Universiteit Leiden

Computer Science

Combining spectral and texture features for aquatic plant species classification

A comparison of remote sensing data

Name:	Yuqi Zhang
Studentnr:	s1959670
Date:	01/06/2019
University supervisors:	Dr. Wojtek Kowalczyk
	Dr. Hao Wang
Company supervisors:	Dr. Ellis Penning
	Sandra Gaytan Aguilar

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Combining spectral and texture features into aquatic plant species classification: A comparison of remote sensing data

by

Yuqi Zhang

to obtain the degree of Master of Computer Science and Advanced Data Analytics

at the Leiden University.

Student number:	s1959670	
Project duration:	Dec 1, 2018 – July 1, 2019	
Thesis committee:	Dr. W. Kowalczyk,	Liacs, supervisor
	Dr. H. Wang,	Liacs
	Dr. E. Penning,	Deltares
	S. G. Aguilar,	Deltares

Deltares

Enabling Delta Life

Abstract

Aquatic plants serves as a great environmental indicator of overall water body health. Currently, Rijkswaterstaat conducts vegetation monitoring every three years. The method, however, is manual, expensive and time consuming. Remote sensing data has the potential to improve the cost effectiveness of the monitoring significantly by image classification. After a comparison of different sources of remote sensing data, this study uses free satellite Sentinel-2 data and selects Randmeren as our study area. A validation map is obtained by projecting field samples to the image after region-based segmentation. The image preprocessing and segmentation are done using different tools in QGIS. Spectral features, including reflectance of each band and vegetation indices, and texture features, including Local Binary Patetrns and Gabor Filters, are extracted. Principal component analysis is then applied to reduce the dimension of the combined feature vectors. A supervised algorithm Random Forest classifier and an unsupervised algorithm K-means clustering are employed for multiple classification tasks respectively. The six classes that are classified are: water, low density plants, floating plants, characeae, submerged plants and pondweed. The unique features of individual species including Doorgroeid fonteinkruid and Kransblad are also investigated in this study. The experimental results showed that the Random Forest classification achieved higher accuracy compared to K-means clustering, with an average accuracy of 87%. Spectral features contributed most in all classification tasks, with an average importance of 44.3% for vegetation indices and 37.3% for reflectance, while texture features contributed least. This research proved that Sentinel-2 data can be used as a tool for efficient aquatic vegetation monitoring in shallow water bodies. For complex regions, however, airborne images are required for more details.

assumptions

***Keywords*—** Remote sensing data, Aquatic plants, Segmentation, Classification, Comparison.

Preface

This graduation project started at the company Deltares and was undertaken from 01 September 2018 to 28 May 2019, as a collaboration with Rijkswaterstaat. The first phase of my research was to specify the objective and main question of this project. After an extensive literature study, multiple meetings with my supervisors at Deltares and Leiden University, I formulated the main question for this research. This research goal was to utilize remote sensing data for efficient aquatic vegetation monitoring.

I want to express my sincere gratitude to my supervisors from Deltares. Firstly, I would like to thank my supervisor Dr. E. Penning, for offering this interesting project to me, helping me combining ecological knowledge into data science, intensive guidance to my thesis writing and speech skills. Secondly, I would like to thank my supervisor S. G. Aguilar for technical support every time I came across problems. Next, I would like to thank Jonathan Nuttall for providing related courses to our interns. Lastly, I also want to thank my supervisor in Leiden University, Dr. W. Kowalczyk for giving useful suggestions and attending meetings at Deltares.

The Hague, May 2019

Contents

List of Figures

List of Tables

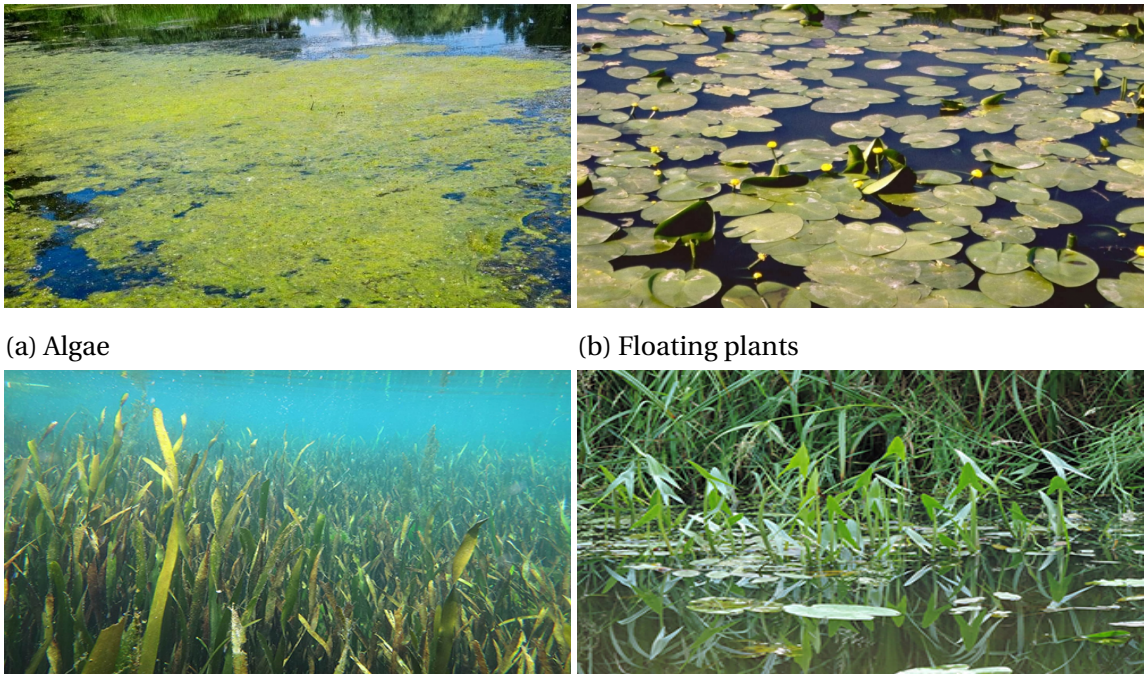
1 Introduction

1.1. Aquatic plants

Aquatic plants are plants adapted to life in water. Their distribution and abundance are controlled by many factor, such as water temperature, water quality, and nutrient enrichment (Dar et al., 2014). Aquatic plants can be grouped into four common class types based on their morphological appearance, as shown in Figure 1.1: algae, floating plants, submerged plants and emerged plants, according to the positioning of their roots and leaves (Lembi, 2009). Algae have no roots, stems or leaves. Their size can range from the microscopic (microalgae), to large seaweeds (macroalgae). Algae usually explode in water with abundant nutrients. A microscopic algae population that colors the water green or yellow green is called a "bloom". Floating plants can be divided into floating-leaved plants and free-floating plants (Freedman and Lacoul, 2006). Free-floating plants are not rooted in the soil. They suspend on the water surface, freely to be moved by wind and water currents. Common species include water-lettuce and duckweed. Floating-leaved plants are rooted in the soil with their leaves floating on the water surface. Common species include water lily and floating pondweeds. Submerged plants are rooted in the soil and grow up through water column. Common species include pondweed and hydrillas. Submerged plants stabilize bottom sediment and reduce turbidity. Emerged plants are rooted in the soil and grow above the water surface. These plants need constant exposure to sunlight. Examples include cattails. Emerged plants attenuate the wave energy and protect shoreline from erosion (Tschirky et al., 2001).

