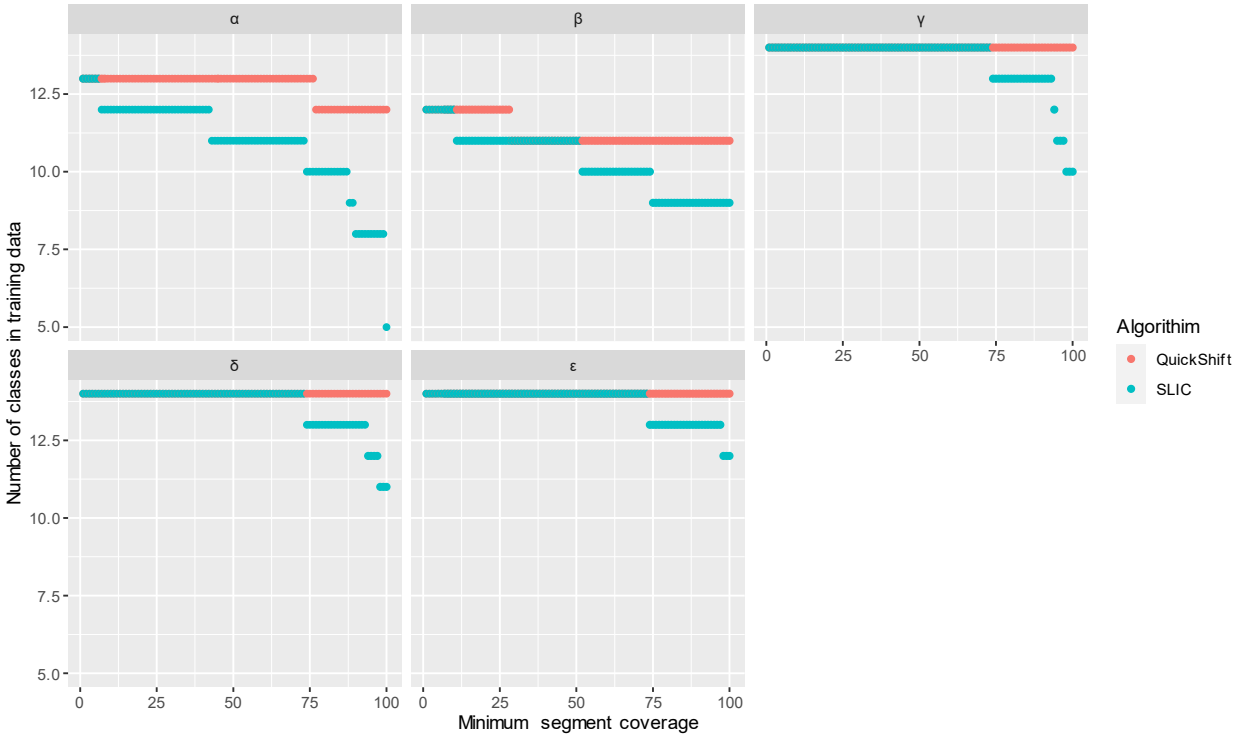


Figure 32 | Relationship between the number of classes present in training data, and the minimum overlap required.

3.6.3 Shannon's Diversity Index

When looking at comparisons of classification accuracies, and why some sites perform better than others, one potential theory to look at is how well distributed are the samples being trained with. Shannon's Diversity Index (Shannon, 1948) is one such method for measuring the diversity of sample sizes, where higher numbers represent a more unbalanced dataset. Shannon's Diversity Index was calculated for both individual sites Table 24, as well as the groups of sites tested Table 25.

Table 24 | Shannon's Diversity Index of sample counts by site

Diversity of Sample Sizes											
	200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
Truth	2.260	2.105	2.066	1.680	2.364	2.563	1.881	1.761	2.424	1.569	1.716
QuickShift	2.060	2.045	2.292	1.696	2.156	2.270	1.297	1.665	2.082	1.419	1.749
SLIC	1.916	1.937	2.220	1.569	2.031	2.173	1.231	1.663	1.996	1.337	1.720

Table 25 | Shannon's Diversity Index of sample counts by site groups

Diversity of Sample Sizes					
	α	β	γ	δ	ε
Truth	2.364	2.480	2.598	2.720	2.660
QuickShift	2.156	2.342	2.596	2.646	2.660
SLIC	2.031	2.295	2.563	2.578	2.649

These diversity indexes can then be compared to both the accuracy of sites Table 26, Table 28, as well as the effects of SMOTE on those accuracies Table 27, Table 29.

Table 26 | Correlation of Shannon's Diversity Index and classification accuracy of individual sites

Algorithm	Correlation
RF	-0.53937
SVM	-0.58521
MNB	-0.19679
KNN	-0.41848

Table 27 | Correlation of Shannon's Diversity Index and SMOTS effects on classification accuracy of individual sites

Algorithm	Correlation
RF	0.000663
SVM	0.224732
MNB	-0.30708
KNN	-0.00788

Table 28 | Correlation of Shannon's Diversity Index and classification accuracy of site groups

Algorithm	Correlation
RF	0.854157
SVM	0.783865
MNB	0.536354
KNN	0.536354

Table 29 | Correlation of Shannon's Diversity Index and SMOTS effects on classification accuracy of site groups

Algorithm	Correlation
RF	0.329936
SVM	0.112355
MNB	0.500587
KNN	-0.07188

From Table 26, a negative correlation between Shannon's Diversity Index, and classification accuracy can be observed suggesting that as sample sizes become more unbalanced accuracy decreases, which is an intuitive solution. However, that is somewhat countered by positive correlation when looking at groups of sites in Table 28, there is not a good explanation for this other than the different groupings of sites may be impacting this outside of the diversity of the sites.

When looking at the correlations with the effects of SMOTE on the data Table 27; does not provide a clear correlation with SVM being positive, and MNB being negative. Table 29 on the other hand does show a consistent high correlation, which shows higher site diversity leads to higher effectiveness of SMOTE at improving accuracy.

3.7 Results Synthesis

3.7.1 How should data be segmented?

While this chapter is focused on classification rather than the segmentation, it must be noted that it is not possible to completely decouple these processes. Into the classification process carried further were the two highest ranked approaches to segmentation. When looking at Figure 27 and Figure 28 graphs with data above the 0 line have better performance for SLIC, while those points below the line Quick Shift performed better. When looking at these figures we can see that SLIC generally performs better than Quick Shift, with a few outliers such as northfraser41 as well as groups β and ε where improvement was not seen until coverages were above 50%; and olson5km where there was little difference in performance. This is unexpected based upon the results of the segmentation indicating Quick Shift having marginally better scores and stability.

3.7.2 Should SMOTE resampling be used?

The use of SMOTE provides generally positive results and thus the general recommendation is that SMOTE is used. The notable exceptions to this can be seen in Figure 29 and Figure 30 are that MNB performs poorly with SMOTE applied, as well as limited effects on SLIC segments with very high coverage and use of the Random Forest Classifier.

3.7.3 Which algorithm should be used for training?

The two algorithms with the highest performance were Random Forest and Support Vector Machines, as can be seen in Figure 25 and Figure 26. Of interest in these figures is that Support Vector Machines performed better on most of the single site tests, while Random Forest performs better on groupings of sites.

3.7.4 Are trained models re-usable?

The results strongly suggest that the current models are not re-usable as each site presents differences in species relative reflectance; looking at Figure 22 we can see for example that 200rd, ChiefLake and Olson5km all have very low reflectance values; without a way of anchoring all data to a relative point moving trained models may be difficult.

3.8 Discussion

Evaluating multiple species accuracy will be based on number of species and their relative proportions. Accuracy decreases as number of species increases; conversely it will increase as the dataset becomes imbalanced. A simple example of this is in the most trivial case where the algorithm simply says all samples are of the most common class, we might expect 50% accuracy for two classes, and 33% accuracy for three classes; however, if 80% of the samples were class 2, we could classify all data as class 2 producing an overall accuracy of 80%.

In practical application looking at the Chief Lake site the leading species was Pine with a 23% representation, however classification accuracies just over 70% were achieved, this demonstrates that the ability to classify tree species is greater than a random distribution.