1.2. The role of aquatic plants in aquatic ecosystems

An aquatic ecosystem is a habitat where organisms live and interact within a water body. Familiar examples are lakes, rivers, and wetlands. Aquatic ecosystems contain a wide variety of organisms



(a) Algae

(b) Floating plants

(c) Submerged plants

(d) Emerged plants

Figure 1.1: Four common types of aquatic plants

including microorganisms such as bacteria and fungi; plants such as macrophytes and riparian vegetation; invertebrates such as worms and insects; vertebrates such as fish and reptiles.[1] Aquatic ecosystems are vital sources of fish for human consumption, are repositories of bio-diversity and are crucial to the functioning of the hydrological cycle (S.K.Jain and V.P.Singh, 2003). Human population growth accompanied by intensified industrial, commercial and residential development, has added pressure to these ecosystems (Council, 1992). Phosphorus fertilizers, toxic chemicals and plastic pollution from human activities are all threats to aquatic species and their habitat. Overexploitation makes fish species increasingly scarce. At the same time, invasive species can cause the extinction of native species and disrupt ecosystem functioning (Chatterjee, 2017). Aquatic plants are a natural part of every ecosystem and serve many purposes in a lake. They produce leaves and stems as food for aquatic organisms, provide oxygen through photosynthesis, aid in nutrient cycling, and provide habitat for invertebrates and fish (Carpenter and Lodge, 1986). All these functions contribute to the diversity of aquatic ecosystems. Therefore, aquatic plants serves as a great environmental indicator of overall water body health.

1.3. The importance of aquatic plants monitoring

Monitoring of aquatic plants is becoming increasingly important to document distribution, and abundance of aquatic plants over time (Bartodziej and Ludlow, 1997; Sytsma, 2008), via regular data

[1] `http://www.ramp-alberta.org/river/ecology/life.aspx`

collection of species and respective habitats of plant communities. Changes include loss of native plants, spread of nonnative plants and habitat alteration. Some species are known as "key species" as they provide food or habitat for animal species. They are vital for the survival of other species through the actions they perform for the maintenance of these species. Their loss or disappear may cause a major chaos in the functioning of the ecosystems and thus are essential to maintain one or more communities. In conservation biology, the term bio-indicator is used for species whose presence or the fluctuation of whose population reflect changes in the environment or in communities of other species (Diop, 2010). Thus it is necessary to quantify plant populations and understand distributions of plant communities. Monitoring can help detect negative changes in early stage and plot control strategies in time.

1.4. Remote sensing data for monitoring aquatic plants

Traditional field-based monitoring of aquatic plants face several challenges, including inaccessibility of areas for boats to collect field samples, vegetation dynamic changes in species and distributions, and budget constraints on field survey (Jakubauskas et al., 2012). These constraints, on the other hand, have increasingly boosted applications of remote sensing data from aircraft and satellite for aquatic vegetation analysis and monitoring. Chen et al. (2018) applied GF-1 satellite data (of four bands and 16 m resolution) to extract submerged aquatic vegetation and obtained an overall accuracy of 91% on four classes classification. This study utilized a decision tree model by first identifying land using red band, identifying emerged vegetation using reflectance difference in green and red bands and NIR peak, then identifying Huangtai algae using NDVI and finally extracting submerged aquatic vegetation by using a concave–convex decision function. However, this study was conducted in a small shallow lake of depth ranging from 0.5 to 2.5 m. The spectral signal of aquatic plants is affected by water turbidity and water depth. This study also discussed that NIR peak and concave characteristic disappeared at 0.5 m with high turbidity, which shown difficulties in extracting aquatic plants underwater. Different with the approach above, Luo et al. (2017) involved life history information of seven submerged plant species from February to October when using satellite images. These seven species reached maximum biomass and died in different seasons. Only three to four species of different biomass were dominant in Taihu lake at the same time. Based on their life information, species were extracted from the satellite image from the month that they reached maximum biomass, and these areas were ignored for further processing. After extraction of all species had finished, separate maps were combined into one map to demonstrated the distribution of submerged plants, and this map reached 68% overall accuracy. Although they gave a novel direction to involve plant life history information, multi-temporal satellite images may have various qualities; besides it was a knowledge-based approach which would be sensitive to weather changes and invasive species, and therefore it was hard to apply to other lakes. In general, airborne images are of higher spatial resolution than satellite images, this gives the ability to extract texture features and to segment images into regions of different aquatic plant species. Husson et al. (2016)

applied Unmanned Aircraft Systems (UAS) images with 5 cm resolution for non-submerged aquatic vegetation classification. Their results shown that combining spectral and texture features would increase classification accuracy as compared to using spectral and texture features alone. Pande-Chhetri et al. (2014) shown that airborne hyperspectral imagery with narrow band-width (7.4 nm) could be an efficient tool for classifying macrophyte communities.

1.5. Project description and research objectives

Rijkswaterstaat (RWS) is the executive agency of the Ministry of Infrastructure and Water Management in the Netherlands. Its main task is working on the management and maintenance of the system of waterways.[2] In 2018, a pilot location has been selected in consultation with RWS. The choice was made for the Eemmeer and the Veluwemeer based on field knowledge and the fact that the regular monitoring was also carried out in the Randmeren in 2018. Figure 1.2 shows the spatial relations of Eemmeer and Veluwemeer in Randmeren. The Eemmeer and Veluwemeer together cover a very diverse vegetation pattern, with the Eemmeer being characterized by a fountain herb composition and the Veluwemeer is dominated by wreath algae.