It must also be noted that the accuracies of this work show lower accuracy rates than other previous work. For example (Csillik et al., 2018) demonstrated a greater than 96% accuracy on the identification of citrus trees; this work in comparison is working with a very diverse ecosystem including more species of trees, understory, and mixture of tree sizes. It is the hope of this paper that even if the results have room for improvement that a methodology for comparison of approaches to segmentation and classification has been demonstrated that would allow for future work to continue to build upon testing more algorithms, with larger datasets. As the fields of Computer Vison and Machine Learning continue to develop a systematic approach to evaluating the effectiveness of new tools in comparison to existing tools will prove valuable for those seeking more comprehensive management of natural resources.

The availability of training data must also be considered when training models. While this research was conducted with what may appear to be a very large dataset; the lack of balance in the data also makes the sample size small in many respects. Many of the under-represented

species were simply removed from classification due to not having enough samples at a given site to be able to split into K-folds, and even those that were included many were very underrepresented potentially leading to low training accuracy. Operationally this introduces the need to be thoughtful during the collection of training data, ensuring training sites both contain representation of species of interest, and that these samples are adequately documented. While more training data is always better, this researcher would suggest targeting a minimum of 100 samples of target species, if possible, per site; this guidance is tempered with the realization this may not be possible due to the high cost of collecting training data.

3.8.1 Considerations of algorithms tested

This chapter reviewed 4 common classification algorithms; while this does not present an exhaustive search of available methods, the algorithms were chosen due to their popularity of use and being mechanically different with the goal of highlighting differences between approaches.

Overall, the highest performance came from SVM and RF. KNN and MNB are faster and more naive algorithms, providing a baseline of very common, and relatively accessible tools.

K-Nearest Neighbors is the simplest of the algorithms used and is commonly used in remote sensing projects due to the simplicity and speed of implementation. Due to the nature of the algorithm, it is relatively robust in highly dimensional data provided that each dimension has either a single representation, or the number of clusters is sufficiently high to capture clusters for each representation of any given class. However, at the same time this simplistic nature also makes the algorithm poor for use with noisy data; especially when the values of a given dimension have overlap between classes.

The Multinominal Naïve Bayes Probabilistic Classifier works by treating each dimension as an independent variable (Murphy, 2006); this can be very advantageous if individual dimensions are particularly noisy, allowing more consistent dimensions to take the lead. This type of classifier is more commonly applied to text classifications than imagery. Both KNN and MNB present issues for this classification problem in that they are looking for mean values on each dimension; something that is particularly challenging to achieve in the case where the calibration of reflectance is not consistent between sites.

State Vector Machines provide a compelling option for the classification of tree species given the collected dataset as they are very well suited to handling data with very high dimensionality, as well as filtering out redundant dimensions. Additionally, the mechanics of calculating dividing lines maximize the distance between clusters as opposed to placing nearest to the mean of a cluster while subtle can make the algorithm much more robust to certain types of noise in the data. However, this is challenged by the continuing theme that SVM expects all elements of a class to present in the same way; a condition that was not true due to the pool calibration abilities of the datasets used in this research. It is believed that should a way of securing better calibration be achieved this could be the preferred method of segmentation. Finally, SVM, like the previous two algorithms has a very fast training time; this is beneficial as over time more training data could be collected and used for efficient retraining of models.

Random Forest was the final model used for training; in comparison to SVM in brief RF has worse handling of high dimensionality, but better handling for multiple representations. Among the methods shown RF is unique in its ability to handle dependent dimensions; this is particularly useful as it can allow the model to respond to changes in exposure through the nested decision trees. Such that a high value on dimension A might lead to a higher threshold on

dimension B and vice versa thus providing a basic correction for image exposure. It should however be remembered that having multiple representations of classes makes the training set effectively smaller as each representation will contain only a portion of the training data for that class.

3.8.2 Future work and areas of improvement

The work here demonstrates the potential of machine learning to be used in a production environment especially with further development on the primary limitations. The first limitation shown is that there needs to be more control of site variability, and it is hypothesized this could be addressed with better calibration of the multispectral cameras used. This has the potential to be as simple as using the more modern revisions of the downwelling light sensors. With the calibration challenges better addressed there would be further room to examine the impacts of the phenological cycle of plants and general soil conditions.

The other primary limitation demonstrated is sample sizes; while the set of training data did include 40,257 labeled trees, due to the highly imbalanced nature of the data many species did not contain sufficient training data. It is proposed that to make the methods presented more robust effort would need to be placed into collecting samples from sites with better representation of the rarer species; it would be ideal to have a minimum of at least 100 of each species with a target closer to 1000 or more.

It is the hope that this research presents the framework for evaluating accuracy of various tools that will allow for future research to continually try new algorithms and compare against existing solutions to provide a path for continual improvement in accuracy as new machine learning techniques are developed and better imagery is available.

Finally, it is believed that the development of technologies whether higher resolution cameras or something like RPAS based LiDAR allowing for a finer scale of structural attributes to be captured may greatly improve these results. Other research papers were reviewed that were able to achieve higher accuracies (Brandtberg, 2007; Csillik et al., 2018). One notable difference in these papers was a focus on mature trees that were easier to discriminate from the understory; thus, removing the need to classify between small deciduous trees and grasses for example.

As more research is conducted the models produced will be useful for ecologists and land managers to provide more detailed information about the trees present in terms of species, number, and size. This information could then provide for new ecological questions to be asked in terms of this more detailed inventory not currently possible. This also allows for positive feedback whereas more data is collected models can be continuously trained to be more accurate.

Chapter 4

Synthesis

4.1 Introduction

The research presented above presents a framework for identification of early seral vegetation. While at present these results provide an accuracy of 52%-83% depending on site and what species are targeted for classification. This accuracy is respectable on its own given the unique challenges of classifying relatively small targets. The methods presented are offered as a framework that will be able to continually integrate with new technologies; and see further increases to quality of results as technology advances.

4.2 Data Collection

As effective machine learning requires that a high-quality dataset be used for training and implementation, this research has also provided the opportunity to think of strategies for effective data collection moving forward.

The first group of suggestions is around what data should be collected, as well as when and where it should be collected. This research had some outliers in terms of some sites not training well with the other sites, there are several possible explanations for this, differences in the phenological cycle, differences in site conditions such as soil nutrients, or inconsistent distribution of species from site to site.

In terms of the phenological cycle the importance of this can be most easily seen as this classification is based on the reflectance of light in various spectrums, something that very visibly changes from summer to fall, and to a greater extent on deciduous trees, during leaf off. As such it would be recommended that future work either develop controls for differences in

phenological cycle such as multiple training sets based on the position in the cycle, or taking measures to ensure all data is captured at the same point within the cycle. Both options do however provide operational challenges as research would require either much larger training sites to obtain sufficient samples, or work with very short and potentially unpredictable data collection windows.

Site condition is another element that has not been able to be adequately explored in this research. As we know that work is being done to monitor forest health using RPAS Multispectral imagery (Fraser and Congalton, 2021) it would then also follow that a trees health will alter its appearance and thus how a given class is represented in machine learning models. An additional complication here is that preliminary research suggests that some disease such as bark beetle infestation can be detected fast than traditional methods using multispectral imaging (Bárta et al., 2021), thus it may not even be clear at the time of analysis if the trees are healthy or not. In terms of site conditions effects on forest health this should be easier to account for by looking at available moisture and soil nutrients as a variable within the dataset.

Finally, in terms of training models, training models on one site then applying to another site adverse impacts on classification accuracy could be observed due to differences in species distribution between these sites. A model may work very well with the set of species on one site but could then be moved to a new site where previously unrepresented species are more plentiful, the model may then have a bias towards not classifying these species due to the imbalanced training data used. A method for selecting training data that strives to balance the training data is more useful than one that focuses more on spatial distribution of sites.

The quality of data collected will also serve to have impacts on how effective classification can be. This accounts for both collection of the imagery, as well as the validation surveys.

For imagery collection it is important that images be sharp, use the same spectral bands on all collected data for which a model is to be applied and make best efforts to control differences in sunlight. The most prominent issues found with data during this research were variations in sharpness within individual sites, with the hypothesis that this was caused by motion blur from tail winds, then getting sharper images on the return path where a slower relative ground speed is present. Some sites also did include visible shadow from changing lighting conditions, high quality light sensors to account for this would be beneficial.

For the survey component, the first and most obvious point is that accuracy matters a bias in segments of trees will affect model accuracy. Additionally, while not observed as an issue in this research it should be a priority to ensure that all training data is correctly labeled. Another issue to be aware of is the need to classify everything in the imagery collected. Most classification algorithms do not have a default option for ‘unknown’ instead opting to classify everything as what it is most comparable to in the training data. As such in initial testing the roads were being classified as trees despite being very different in appearance; this was solved after the fact by adding a bare earth class to the training data. Caution is urged here at the data collection stage to identify all types of vegetation and not just target species as adding to the training set after the fact may not be an option, and non-target vegetation may exhibit similar characteristics to target species.

These suggestions can be reduced to the summary that training data must include representations for all data that the model is to be applied to, in all its aspects including species,

health and phenological cycle. In order to account for reductions in accuracy due to highly imbalanced data, at the time of data collection either training data collected in quantities sufficient for training, or site selection must be done to limit appearance underrepresented species of little consequence from appearing in the training data. And then data quality must be ensured, any errors in training data will lead to flawed models regardless of how good the model is. While minor errors can be mitigated simply by having large data sets, attention to accuracy is critical at the training stage.

4.3 Evaluation of Segmentation

When looking at the accuracy of segmentation, it is very hard to develop a comparison to other work, as other research did not reveal any consistent methods for measuring the quality of segmentation. However, this paper does propose some possible equations to address this. To be clear it is not the concept of rating segmentation that is novel, but rather presenting the quality as a single score. While metrics such as false segments, and false merges are common, they pose a problem for automated optimization processes due to the ambiguity of which result is better based upon multiple competing metrics.

While there is undoubtedly room for improvement upon the algorithms presented here in terms of fine tuning their effectiveness. The segments produced were usable for producing viable classification in the following stages of analysis. And potentially more importantly a framework has been produced for the rapid evaluation of other algorithms. The core idea here is that by producing standardized testing frameworks as research advances segmentation methods novel approaches should be able to be calculated and examined against previous results to determine an optimal method that can evolve over time. This also allows for automated retraining and testing as datasets grow.

It was found that two of the proposed metrics Metric 1 *MAX (ARP - FM)* and Metric 3 *MAX (-FM - 0.5 * FS))* were the most useful. Both focusing on reducing the number of false merges; with the difference being metric 1 focuses on the number of pixels correctly segmented, where Metric 2 focuses on the number of false splits. In this case Metric 1 in theory could be a better representation of land cover, while Metric 3 may be a better representation of individual stems. As such while the are industry trends to wards per stem forest management (Gray et al., 2021; Seidl et al., 2012) it would seem logical to focus on the refinement of metric 3 as an ongoing measurement of segmentation quality.

All the algorithms tested produced an over segmentation based upon the developed scoring metrics. Given this process is based solely on reflectance data and not texture or height of trees this is likely as good as could be expected due to the often-ambiguous boundaries between small trees and the understory. There are many successful applications of individual tree segmentation (Jing et al., 2012; Morsdorf et al., 2003; Zhang et al., 2015) these works all work with more mature trees that can be more easily discerned from noise in the data. One issue with the early seral vegetation is that the elevation models produced due not accurately reflect heights of vegetation after noise filtering is completed, this is due to both the resolution used, and movement in vegetation cause by even a small breeze.

4.4 Effectiveness of Classification

Measuring the effectiveness of classification can be a complex topic hard to reduce accuracy to a single number. Effectiveness will vary based upon how well balanced the data is, the quality of imagery collected, and quality of segmentation and training labels provided. When reviewing the results below it is important to remember that the data below represents real world

data and not all of these variables can be controlled, thus producing some natural variance from site to site in terms of classification accuracy.

Table 30 |Classification accuracy of target species using SVM, with top row showing total number of training samples.

Site	200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
Total Samples	11529	3774	5649	1521	12861	1398	1077	8331	1677	1107	1862
SVM	65%	83%	74%	78%	65%	68%	80%	82%	52%	79%	63%
SVM SMOTE	65%	85%	71%	77%	64%	74%	88%	78%	NA	83%	59%

Table 31 |Classification accuracy of target species using SVM, with top row showing percentage of samples that are target species.

Site	200rd	700rd	Alezza	Bend05km	ChiefLake	ConifexH47	ConifexK14	NorthFraser11	NorthFraser41	NorthFraser50	Olson5km
% Target Samples	18%	52%	37%	42%	12%	19%	48%	42%	16%	42%	20%
Target SVM	61%	91%	69%	64%	44%	37%	81%	79%	18%	71%	43%
Target SVM SMOTE	59%	96%	68%	68%	57%	43%	83%	77%	NA	71%	49%

Looking at Table 30 above we can see that simply having a larger training dataset does not ensure better accuracy; however, Table 31 does show that when attempting to classify target species there is a very strong correlation (0.929) between the prevalence of target species and ability to classify those species. One possible explanation is that the variance in number of samples for each species is very important as an example of this if Shannon's Diversity index is calculated for the number of each sample type per site as in Table 24 we can see a correlation of -0.585 with SVM's classification accuracy, this correlation suggests that sites with less diversity have a better classification accuracy. It is still an open question if the need for homogeneity is only in the training data or if more uniform sites would classify better even with a homogenous training set. SMOTE did increase the accuracy of classification on average, suggesting that it is just the training that needs to be homogenous, but with so many species being extremely underrepresented this is hard to discern from the current dataset.

For those sites able to achieve the 78% and higher accuracies RPAS imagery is a relatively effective method for classification of entire sites. In comparison the other options for

classification would be manual surveys of small plots (typically a circle with a 3.99m radius, hereafter referred to as a 3.99 plot) (Raymer, 2001) and interpolating over the entire land base which would still contain errors; or the prohibitively expensive manual survey of every tree which while would have 100% accuracy could realistically be accomplished. Another interesting avenue for future research would be to look at comparisons of RPAS classification to 3.99 plots, as well as looking at a potential for a hybrid approach.

4.5 What information do these results provide to forest managers?

Just as important as developing methods for data collection and classification is to develop an understanding of what the information can be used for. The methods presented in this research classify ground cover by species type, which does have an important distinction from classifying trees. Due to the relatively high number of false splits allowed at the segmentation stage this is not a method for counting trees, only how much area each species covers. Due to the nature of early seral vegetations small size and lack of defined edges in RPAS imagery, the orthomosaics produced from a photogrammetry workflow result in an inability to separate from grass and brush in the understory. Further the small size of early seral vegetation, results in the height being filtered out as noise in the photogrammetry workflow resulting in a lack of data on the vertical dimension; that could be used derive tree heights or total volume of biomass.

The information collected is however still very useful from an ecological and forest management perspective as it does allow for determining a general idea of the forest makeup by area; if this is combined with knowledge of how recently logging has occurred it may be possible to get good estimates on food supply for ungulates as an example. This information is also useful for monitoring changes over time; RPAS is a relatively cost-effective method for collecting data

and changes in surface cover vegetation over time can provide indicators of how the forest is developing.

This methodology also allows for a more holistic data collection, while in some ways the need to classify every type of vegetation to produce accurate models could be a disadvantage in some cases. The collection of information on every type of vegetation instead of just merchantable timber for example could lead to data sets that can be used for multiple aspects of forest management, including not just how profitable logging could be but also what other ecosystem services are provided, and monitoring how much biodiversity is present in our forests.

4.6 Framework for evaluation

The technologies used in remote sensing continue to advance; and have done so since the beginning of this research. As new approaches are implemented into working processes, they will continually have room for improvement with modern advancements. This demonstrates the need for processes in place that can be used to evaluate these new advancements in a timely manner providing for faster integrations. This research should not be understood as a prescriptive method for the classification; but rather a framework for considering how the tools can be applied.

While working on this research early on it became clear there is not a simple and definitive way for saying how good the segmentation of an image is. One of the primary factors for this is that different problems may be more affected by different types of errors. The optimal segmentation is the segmentation that produces the highest accuracy classification. The most trivial solution to this problem would be to train models-based segmentations were the parameters used to generate the segments spanned the range of possibilities. However, given the computational time needed to train machine learning algorithms, combining all the potential

parameters of both segmentation and classification would be nearly impossible to compute with current technology. For this reason, these two steps were decoupled.