Figure 1.2: Randmeren in Satellite image

[2] https://nl.wikipedia.org/wiki/Rijkswaterstaat

The development in multispectral and hyperspectral sensors have boosted their applications in monitoring vegetation in terrestrial and small-scale water systems (Allan, 1990). These sensors are used in combination with different platforms: satellite and aircraft. The satellite images are relatively easy to obtain, but may not have the desired spatial resolution. The aircraft images may be relatively more labor intensive and expensive to obtain, but of a higher spatial resolution. An assessment must therefore be made of which method would give the most optimal results, and which types of images most closely matches the aim of actually making the monitoring more efficient.

In this project, two types of images have been taken from an airplane, including multispectral and hyperspectral images of high resolution (0.2×0.2 m), and most of test sections have been flown successfully. However, the switch from small regions to entire water body has not been entirely successful. Scarce bank reference in the middle of large water bodies led to difficulty in obtaining a flat and fully covered scene. On the other hand, satellite images such as Sentinel-2, though with coarse resolution (10×10 m), are free and flat-covering the whole region of interest. Therefore, Sentinel-2 data can be the first step to explore the practical applicability of remote sensing technologies in aquatic vegetation mapping.

With the current situation stated above the research objectives are:

1. To classify aquatic plant species in Randmeren using Sentinel-2 satellite images, estimate its interpretation and classification accuracy of aquatic plant species by using field observations as validation data and then discuss whether Sentinel-2 data can be an alternative of mapping and monitoring aquatic vegetation.

2. To compare the performance and cost-efficiency of Sentinel-2 images, airborne multispectral images and airborne hyperspectral images. Since airborne images only cover small parts of Randmeren, they are not used for classification all plant species, but for the comparison of feature quality, segmentation quality with satellite images at locations that they are overlapped. Then discuss the potentials and possibilities that whether and on which aspects two types of airborne images can improve the classification accuracy.

1.6. Research questions

1. What are the unique properties to discriminate aquatic plant species?

2. Can we extract those unique properties from Sentinel-2 images?

3. What classification result can Sentinel-2 images yield?

4. Are field data necessary in remote sensing techniques?

5. To what extent do spectral and spatial resolution influence the classification result respectively?

6. Can Sentinel-2 data be an alternative to airborne images?

7. Can remote sensing techniques be used for the large water bodies as a tool to achieve faster and more efficient monitoring of the vegetation?

1.7. Outline

Chapter 2 explains background of remote sensing and processing techniques that are related to this project. Chapter 3 gives detailed description of materials and methods we used. Chapter 4 describes the results from those methods in detail. The results are discussed in chapter 5 and a final conclusion with answers to the primary research questions and recommendations for future works are given in chapter 6.

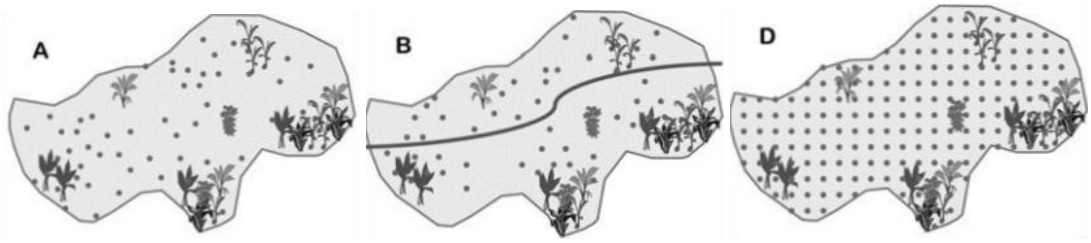
2

Background

2.1. Field surveys and sampling methods

2.1.1. Sampling design

An appropriate sampling design makes sure that collected data are suitable for quantifiable analyses (Madsen and Wersal, 2012). The most common sampling designs for ecological data are random sampling, stratified sampling, and systematic sampling as shown in figure 2.1. The completely random design removes biases caused by the subjective selection of sampling locations. However, the points selected from this method may be placed in inaccessible areas and thus the completely random design is not suitable for large water bodies. A stratified random design places sample locations in homogeneous sections which are typically divided by a stream channel. In each section, points are randomly distributed. The systematic sampling design places sample locations based on a grid and covers the entire water body.

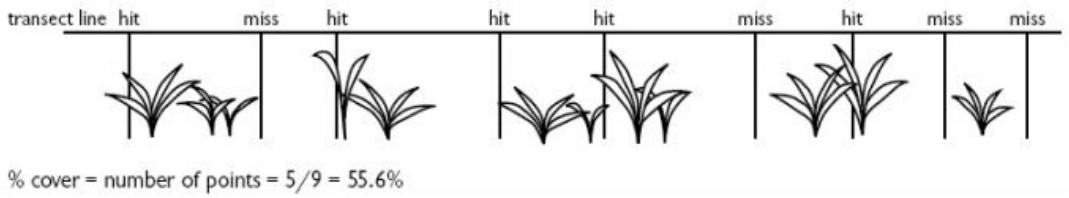


(a) completely random (b) stratified random (c) systematic designs

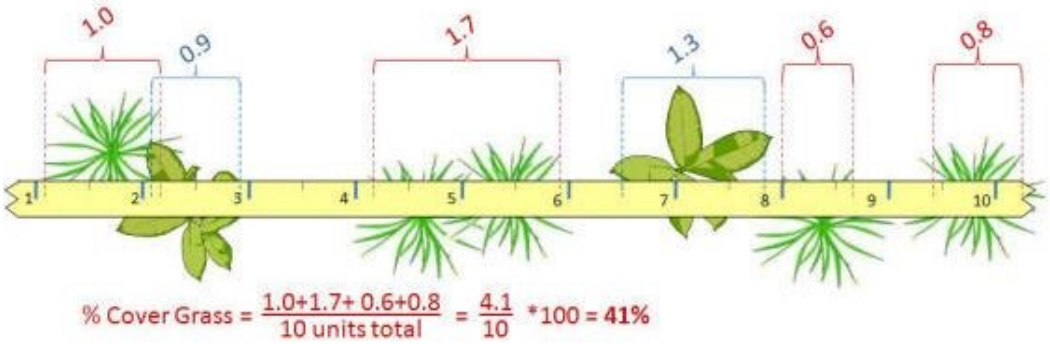
Figure 2.1: Common sampling designs (Madsen and Wersal, 2012)

2.1.2. Sampling methods

Field surveys are conducted based on a given sampling design. Methods can be quantifiable and non-quantifiable. Visual observation above water surface is the simplest way to estimate aquatic plant cover. This method, however, is non-quantifiable, subjective and highly variable among observers. Two common quantitative methods are point intercept and line intercept. Examples are shown in figure 2.2. Point intercept is designed to avoid subjectivity, by making measurements at pre-selected and regularly spaced grid points, with the help of a Global Positioning System (GPS) and Geographic Information Systems (GIS) technologies. On each point a plant rake is used to sample submerged vegetation; floating and emerged vegetation are recorded as well. The grid interval between points depends on the size of study sites (Madsen, 1999). Line intercept method is similar to the point intercept method, however, with transects to collect data. Line intercept method requires less technology than the point intercept method. Generally, several transects are deployed in the study area to determine the percent occurrence of the species in the plant community. The point intercept survey works well to estimate the cover and distribution of submerged plants in whole lake monitoring (Hauxwell et al., 2019). However, the line intercept can be more effective in monitoring and assessing emerged and floating plants in small areas as their distributions are typically more concentrated than submerged plants. RWS uses the grid-based point intercept methods for regular monitoring and details will be explained in chapter 3.