Providing a quantitative metric for scoring the segmentation is a critical piece to producing automated training pipelines. As spatial and spectral resolutions change with different sensors different hyper-parameters are needed to produce optimal results. Determining a consistent way of weighing the importance of the different error measurements such as false splits and false merges is essential for performing fair comparisons between sensors.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2010. SLIC Superpixels [WWW Document]. Infoscience. URL http://infoscience.epfl.ch/record/149300 (accessed 3.21.21).

Arganda-Carreras, I., Turaga, S.C., Berger, D.R., Cireşan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J., Laptev, D., Dwivedi, S., Buhmann, J.M., Liu, T., Seyedhosseini, M., Tasdizen, T., Kamentsky, L., Burget, R., Uher, V., Tan, X., Sun, C., Pham, T.D., Bas, E., Uzunbas, M.G., Cardona, A., Schindelin, J., Seung, H.S., 2015. Crowdsourcing the creation of image segmentation algorithms for connectomics. Front. Neuroanat. 9. https://doi.org/10.3389/fnana.2015.00142

Baleshta, K.E., Simard, S.W., Roach, W.J., 2015. Effects of thinning paper birch on conifer productivity and understory plant diversity. Scandinavian Journal of Forest Research 0, 1–47. https://doi.org/10.1080/02827581.2015.1048715

Bárta, V., Lukeš, P., Homolová, L., 2021. Early detection of bark beetle infestation in Norway spruce forests of Central Europe using Sentinel-2. International Journal of Applied Earth Observation and Geoinformation 100, 102335. https://doi.org/10.1016/j.jag.2021.102335

Beaudry, L., Coupe, R., Delong, C., Pojar, J., 1999. Plant Indicator Guide for Northern British Columbia: Boreal, Sub-Boreal, and Subalpine Biogeoclimatic Zones: BWBS, SBS, SBPS, and northern ESSF [WWW Document]. URL https://www.for.gov.bc.ca/hfd/pubs/docs/Lmh/Lmh46.htm (accessed 4.1.24).

Biswas, D., Su, H., Wang, C., Blankenship, J., Stevanovic, A., 2017. An Automatic Car Counting System Using OverFeat Framework. Sensors 17, 1535. https://doi.org/10.3390/s17071535

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92. Presented at the fifth annual workshop, ACM Press, Pittsburgh, Pennsylvania, United States, pp. 144–152. https://doi.org/10.1145/130385.130401

Boukherroub, T., LeBel, L., Ruiz, A., 2017. A framework for sustainable forest resource allocation: A Canadian case study. Omega, New Research Frontiers in Sustainability 66, 224–235. https://doi.org/10.1016/j.omega.2015.10.011

Brandtberg, T., 2007. Classifying individual tree species under leaf-off and leaf-on conditions using airborne lidar. ISPRS Journal of Photogrammetry and Remote Sensing 61, 325–340. https://doi.org/10.1016/j.isprsjprs.2006.10.006

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

British Columbia, Forest Inventory and Monitoring Program, British Columbia, Resources Information Standards Committee, British Columbia, Forest Analysis and Inventory Branch, British Columbia, Ministry of Forests and Range, Growth and Yield Program, 2007. Forest Inventory and Monitoring Program: growth and yield standards and procedures. Resources Information Standards Committee, Victoria, B.C.

Brown, G.S., Rettie, W.J., Brooks, R.J., Mallory, F.F., 2007. Predicting the impacts of forest management on woodland caribou habitat suitability in black spruce boreal forest. Forest Ecology and Management 245, 137–147. https://doi.org/10.1016/j.foreco.2007.04.016

Bugmann, H.K.M., Yan, X.D., Sykes, M.T., Martin, P., Lindner, M., Desanker, P.V., Cumming, S.G., 1996. A comparison of forest gap models: Model structure and behaviour. Climatic Change 34, 289–313.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321–357. https://doi.org/10.1613/jair.953

Comaniciu, D., Meer, P., 2002. Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619. https://doi.org/10.1109/34.1000236

Coops, N.C., Tompaski, P., Nijland, W., Rickbeil, G.J.M., Nielsen, S.E., Bater, C.W., Stadt, J.J., 2016. A forest structure habitat index based on airborne laser scanning data. Ecological Indicators 67, 346–357. https://doi.org/10.1016/j.ecolind.2016.02.057

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach Learn 20, 273–297. https://doi.org/10.1007/BF00994018

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R.V., Paruelo, J., Raskin, R.G., Sutton, P., van den Belt, M., 1997. The value of the world's ecosystem services and natural capital. Nature 387, 253–260. https://doi.org/10.1038/387253a0

Cristianini, N., Shawe-Taylor, J., 2000. An introduction to support vector machines: and other kernel-based learning methods. Cambridge University Press, Cambridge ; New York.

Csillik, O., Cherbini, J., Johnson, R., Lyons, A., Kelly, M., 2018. Identification of Citrus Trees from Unmanned Aerial Vehicle Imagery Using Convolutional Neural Networks. Drones 2, 39. https://doi.org/10.3390/drones2040039

D'Amato, A.W., Bradford, J.B., Fraver, S., Palik, B.J., 2011. Forest management for mitigation and adaptation to climate change: Insights from long-term silviculture experiments. Forest Ecology and Management 262, 803–816. https://doi.org/10.1016/j.foreco.2011.05.014

Dhar, A., 2013. Birch (Betula papyrifera) white spruce (Picea glauca) interactions in mixedwood stands: implications for management. Journal of Forest Science 59, 137–149.

Duan, K.-B., Keerthi, S.S., 2005. Which Is the Best Multiclass SVM Method? An Empirical Study, in: Oza, N.C., Polikar, R., Kittler, J., Roli, F. (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 278–285. https://doi.org/10.1007/11494683_28

Dumitru, C.O., Andrei, V., Schwarz, G., Datcu, M., 2019. MACHINE LEARNING FOR SEA ICE MONITORING FROM SATELLITES. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLII-2/W16, 83–89. https://doi.org/10.5194/isprs-archives-XLII-2-W16-83-2019

Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition 58, 121–134. https://doi.org/10.1016/j.patcog.2016.03.028

Evangelista, P.F., Embrechts, M.J., Szymanski, B.K., 2006. Taming the Curse of Dimensionality in Kernels and Novelty Detection, in: Abraham, A., de Baets, B., Köppen, M., Nickolay, B. (Eds.), Applied Soft Computing Technologies: The Challenge of Complexity, Advances in Soft Computing. Springer, Berlin, Heidelberg, pp. 425–438. https://doi.org/10.1007/3-540-31662-0_33

Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient Graph-Based Image Segmentation. International Journal of Computer Vision 59, 167–181. https://doi.org/10.1023/B:VISI.0000022288.19776.77

Forest Health Aerial Survey Manual, 2012. 65.

Fradkov, A.L., 2020. Early History of Machine Learning. IFAC-PapersOnLine 53, 1385–1390. https://doi.org/10.1016/j.ifacol.2020.12.1888
Fraser, B.T., Congalton, R.G., 2021. Monitoring Fine-Scale Forest Health Using Unmanned Aerial Systems (UAS) Multispectral Models. Remote Sensing 13, 4873. https://doi.org/10.3390/rs13234873
Friedman, J.H., 1997. On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality 23.
Government of British Columbia, n.d. VRI - HISTORICAL Vegetation Resource Inventory (2002 - 2022) - Open Government Portal [WWW Document]. URL https://open.canada.ca/data/en/dataset/02dba161-fdb7-48ae-a4bb-bd6ef017c36d (accessed 12.9.24).
Gray, A.N., McIntosh, A.C.S., Garman, S.L., Shettles, M.A., 2021. Predicting canopy cover of diverse forest types from individual tree measurements. Forest Ecology and Management 501, 119682. https://doi.org/10.1016/j.foreco.2021.119682
Gupta, D.D., Nair, L.M., 2005. IMPROVING OCR BY EFFECTIVE PRE-PROCESSING AND SEGMENTATION FOR DEVANAGIRI SCRIPT:A QUANTIFIED STUDY. . Vol. 52.
Holmgren, J., Persson, Å., Söderman, U., 2008. Species identification of individual trees by combining high resolution LiDAR data with multi-spectral images. International Journal of Remote Sensing 29, 1537–1552. https://doi.org/10.1080/01431160701736471
Jing, L., Hu, B., Noland, T., Li, J., 2012. An individual tree crown delineation method based on multi-scale segmentation of imagery. ISPRS Journal of Photogrammetry and Remote Sensing 70, 88–98. https://doi.org/10.1016/j.isprsjprs.2012.04.003
Lemprière, T.C., Kurz, W.A., Hogg, E.H., Schmoll, C., Rampley, G.J., Yemshanov, D., McKenney, D.W., Gilsenan, R., Beatch, A., Blain, D., Bhatti, J.S., Krcmar, E., 2013. Canadian boreal forests and climate change mitigation. Environ. Rev. 21, 293–321. https://doi.org/10.1139/er-2013-0039
McGrath, M.J., Luyssaert, S., Meyfroidt, P., Kaplan, J.O., Bürgi, M., Chen, Y., Erb, K., Gimmi, U., McInerney, D., Naudts, K., Otto, J., Pasztor, F., Ryder, J., Schelhaas, M.-J., Valade, A., 2015. Reconstructing European forest management from 1600 to 2010. Biogeosciences 12, 4291–4316. https://doi.org/10.5194/bg-12-4291-2015
Meilă, M., 2007. Comparing clusterings—an information based distance. Journal of Multivariate Analysis 98, 873–895. https://doi.org/10.1016/j.jmva.2006.11.013
Millennium Ecosystem Assessment, 2005. Ecosystems and Human Well-Being: Synthesis.
Morsdorf, F., Meier, E., Allgower, B., 2003. CLUSTERING IN AIRBORNE LASER SCANNING RAW DATA FOR SEGMENTATION OF SINGLE TREES 7.
Murphy, K.P., 2006. Naive Bayes classifiers.
Nowak, D.J., Hirabayashi, S., Bodine, A., Greenfield, E., 2014. Tree and forest effects on air quality and human health in the United States. Environmental Pollution 193, 119–129. https://doi.org/10.1016/j.envpol.2014.05.028
Oettel, J., Lapin, K., 2021. Linking forest management and biodiversity indicators to strengthen sustainable forest management in Europe. Ecological Indicators 122, 107275. https://doi.org/10.1016/j.ecolind.2020.107275
Pearce, D.W., 2001. The Economic Value of Forest Ecosystems. Ecosystem Health 7, 284–296. https://doi.org/10.1046/j.1526-0992.2001.01037.x
Pettis, B.T., 2023. reCAPTCHA challenges and the production of the ideal web user. Convergence 29, 886–900. https://doi.org/10.1177/13548565221145449

Pitt, D.G., Comeau, P.G., Parker, W.C., MacIsaac, D., McPherson, S., Hoepting, M.K., Stinson, A., Mihajlovich, M., 2010. Early vegetation control for the regeneration of a single-cohort, intimate mixture of white spruce and trembling aspen on upland boreal sites. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere 40, 549–564. https://doi.org/10.1139/X10-012
Pollard, D., 1982. Quantization and the method of k-means. IEEE Transactions on Information Theory 28, 199–205. https://doi.org/10.1109/TIT.1982.1056481
Raymer, B., 2001. Juvenile Spacing Quality Inspection. Forest Renewal BC.
Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767 [cs].
Schuster, C., Förster, M., Kleinschmit, B., 2012. Testing the red edge channel for improving land-use classifications based on high-resolution multi-spectral satellite data. International Journal of Remote Sensing 33, 5583–5599. https://doi.org/10.1080/01431161.2012.666812
Seely, H.E., 1934. AERIAL PHOTOGRAPHY IN FOREST SURVEYS. The Forestry Chronicle 10, 226–229. https://doi.org/10.5558/tfc10226-4
Seidl, R., Rammer, W., Scheller, R.M., Spies, T.A., 2012. An individual-based process model to simulate landscape-scale forest ecosystem dynamics. Ecological Modelling 231, 87–100. https://doi.org/10.1016/j.ecolmodel.2012.02.015
Seydi, S.T., Saeidi, V., Kalantar, B., Ueda, N., Halin, A.A., 2022. Fire-Net: A Deep Learning Framework for Active Forest Fire Detection. Journal of Sensors 2022, e8044390. https://doi.org/10.1155/2022/8044390
Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
Takumi, K., Watanabe, K., Ha, Q., Tejero-De-Pablos, A., Ushiku, Y., Harada, T., 2017. Multispectral Object Detection for Autonomous Vehicles, in: Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Thematic Workshops '17. Association for Computing Machinery, New York, NY, USA, pp. 35–43. https://doi.org/10.1145/3126686.3126727
Taunk, K., De, S., Verma, S., Swetapadma, A., 2019. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification, in: 2019 International Conference on Intelligent Computing and Control Systems (ICCS). Presented at the 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747
Terry, E.L., McLellan, B.N., Watts, G.S., 2000. Winter habitat ecology of mountain caribou in relation to forest management. Journal of Applied Ecology 37, 589–602. https://doi.org/10.1046/j.1365-2664.2000.00523.x
Tin Kam Ho, 1995. Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition. Presented at the Proceedings of 3rd International Conference on Document Analysis and Recognition, pp. 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994
Tompalski, P., Coops, N.C., White, J.C., Wulder, M.A., Pickell, P.D., 2015. Estimating Forest Site Productivity Using Airborne Laser Scanning Data and Landsat Time Series. Canadian Journal of Remote Sensing 41, 232–245. https://doi.org/10.1080/07038992.2015.1068686
van Leeuwen, M., Hilker, T., Coops, N.C., Frazer, G., Wulder, M.A., Newnham, G.J., Culvenor, D.S., 2011. Assessment of standing wood and fiber quality using ground and airborne

laser scanning: A review. Forest Ecology and Management 261, 1467–1478. https://doi.org/10.1016/j.foreco.2011.01.032
Vedaldi, A., Soatto, S., 2008. Quick Shift and Kernel Methods for Mode Seeking, in: Forsyth, D., Torr, P., Zisserman, A. (Eds.), Computer Vision – ECCV 2008, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 705–718. https://doi.org/10.1007/978-3-540-88693-8_52
Walsh, S.J., 1980. Coniferous tree species mapping using LANDSAT data. Remote Sensing of Environment 9, 11–26. https://doi.org/10.1016/0034-4257(80)90044-9
Weisberg, P.J., Bugmann, H., 2003. Forest dynamics and ungulate herbivory: from leaf to landscape. Forest Ecology and Management 181, 1–12.
Whitman, E., Parisien, M.-A., Price, D.T., St-Laurent, M.H., Johnson, C.J., DeLancey, E.R., Arseneault, D., Flannigan, M.D., 2017. A framework for modeling habitat quality in disturbance-prone areas demonstrated with woodland caribou and wildfire. 8, e01787. https://doi.org/10.1002/ecs2.1787
Wulder, M., 2003. EOSD Land Cover Classification Legend Report (No. V2). Victoria, B.C.
Yancho, J.M.M., Coops, N.C., Tompalski, P., Goodbody, T.R.H., Plowright, A., 2019. Fine-Scale Spatial and Spectral Clustering of UAV-Acquired Digital Aerial Photogrammetric (DAP) Point Clouds for Individual Tree Crown Detection and Segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, 4131–4148. https://doi.org/10.1109/JSTARS.2019.2942811
Zhang, C., Zhou, Y., Qiu, F., 2015. Individual Tree Segmentation from LiDAR Point Clouds for Urban Forest Inventory. Remote Sensing 7, 7892–7913. https://doi.org/10.3390/rs70607892
Zheng, G., Chen, J.M., Tian, Q.J., Ju, W.M., Xia, X.Q., 2007. Combining remote sensing imagery and forest age inventory for biomass mapping. Journal of Environmental Management, Carbon Sequestration In China's Forest Ecosystems 85, 616–623. https://doi.org/10.1016/j.jenvman.2006.07.015

Appendix

Appendix A | List of all species recorded in data collection.