(a) Point intercept



(b) Line intercept

Figure 2.2: Two common quantitative methods[1]

[1] https://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%208(Cover)/8_3_Points.htm

2.2. Remote sensing principles

2.2.1. Optical remote sensing

Optical remote sensing takes images of the earth's surface by detecting the solar radiation reflected from materials on the ground, as illustrated in figure 2.3. The solar radiation is measured by use of visible, near infrared (NIR) and short-wave infrared (SWIR) sensors, as shown in figure 2.4. Materials with different attributes such as surface texture, color, and structure may have different reflectence patterns at different wavelengths. Thus, unique spectral reflectance signatures of different types of materials can be extracted in the remotely sensed images.

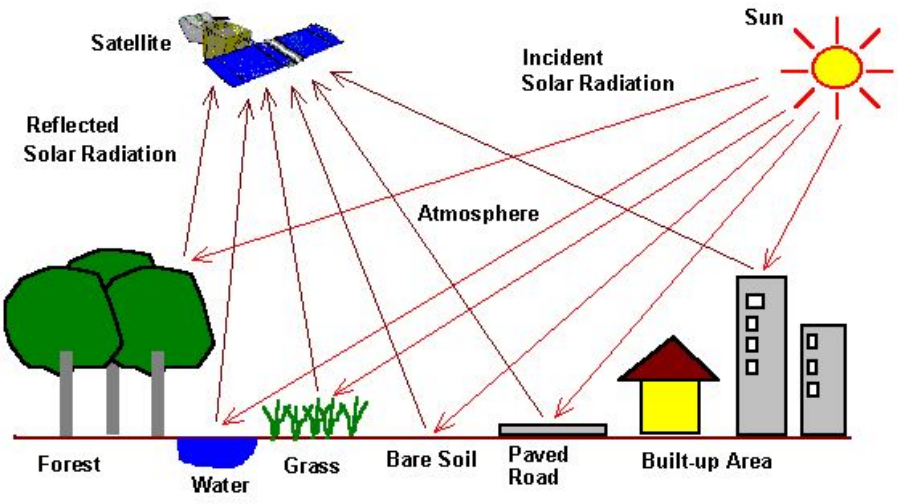


Figure 2.3: Illustration of remote sensing principles

2.2.2. Remote sensing imagery

Remote sensing data are available from a range of sources and data acquisition techniques, which can be collected from the ground, the air using aircraft or from space using satellite. Both satellite imagery and aerial photography provide rich sources of information of the earth surface.

Satellite Imagery

Satellites orbit around the earth continuously and are capable to map a large area in relatively small amount of time. Satellite imagery can take images of anywhere on the earth without the need to cross border or get access to restricted areas. Satellite data are often of a lower spatial resolution than aerial photographs, as satellites are generally hundred kilometers above. Military satellite more likely have a higher resolution but this imagery has not become publicly available. Most satellite imagery is between 5m and 1000m in resolution (Liang et al., 2012). Due to the position in orbit and

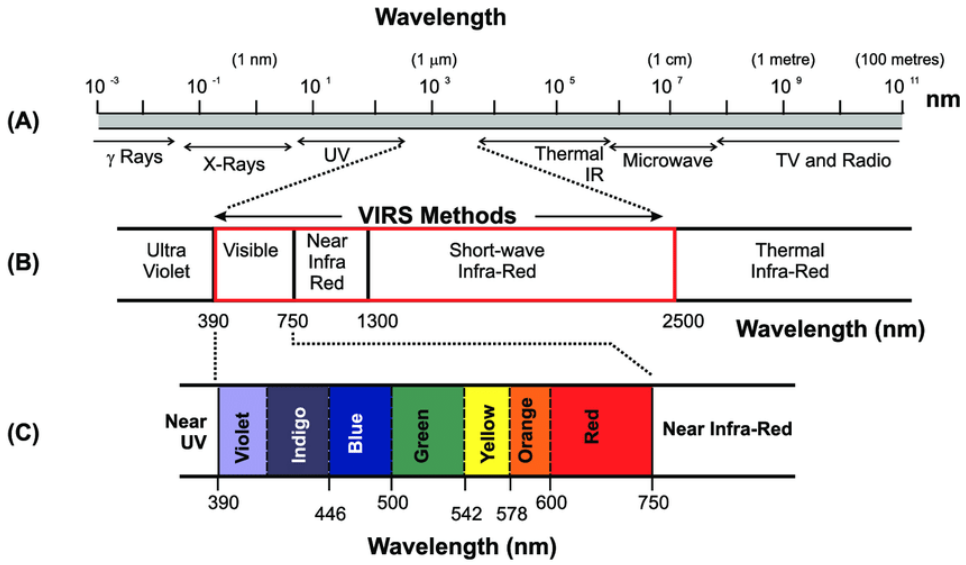


Figure 2.4: The electromagnetic spectrum

large scale of area covered, satellites have more weather conditions to cope with than aerial photography. Another problem is that when improved technology is released it is hard to change sensors in satellites in orbit. The cost of satellite data is straightforward and there are lots of satellite data freely available. Satellite imagery is cost-effective is often used by governments, large corporations and educational institutions for large-scale scientific applications.

Aerial Photography

Aerial photography is a time consuming process as an airplanes fly back and forth to take overlapping images. Traveling time and procedure of getting access procedure in some cases may also cause a delay for airborne data acquisition teams. Aerial photography has its advantage of being closer to the subject, as the flying height can be adjusted according to the required level of detail. The pixel size can be 1-3m or less. High-resolution imagery, however, also requires a suitable system for storing a complete collection of images, and the ability to handle the large data files and process them. Aerial photography can wait for an improved weather and post-processing stage can rectify effects from thin or high level cloud. Latest cameras or sensors can be utilized easily in aerial photography. The cost of aerial photography depends on many variables such as location, resolution and accuracy (Buchroithner, 2000). Thus, aerial photography is suitable for small-scale commercial applications due to its high cost per unit area of ground coverage (Buchroithner, 2000).

2.2.3. Data characteristics

Spatial, spectral and temporal resolutions are main qualities of remote sensing data.

Spatial resolution represents the square area on the ground that each data point covers. It describes how much detail in an image is visible to human eyes. Smaller pixels in a raster image show more

details on the ground, while larger pixels appear more coarse. High resolution is suitable for texture analysis. For instance, Tsai and Chou (2006) took advantages of high resolution (2m and 0.61m) satellite images to extract texture features for the detection of invasive plant species.

Spectral resolution refers to the spectral width of a dataset that contains a certain number of spectral bands. For instance, the sensor of panchromatic imagery is a single channel detector and the resulting image contains only one band. Multispectral images that usually contains a few spectral bands are created using a multichannel detector. Hyperspectral imagery acquires images in hundreds of continuous spectral bands. These narrow numerous can show subtle variations in reflected energy across the entire electromagnetic spectrum. Therefore, hyperspectral images shows more spectral characteristics of vegetation than multispectral images (Adam et al., 2010).

Temporal resolution refers to the amount of time needed to revisit and acquire data for the exact same location. In the case of the Sentinel-2 satellites, the temporal resolution is five days, while aerial photography is generally one-time operation.

2.2.4. Data preprocessing

Preprocessing is a crucial step and normally conducted prior to the main data analysis and varies depending on the type of sensor used. Below are common preprocessing steps:

Geometric correction corrects platform distortions due to the movement and noise and improve the positional accuracy of images. Real world coordinates and relating information are embedded into raw images in the georeferencing process.

Radiometric calibration takes into accounts directional measurements such as the position of the sensor and the sun, irradiance measurements and camera model (Kobayashi and Sanga-Ngoie, 2009). These considerations are intended to interpret the digital values as physically meaningful, quantitative value like radiance or reflectance.

Atmospheric correction eliminates the effects of atmospheric scattering on energy measured by sensors and determines the true surface reflectance values (Hadjimitsis et al., 2010).

2.3. Spectral properties of vegetation

The solar radiance, emitted from the Sun hits a target on the ground and then is transmitted, absorbed or reflected. The characteristics of the reflectance behaviors vary among different groups of photosynthetic organisms in the visible and near infrared spectral region (Kiang et al., 2007). Satterwhite and Henley (1990) studied the spectral curves of vegetation, soils, rocks, and human - made materials and demonstrated their unique spectral properties. Generally, the reflectance of vegetation is relatively low in the visible region. This feature is caused by plant pigments such as chlorophyll. Chlorophyll strongly absorbs energy in the blue and red wavelengths and reflects more

green wavelengths. That is the reason why human eyes perceive healthy vegetation as green. In the near infrared region, the reflectance is much higher due to the cellular structure of the leaves, specially mesophyll. Thus, vegetation can be identified by the high NIR reflectance and low visible reflectance. In the shortwave infrared wavelengths, the reflectance refers to the water content of the vegetation and its structure. Water absorbs light at wavelengths around 1.45, 1.95 and 2.50 µm. The characteristic of reflectance patterns is illustrated in figure 2.5. Clear water can be easily recognised due to the high absorption of wavelengths longer than 0.8 µm. Therefore, aquatic plants are distinguishable from water body and plant species can be identified if the spectral resolution of remote sensing data is sufficient to distinguish its spectrum from those of other species.

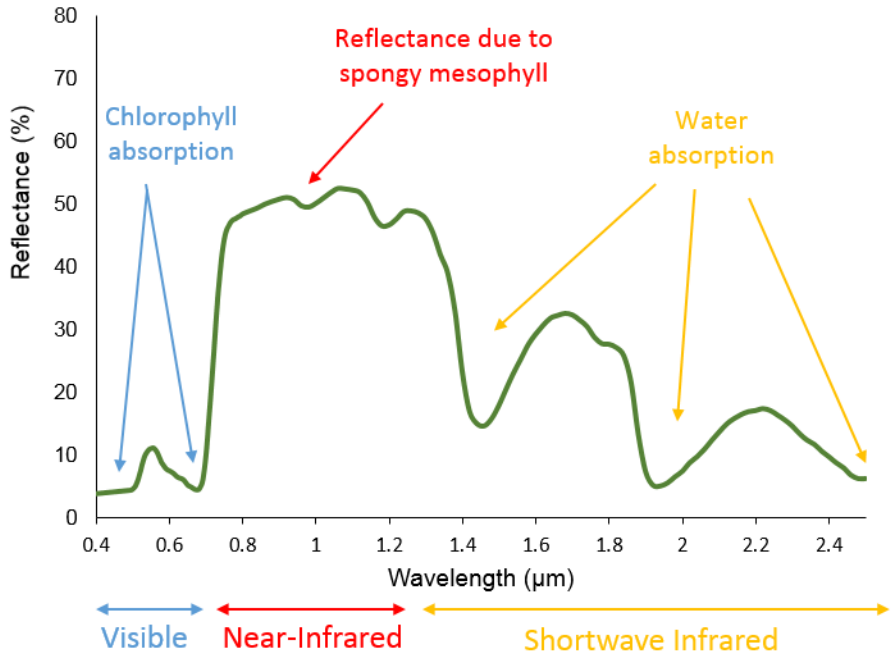


Figure 2.5: Vegetation spectral reflectance curves[2]

The use of appropriate features to differentiate plant species is fundamental for classification problems. From the literature review, the use of vegetation indices is dominant method for interpreting aquatic vegetation properties from remote sensing data. Vegetation indices combine reflectance in two or more spectral bands to model plants. Vegetation indices accentuates the spectral properties of green plants and serve as a good spectral feature for differentiating plant species. For example, Tian et al. (2010) differentiated aquatic plant communities using Normalized Difference Vegetation index (NDVI), near-infrared (NIR)-Green Angle index (NGAI) and normalized water absorption depth (DH) and achieved an overall accuracy of 79%. Ma and Zhou (2018) constructed indices such as Slope index and Angle index and shown better performance than conventional vegetation indices.