Latin Name	sp_code	Common Name	Class
Amelanchier alnifolia	Aa	Saskatoon	plant
Aruncus dioicus	Ad	Goat's Beard	plant
Acer glabrum	Ag	Douglas Maple	deciduous
Alnus sp	Al	Alder	deciduous
Anaphalis margaritacea	Am	Pearly Everlasting	plant
Aralia nudicaulis	An	Wild Sarsaparilla	plant
Actaea rubra	Ar	Baneberry	plant
Arctostaphylos uva-ursi	Auu	Kinnikinnick	plant
Cornus canadensis	Cc	Bunchberry	plant
Crataegus douglasii	Cd	Black Hawthorn	plant
Castilleja miniata	Cm	Red Paintbrush	plant
Corylus cornuta	Coco	Beaked Hazelnut	plant
Cornus stolonifera	Cs	Red-osier Dogwood	plant
Disporum hookeri	Dh	Hooker's Fairybells	plant
Epilobium angustifolium	Ea	Fireweed	plant
Equisetum sp	Eq	Horsetail	plant
Geocaulon lividum	Gl	Bastard Toad-flax	plant
Heracleum lanatum	Hl	Cow-parsnip	plant
Juniperus communis	Jc	Common Juniper	plant
Lysichiton americanum	La	Skunk Cabbage	plant
Linnaea borealis	Lb	Twinflower	plant
Ledum groenlandicum	Lg	Labrador Tea	plant
Lonicera involucrata	Li	Black Twinberry	plant
Lupinus sp	Lu	Lupine	plant
Lycopodium annotinum	Lya	Stiff Clubmoss	plant
Mitella nuda	Mn	Common Mitrewort	plant
Paxistima myrsinites	Pm	Falsebox	plant
Petasites palmatus	Pp	Palmate Coltsfoot	plant
Rosa acicularis	Ra	Prickly Rose	plant
Rubus idaeus	Ri	Red Raspberry	plant
Ribes lacustre	Rl	Black Gooseberry	plant
Rubus parviflorus	Rp	Thimbleberry	plant
Symphoricarpos albus	Sa	Common Snowberry	plant
Streptopus amplexifolius	Sap	Clasping Twistedstalk	plant
Sambucus racemosa	Sar	Red Elderberry	plant
Spiraea betulifolia	Sb	Birch-leaved Spirea	plant
Shepherdia canadensis	Sc	Soopolallie	plant

Spiraea douglasii ssp. menziesii	Sd	Douglas Spirea (pink / hardhack)	plant
Smilacina racemosa	Sr	False Solomon's-seal	plant
Sorbus sp	Ss	Mountain-ash	plant
Viburnum edule	Ve	Highbush-cranberry	plant
Vaccinium membranaceum	Vm	Black Huckleberry	plant
Veratrum viride	Vv	Indian Hellebore	plant
Salix sp	W	Willow	deciduous
Populus balsamifera	Ac	Cottonwood	deciduous
Populus tremuloides	At	Trembling Aspen	deciduous
Abies lasiocarpa	Bl	Subalpine Fir	conifer
Betula papyrifera	Ep	Paper Birch	deciduous
Pseudotsuga menziesii	Fd	Douglas-fir	conifer
Pinus contorta	Pl	Lodgepole Pine	conifer
Picea glauca x engelmannii	Sx	Hybrid White Spruce	conifer

Latin Name	**sp_code**	**Common Name**	**Class**
Amelanchier alnifolia	Aa	Saskatoon	plant
Aruncus dioicus	Ad	Goat's Beard	plant
Acer glabrum	Ag	Douglas Maple	deciduous
Alnus sp	Al	Alder	deciduous
Anaphalis margaritacea	Am	Pearly Everlasting	plant
Aralia nudicaulis	An	Wild Sarsaparilla	plant
Actaea rubra	Ar	Baneberry	plant
Arctostaphylos uva-ursi	Auu	Kinnikinnick	plant
Cornus canadensis	Cc	Bunchberry	plant
Crataegus douglasii	Cd	Black Hawthorn	plant
Castilleja miniata	Cm	Red Paintbrush	plant
Corylus cornuta	Coco	Beaked Hazelnut	plant
Cornus stolonifera	Cs	Red-osier Dogwood	plant
Disporum hookeri	Dh	Hooker's Fairybells	plant
Epilobium angustifolium	Ea	Fireweed	plant
Equisetum sp	Eq	Horsetail	plant
Geocaulon lividum	Gl	Bastard Toad-flax	plant
Heracleum lanatum	Hl	Cow-parsnip	plant
Juniperus communis	Jc	Common Juniper	plant
Lysichiton americanum	La	Skunk Cabbage	plant
Linnaea borealis	Lb	Twinflower	plant
Ledum groenlandicum	Lg	Labrador Tea	plant
Lonicera involucrata	Li	Black Twinberry	plant
Lupinus sp	Lu	Lupine	plant
Lycopodium annotinum	Lya	Stiff Clubmoss	plant

Mitella nuda	Mn	Common Mitrewort	plant

Appendix B | Site Attributes

Site	200rd_13km	700rd_28km	alezza_lake	bend_0-5km	chief_lake_site_1	conifex_h47	conifex_k14	north_fraser_11km	north_fraser_41km	north_fraser_50km	north_olson_5km
SP_CD_1	Pl	Cs	Sx	Sx	Pl	Li	Bl	Sx	Ri	Sx	Bl
SP_COUNT_1	1267	399	791	251	969	106	104	1153	118	163	309
SP_PCT_1	32.97	31.72	42.01	49.51	22.6	22.75	28.97	41.52	21.11	44.17	32.39
SP_CD_2	At	Sx	Ep	Ep	Fd	Bl	W	Ep	Li	Ep	Pl
SP_COUNT_2	564	288	337	159	770	88	88	751	104	91	252
SP_PCT_2	14.68	22.89	17.9	31.36	17.96	18.88	24.51	27.04	18.6	24.66	26.42
SP_CD_3	Li	Ep	W	W	Al	Al	At	W	Mixed	W	Fd
SP_COUNT_3	389	221	198	49	555	75	66	225	46	43	121
SP_PCT_3	10.12	17.57	10.52	9.66	12.95	16.09	18.38	8.1	8.23	11.65	12.68
CON_COUNT	1708	373	976	257	2189	126	129	1386	9	174	742
CON_PCT	44.44	29.65	51.83	50.69	51.06	27.04	35.93	49.91	1.61	47.15	77.78
DECD_COUNT	1035	287	587	226	1092	146	161	1048	87	142	193
DECD_PCT	26.93	22.81	31.17	44.58	25.47	31.33	44.85	37.74	15.56	38.48	20.23
PLANT_COUNT	1100	598	320	24	1006	194	69	343	463	53	19
PLANT_PCT	28.62	47.54	16.99	4.73	23.47	41.63	19.22	12.35	82.83	14.36	1.99
Tree_size	0.849268	1.044152	1.021664	0.0872799	1.265636	0.901288	0.998359	0.912881	1.538062	1.279441	1.111404
Tree_size_variation	2.498069	2.47825	5.773248	5.26019	5.97062	3.355597	4.203311	1.617215	6.820703	5.500881	3.99158
tree_count	3843	1258	1883	507	4287	466	359	2777	559	369	954
AVG_ELEVATION	714	713	758	577	674	1118	1009	596	522	492	732
Var_ELEVATION	1.756	16.778	6.424	9.331	10.454	4.981	7.082	7.693	3.93	5.405	3.118