[2] `http://gsp.humboldt.edu/OLM/Courses/GSP_216_Online/lesson2-1/vegetation.html`

2.4. Influence of water depth and turbidity

The distinctive spectral characteristics of plant species has been used to develop vegetation indices. With the advance of remote sensing imagery, plant species are distinguishable by extracting vegetation indices from high-quality remote sensing images. Aquatic plants and their spectral characteristics, however, may not as easily distinguishable as terrestrial vegetation because the overlying water interferes with the vegetation signals. The energy from the sun is partially absorbed by water and the reflectance from aquatic plants underwater is scattered from water column. Lyzenga (1978) proposed a model called Depth Invariant Index (DII) to reduce the effect of light attenuation in the water column. This technique has been tested on various water environment using various imageries with satisfactory accuracies. However, Manessa et al. (2016) pointed out that the DII method is not capble to remove the water column effect in complex shallow water environments. A slight water quality variation (0.2-0.25 mg/m3) shows an significant effect on the performance of the DII model. Sagawa et al. (2010) proposed a new index for radiometric correction called Bottom Reflectance Index (BRI). The BRI overcomes the problem of DII when transparency decreases. The BRI utilizes attenuation coefficient of each band in different depths. These coefficients are measured by Profiler Reflectance Radiometer. In his experiments, Sagawa proved that BRI improved the accuracy of DII from 65% to 83% in the same case study. Turbidity, an water quality indicator, is caused by suspended organic and non organic sediments, and chlorophyll, etc. Water turbidity physically block the penetration of irradiance through the water column. Suspended sediments increase the radiance emergent from water surface in the visible and NIR spectrum (C Ritchie et al., 2003). Chlorophyll-a is an indicator of algae bloom. Algae floats on above underwater plants, which makes monitoring aquatic plants more difficult. A number of publications have studied water quality monitoring using remote sensing images (Arabi et al., 2016; Perivolioti et al., 2016). However, the literature on removing the spectral impacts of water turbidity is still scarce.

2.5. Texture features

When classifying individual plant communities, spectral features are usually used without considering mixed plant communities. Some studies have presented the efficiency of texture for interpreting the internal construction of different regions and improving vegetation classification accuracy (Wang et al., 2018). Texture features give us community structure information about specific region, not only colors, but the spatial arrangement of the colors or intensities. The Grey Level Co-occurrence Matrix (GLCM) is a commonly used texture operator and has been successfully applied to vegetation classification (Maillard, 2005). GLCM computes spatial relationship of pixels by extracting statistical measures (contrast, correlation, entropy, homogeneity) (Haralick et al., 1973). Local Binary Pattern (LBP) is a powerful visual descriptor and is widely used in computer vision applications, such as pedestrian detection (Yang et al., 2012) and crowd density estimation (K Pai et al., 2017). LBP analyses the grey value changes between a pixel and its neighbours, rather than

following one pixel in a particular direction. Studies applying LBP to the field of vegetation classification also shown great performance and got 85% overall accuracy (Wang et al., 2018). Gabor filters, claimed to be similar to human visual systems, are orientation-sensitive filters. Gabor Filters can effectively characterize repetitive structure in the Fourier domain. Chaki (2012) classified plant species by discriminating leaf shapes using Gabor filters. As the Gabor filters usually cover all possible orientations and produce large feature size, this technique also requires dimensions reduction.

2.6. Managing high dimensionality of feature space

Combining multiple texture features may cause high-dimensional mathematical problems. Dealing with these problems requires high computational time and the large amount of memory. More feature do not necessarily improve classification accuracy. In some cases, as the number of feature grows, the performance of our classifier would actually be degraded, if there are only a fixed number of training samples . This phenomenon is illustrated in figure 2.6.

The high dimensionality problem has pushed usage of dimensionality reduction techniques (Jindal and Kumar, 2017). Dimensionality reduction techniques aim at reducing or transforming the original feature space into another space of lower dimensionality. Principal Component Analysis (PCA) is a well-known dimensionality reduction technique. PCA uses an orthogonal transformation to transform original features into lower dimensional feature space by linearly combining original features. These new set of features are known as principle components and are uncorrelated. The first principal component has the largest possible variance. PCA is sensitive to the scale of features. Minimum Noise Fraction (MNF) transformation was first introduced by Green et al. (1988) as an extension of PCA. MNF is invariant with respect to scale changes in features. Studies have shown that MNF performs better than PCA in hyperspectral imagery denoising, as MNF maximizes the ratio of the signal to noise while PCA not necessarily (Ibarrola et al., 2017; Luo et al., 2016).

2.7. Image segmentation

Differentiating plant species in aquatic environments by traditional pixel-based image classification may be challenging. First, heterogeneity in local pixel values, spectral similarities among plant species and variable water transparency and often lead to noises in discrimination of classes (Wang et al., 2004). Second, 10-60 m resolution Sentinel-2 data is insufficient to detect fine-scale aquatic plant features and community boundaries; and mixed spectral reflectance pattern of different communities at pixel scales may reduce classification accuracy (C. Frohn et al., 2009; Dronova, 2015). Finally, pixel-based approaches may be limited by the lack of ecologically meaningful information on spatial context and class neighbourhood relationships in the pixel scale (Autrey and Reif, 2011).

[3] http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

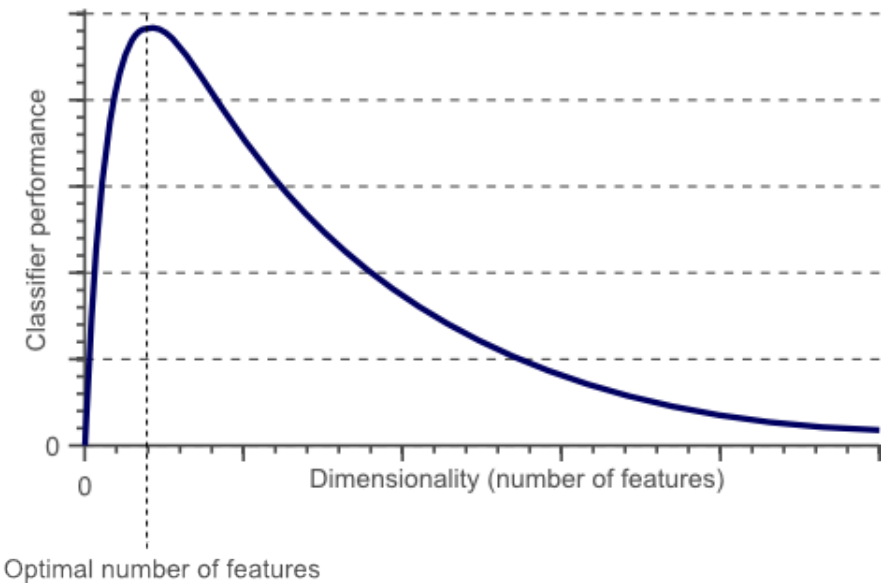


Figure 2.6: As the dimensionality increases, the classifier's performance increases until the optimal number of features is reached. Further increasing the dimensionality results in a decrease in classifier performance.[3]

In recent years, image segmentation has become popular in image classification, as a number of studies have proven to have better results than that from the pixel-based image analysis (Blaschke, 2010). Segmentation is a process that partitions an image into non-overlapped segments in image space by considering reflectance values, texture features and locations. Individual plant communities are expected to be segmented as individual segments by incorporating both texture variables and spectral values. Local variations can be smoothed by which noise can be relatively reduced (Kim et al., 2011).

There are three strategies for segmentation: Point-based segmentation; Edge-based segmentation and Region-based segmentation. In point-based segmentation, pixels are allocated to segments by applying threshold operations, which combine pixels that are similar in the range of values in which a pixel lies. However, this strategy is not suitable for remote sensing data, due to the varying reflectance within the image. Edge-based segmentation uses edge-detection algorithms. Pixels that are completely surrounded by edges are allocated to a segment. This strategy results in two types of pixels: edge pixels and segment pixels. Small segments may end up with complete edge pixels. Thus this strategy is not suitable in areas with varying density vegetation. Region-based segmentation splits entire image into lots of small polygons, then separate polygons are interatively merged to one segment as long as it satisfies the homogeneity criteria. In remote sensing applications, this criteria looks at the spectral homogeneity. In our project, region-based segmentation was adopted and processed in QGIS. The software QGIS will be introduced in detail in the end of this chapter.

Image segmentation has been used in machine learning techniques. Features can be extracted from segments or sliding windows and then serve as input of the classifier (Cheng and Han, 2016). Com-

monly used features in remote sensing applications are reflectance, vegetation indices, and texture features. Reflectance is digital value from satellite or airborne images that have been calibrated and atmospherically corrected as described in section 2.2.4. Vegetation indices are spectral characteristics of different materials, as described in section 2.3. Texture features refers to the appearance of a segment given be the shape, density and spatial arrangement, including GLCM and LBP. In segment-level feature extraction, spectral features are processed as statistics, including mean, standard deviation and moments (Li et al., 2010; Wood et al., 2012); arbitrary-shaped segments are extended to rectangular area by padding zero or mean value when extracting texture features (Li et al., 2010). The feature vector of each segment serves as an input of the classifier. In sliding-window feature extraction for vegetation mapping, features are extracted for individual pixels. For each pixel, its features are extracted in a user-defined window size such as 7×7 (Murray et al., 2010). The window is moving until features of all pixels are extracted (Wang et al., 2018). The feature vector of each pixel serves as an input of the classifier.

2.8. Classification of aquatic plant species

Various supervised classifiers have been used to classify aquatic plant species. The Maximum Likelihood (ML) classifier calculates the likelihood that a pixel belongs to a specific class and assigns the pixel to the class with maximum probability (Pande-Chhetri et al., 2014). The Support Vector Machine (SVM) maps input data from the initial space to a usually significantly higher dimensional space, then find hyperplanes with the largest margin to separate from the whole training data (Li et al., 2010). The Spectral Angular Mapper (SAM) calculates the angle between the two spectra and compares the angle with reference spectrum, and assigns the pixel to the reference class that yields the smallest angle (Kruse et al., 1993; Pande-Chhetri et al., 2014). Random Forest (RF) classifier is an ensemble classifier based on a large number of decision trees, in which each tree contributes one vote, and the final classification results are obtained by maximum voting (Zhou et al., 2018). Pande-Chhetri et al. (2014) found that SVM and SAM performed better than the traditional ML classifier in classifying (of 10% higher accuracy) and detecting (of 20% higher accuracy) aquatic vegetation. The choice of classifiers is very important for the identification of vegetation classes. The RF classifier has shown to be able to achieve high classification accuracy even when applied to analyze data with stronger noise (Breiman, 2001; Dietterich, 2000). This advantage can be beneficial for large water bodies with various local conditions. These noise, on the other hand, may affect the process of selecting the SVM hyperparameters and further decrease the SVM performance (Nalepa and Kawulok, 2018). Therefore, we are going to explore the possibilities of RF classifier for aquatic vegetation mapping application.

2.9. QGIS

QGIS is a free and open-source geographic information system (GIS) desktop application[4] with a graphical user interface. QGIS can be used to display map layers of the real world such as forests or lakes; embed spatial information into non-geographical images; edit and analysis geographical images. QGIS supports many types of data, such as vector data and raster data. Vector data is used to represent points, lines and areas in a certain Coordinate Reference Systems (CRS) such as using latitude and longitude data. Raster data consist of grid cells from aerial photography or satellite imagery, where each cell contains information such as reflectance. Web services, open-source GIS packages, including GRASS GIS, and MapServer, and many APIs such as Python, C and GDAL are also integrated into QGIS.

2.9.1. Raster analysis

QGIS allows various analysis based on rasters. For example, calculations based on existing raster pixel values, such as vegetation indices calculations, can be performed in Raster Calculator function. Several raster inputs can be re-projected to the same CRS and resampled to the same cell size or spatial resolution. In QGIS, users can also do conversions between vectors and rasters; create polygons from a raster and export shapefiles; and clip regions of interest using coordinates or a mask layer.

2.9.2. Semi-Automatic Classification

The Semi-Automatic Classification Plugin (SCP) is a free open source plugin for QGIS. It provides several tools for the download of free and commonly used satellite images (Landsat, Sentinel-2, Sentinel-3, ASTER, MODIS). It also allows preprocessing and classifications of satellite images.

2.9.3. GRASS GIS

GRASS GIS is also a Geographic Information System (GIS) software and can be integrated in QGIS. Generally, QGIS more focuses on simple raster calculations and visualization, while GRASS GIS is more suitable for larger scientific data computations, such as segmentation.

[4] `https://en.wikipedia.org/wiki/QGIS`

3

Materials and Methods

3.1. Study area

The Randmeren, also called border lakes in English, are a series of lakes around the Flevopolder that aim to isolate the water management of the polder from the surrounding land. The Randmeren consist of the following lakes: IJmeer, Gooimeer, Eemmeer, Nijkerkernauw, Nuldernauw, Wolderwijd, Veluwemeer, Drontermeer, Vossemeer, Ketelmeer, Ramsdiep, Zwarte Meer, Kadoelermeer and Vollenhovermeer[1]. Rijkswaterstaat (RWS) takes responsibility to maintain the clearness and avoid full spread in Randmeren based on European law requirements. Regular monitoring is carried out in the Randmeren every three years. Every time the field survey spends two months. This workload makes it impossible to detect the dynamic vegetation changes during the year. The Randmeren was selected as our study area because there is a need to improve the monitoring efficiency. The whole Randmeren measure 60 kilometres length and 49.8 kilometres width. The water transparency ranges from 0.2 m to 7.3 m with average transparency of 1.1 m. Figure 3.1 shows the region of interest highlighted in dark orange. As the Randmeren consist of several water bodies, local water quality varies a lot. For instance, the avergae water depth of Veluwemeer is 1.55 m. It is fairly shallow and crystal clear; while water in Eemmeer east is often fully covered by algae which makes the water turbidity higher than Veluwemeer.