Appendix C | SLIC Site-Specific Hyper Parameters

Metric	Site	Score	*Compactness*	*Edges/ha*
1	200rd_13km	-0.16414	*0.003052*	*13300*
	700rd_28km	-0.11134	*0.003052*	*12700*
	alezza_lake	-0.23842	*0.390625*	*13600*
	bend_0-5km	-0.07174	*6.25*	*13600*
	chief_lake_site_1	-0.12559	*0.003052*	*13700*
	conifex_h47	-0.05464	*0.001526*	*10100*
	conifex_k14	-0.02676	*0.003052*	*13700*
	north_fraser_11km	-0.15121	*0.003052*	*12700*
	north_fraser_41km	-0.14689	*0.003052*	*12600*
	north_fraser_50km	-0.09012	*0.001526*	*13900*
	north_olson_5km	-0.06955	*0.001526*	*13400*
2	200rd_13km	-0.67233	*0.003052*	*1800*
	700rd_28km	-0.43038	*0.003052*	*1500*
	alezza_lake	-0.72822	*0.006104*	*3000*
	bend_0-5km	-0.48668	*0.024414*	*1000*
	chief_lake_site_1	-1.14613	*0.001526*	*500*
	conifex_h47	-0.25366	*0.001526*	*1600*
	conifex_k14	-0.14643	*0.001526*	*1400*
	north_fraser_11km	-0.47346	*0.003052*	*3800*
	north_fraser_41km	-0.4463	*0.001526*	*500*
	north_fraser_50km	-0.29921	*0.001526*	*500*
	north_olson_5km	-0.22939	*0.001526*	*500*
3	200rd_13km	-0.3148	*0.003052*	*5200*
	700rd_28km	-0.22408	*0.003052*	*3800*
	alezza_lake	-0.47856	*0.012207*	*5500*
	bend_0-5km	-0.28838	*0.195313*	*2000*
	chief_lake_site_1	-0.6889	*0.003052*	*1100*
	conifex_h47	-0.10666	*0.001526*	*3500*
	conifex_k14	-0.0451	*0.001526*	*2400*
	north_fraser_11km	-0.22789	*0.003052*	*8700*
	north_fraser_41km	-0.4463	*0.001526*	*500*
	north_fraser_50km	-0.29921	*0.001526*	*500*
	north_olson_5km	-0.22939	*0.001526*	*500*
4	200rd_13km	-0.3148	*0.003052*	*5300*
	700rd_28km	-0.1645	*0.003052*	*6500*
	alezza_lake	-0.47856	*0.012207*	*5500*
	bend_0-5km	-0.21667	*0.006104*	*3200*
	chief_lake_site_1	-0.6889	*0.003052*	*1100*
	conifex_h47	-0.09455	*0.001526*	*4800*
	conifex_k14	-0.03739	*0.572344*	*2400*
	north_fraser_11km	-0.22789	*0.003052*	*8200*
	north_fraser_41km	-0.4463	*0.001526*	*500*
	north_fraser_50km	-0.29921	*0.001526*	*500*
	north_olson_5km	-0.22939	*0.001526*	*500*

Appendix D | Quick Shift Site-Specific Hyperparameters

Metric	Site	Score	*Ratio*	*Kernel Size*	*Sigma*
1	200rd_13km	-0.11529	*0.9*	*1*	*1*
	700rd_28km	-0.0923	*0.6*	*1*	*5*
	alezza_lake	-0.10861	*0*	*1*	*5*
	bend_0-5km	-0.03424	*0.5*	*1*	*1*
	chief_lake_site_1	-0.07106	*0.9*	*1*	*3*
	conifex_h47	-0.03572	*0.4*	*1*	*1*
	conifex_k14	-0.05452	*0.8*	*1*	*3*
	north_fraser_11km	-0.07675	*0.1*	*1*	*1*
	north_fraser_41km	-0.11377	*0.3*	*1*	*5*
	north_fraser_50km	-0.07619	*0.6*	*1*	*3*
	north_olson_5km	-0.05211	*0.7*	*1*	*3*
2	200rd_13km	-0.82122	*0.9*	*5*	*3*
	700rd_28km	-0.55086	*0.9*	*5*	*11*
	alezza_lake	-0.95105	*0.6*	*5*	*9*
	bend_0-5km	-0.33437	*0.4*	*5*	*9*
	chief_lake_site_1	-0.50668	*0.1*	*5*	*11*
	conifex_h47	-0.4337	*0.7*	*5*	*1*
	conifex_k14	-0.37687	*0.8*	*5*	*5*
	north_fraser_11km	-0.82614	*0.1*	*5*	*3*
	north_fraser_41km	-0.33876	*0.9*	*5*	*1*
	north_fraser_50km	-0.22037	*0.9*	*5*	*1*
	north_olson_5km	-0.16153	*0.9*	*5*	*7*
3	200rd_13km	-0.43327	*0.6*	*3*	*1*
	700rd_28km	-0.3214	*0.6*	*3*	*9*
	alezza_lake	-0.52079	*0.5*	*3*	*9*
	bend_0-5km	-0.33437	*0.4*	*5*	*9*
	chief_lake_site_1	-0.50668	*0.1*	*5*	*11*
	conifex_h47	-0.19262	*0.6*	*3*	*3*
	conifex_k14	-0.37687	*0.8*	*5*	*5*
	north_fraser_11km	-0.40515	*0.1*	*3*	*3*
	north_fraser_41km	-0.33876	*0.9*	*5*	*1*
	north_fraser_50km	-0.21943	*0.8*	*5*	*1*
	north_olson_5km	-0.16153	*0.9*	*5*	*7*
4	200rd_13km	-0.43373	*0.5*	*3*	*1*
	700rd_28km	-0.32147	*0.4*	*3*	*9*
	alezza_lake	-0.52075	*0.5*	*3*	*7*
	bend_0-5km	-0.33426	*0.8*	*5*	*11*
	chief_lake_site_1	-0.50668	*0.1*	*5*	*11*
	conifex_h47	-0.19829	*0.3*	*3*	*3*
	conifex_k14	-0.20615	*0.8*	*3*	*3*
	north_fraser_11km	-0.40515	*0.1*	*3*	*3*
	north_fraser_41km	-0.33876	*0.9*	*5*	*1*
	north_fraser_50km	-0.21943	*0.8*	*5*	*1*
	north_olson_5km	-0.16153	*0.9*	*5*	*7*

Appendix E | Felzenszwalb's Efficient Graph Site-Specific Hyperparameters

Metric	Site	*Score*	*Scale*
1	200rd_13km	*-0.13814*	*0.0425*
	700rd_28km	*-0.10638*	*0.1025*
	alezza_lake	*-0.14963*	*0.0775*
	bend_0-5km	*-0.05281*	*0.05*
	chief_lake_site_1	*-0.08896*	*0.0325*
	conifex_h47	*0.00665*	*0.2375*
	conifex_k14	*-0.01532*	*0.1825*
	north_fraser_11km	*-0.09844*	*0.1075*
	north_fraser_41km	*-0.12722*	*0.055*
	north_fraser_50km	*-0.08017*	*0.06*
	north_olson_5km	*-0.05455*	*0.03*
2	200rd_13km	*-3.29638*	*0.28*
	700rd_28km	*-2.92586*	*0.4275*
	alezza_lake	*-3.4425*	*0.505*
	bend_0-5km	*-2.77591*	*0.585*
	chief_lake_site_1	*-6.61233*	*0.995*
	conifex_h47	*-2.34211*	*0.575*
	conifex_k14	*-2.1713*	*0.505*
	north_fraser_11km	*-2.8154*	*0.615*
	north_fraser_41km	*-8.6102*	*0.9975*
	north_fraser_50km	*-8.78572*	*0.9925*
	north_olson_5km	*-8.96649*	*0.985*
3	200rd_13km	*-1.68415*	*0.2025*
	700rd_28km	*-1.4949*	*0.3125*
	alezza_lake	*-1.6962*	*0.3125*
	bend_0-5km	*-1.42957*	*0.3*
	chief_lake_site_1	*-3.35091*	*0.345*
	conifex_h47	*-1.13341*	*0.4475*
	conifex_k14	*-1.11794*	*0.3625*
	north_fraser_11km	*-1.44071*	*0.31*
	north_fraser_41km	*-4.10873*	*0.9975*
	north_fraser_50km	*-4.0638*	*0.9925*
	north_olson_5km	*-4.15368*	*0.985*
4	200rd_13km	*-2.16572*	*0.225*
	700rd_28km	*-2.13797*	*0.31*
	alezza_lake	*-2.15994*	*0.3575*
	bend_0-5km	*-2.05404*	*0.2925*
	chief_lake_site_1	*-3.68367*	*0.43*
	conifex_h47	*-1.84296*	*0.37*
	conifex_k14	*-1.85426*	*0.275*
	north_fraser_11km	*-2.06405*	*0.31*
	north_fraser_41km	*-4.54648*	*0.9975*
	north_fraser_50km	*-4.53363*	*0.9925*
	north_olson_5km	*-4.62563*	*0.985*

Appendix F | Mean Shift Site Scores

Metric	Site	Score
1	200rd_13km	-0.16414
	700rd_28km	-0.11134
	alezza_lake	-0.23842
	bend_0-5km	-0.07174
	chief_lake_site_1	-0.12559
	conifex_h47	-0.05464
	conifex_k14	-0.02676
	north_fraser_11km	-0.15121
	north_fraser_41km	-0.14689
	north_fraser_50km	-0.09012
	north_olson_5km	-0.06955
2	200rd_13km	-0.67233
	700rd_28km	-0.43038
	alezza_lake	-0.72822
	bend_0-5km	-0.48668
	chief_lake_site_1	-1.14613
	conifex_h47	-0.25366
	conifex_k14	-0.14643
	north_fraser_11km	-0.47346
	north_fraser_41km	-0.4463
	north_fraser_50km	-0.29921
	north_olson_5km	-0.22939
3	200rd_13km	-0.3148
	700rd_28km	-0.22408
	alezza_lake	-0.47856
	bend_0-5km	-0.28838
	chief_lake_site_1	-0.6889
	conifex_h47	-0.10666
	conifex_k14	-0.0451
	north_fraser_11km	-0.22789
	north_fraser_41km	-0.4463
	north_fraser_50km	-0.29921
	north_olson_5km	-0.22939
4	200rd_13km	-0.3148
	700rd_28km	-0.1645
	alezza_lake	-0.47856
	bend_0-5km	-0.21667
	chief_lake_site_1	-0.6889
	conifex_h47	-0.09455
	conifex_k14	-0.03739
	north_fraser_11km	-0.22789
	north_fraser_41km	-0.4463
	north_fraser_50km	-0.29921
	north_olson_5km	-0.22939

Appendix G | Aggregate Hyperparameters for SLIC

SLIC	Compactness	Segments/ha
Metric 1	0.003051758	13900
Metric 2	0.003051758	500
Metric 3	0.003051758	2300
Metric 4	0.003051758	2300

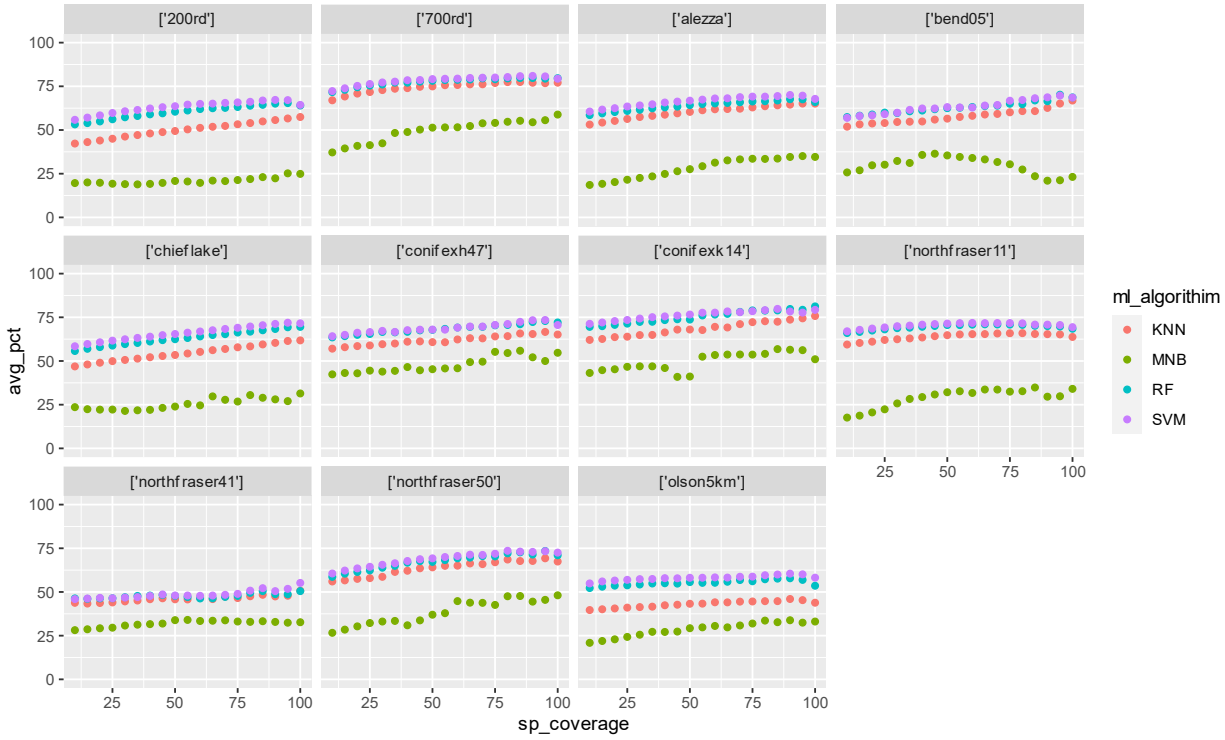
Appendix H | Aggregate Hyperparameters for QuickShift

QuickShift	Ratio	Kernel	Sigma
Metric 1	0.5	1	1
Metric 2	0.6	5	9

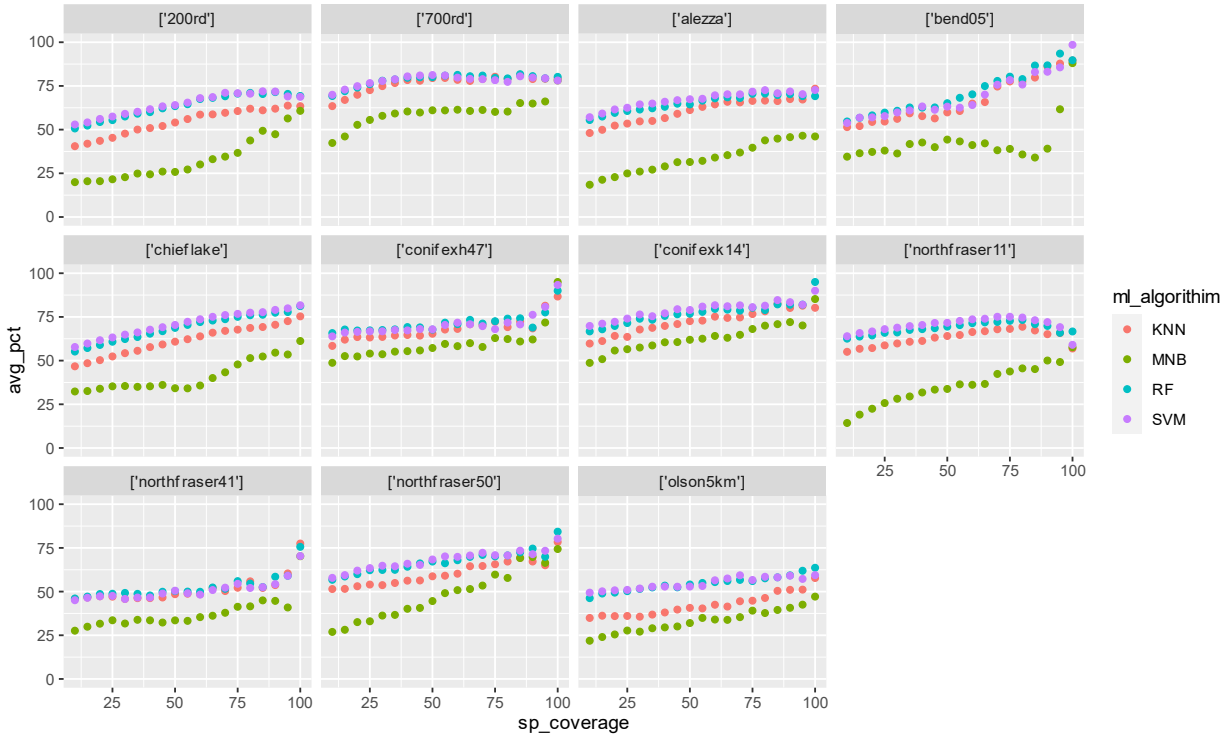
Metric 3	0.3	5	9
Metric 4	0.1	5	5

Appendix I | Aggregate Hyperparameters for F-Graph

F-Graph	Scale
Metric 1	0.045
Metric 2	0.635
Metric 3	0.5
Metric 4	0.4825



Appendix J | Classification Accuracy of Quick Shift Segments on individual sites.



Appendix K | Classification Accuracy of SLIC Segments on individual sites.