3.2. Field data

The current vegetation survey is conducted by Rijkswaterstaat every three years from mid June to mid July. Systematic sampling design was applied in this field survey. 2889 study plots were selected

[1] `https://nl.wikipedia.org/wiki/Randmeer_(water)`

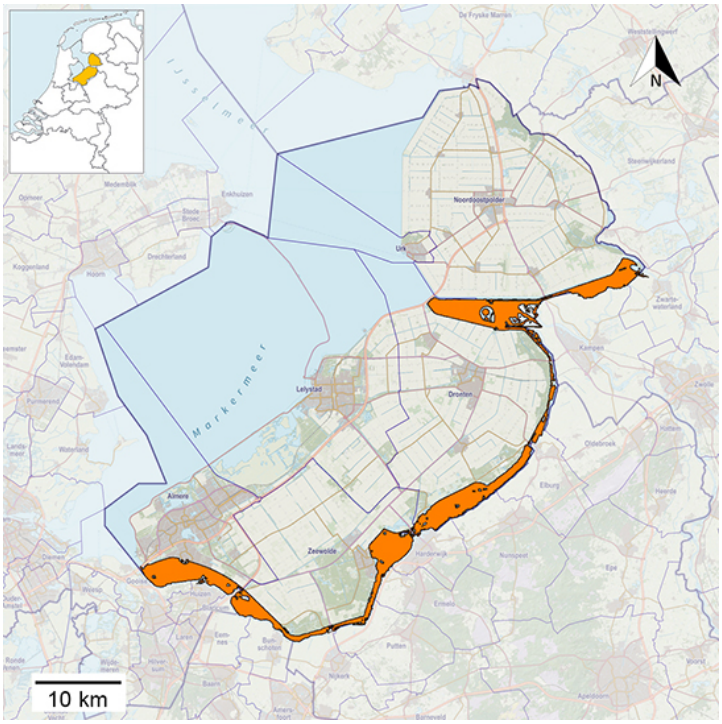


Figure 3.1: The region of Flevoland is highlighted in light orange in upper left map of Netherlands. The study site is highlighted in dark orange. Island area have been masked.

in 2018. These point measurements were carried out on a fixed grid of 200 * 200m, where after a number of empty observations the edges were not measured any further. Each plot covered 1 × 1 square meter, as illustrated in figure 3.2. For each plot, the field survey team recorded its coordinates in Amersfoort coordinate system, water depth, water transparency, cover of 31 plant species and total vegetation cover. The total vegetation cover of all plots and distribution of all aquatic plant species were visualized as shown in figure 3.3. There are four dominant species: Doorgroeid fonteinkruid (DF), Kransblad (KB), Draadwier and Schedefonteinkruid. DF and KB grow the whole water column, which makes shipping harder. Therefore, mapping and monitoring these species is urgent and significant.

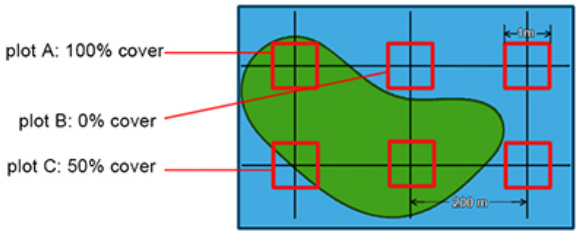
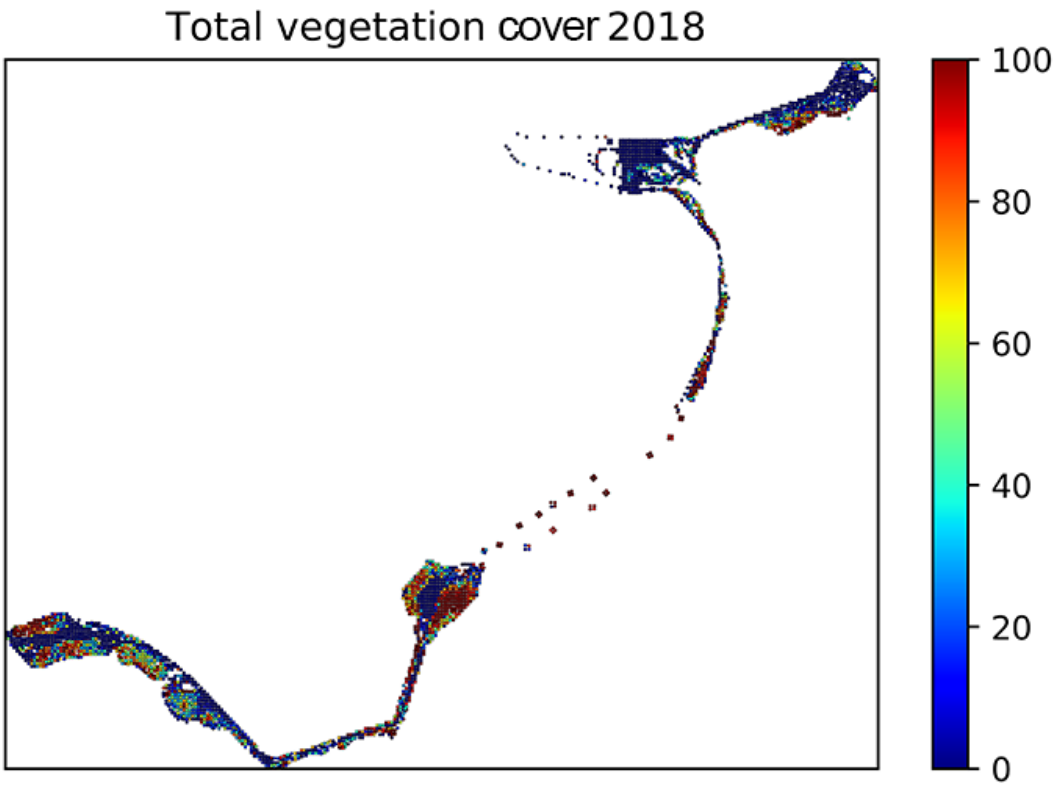
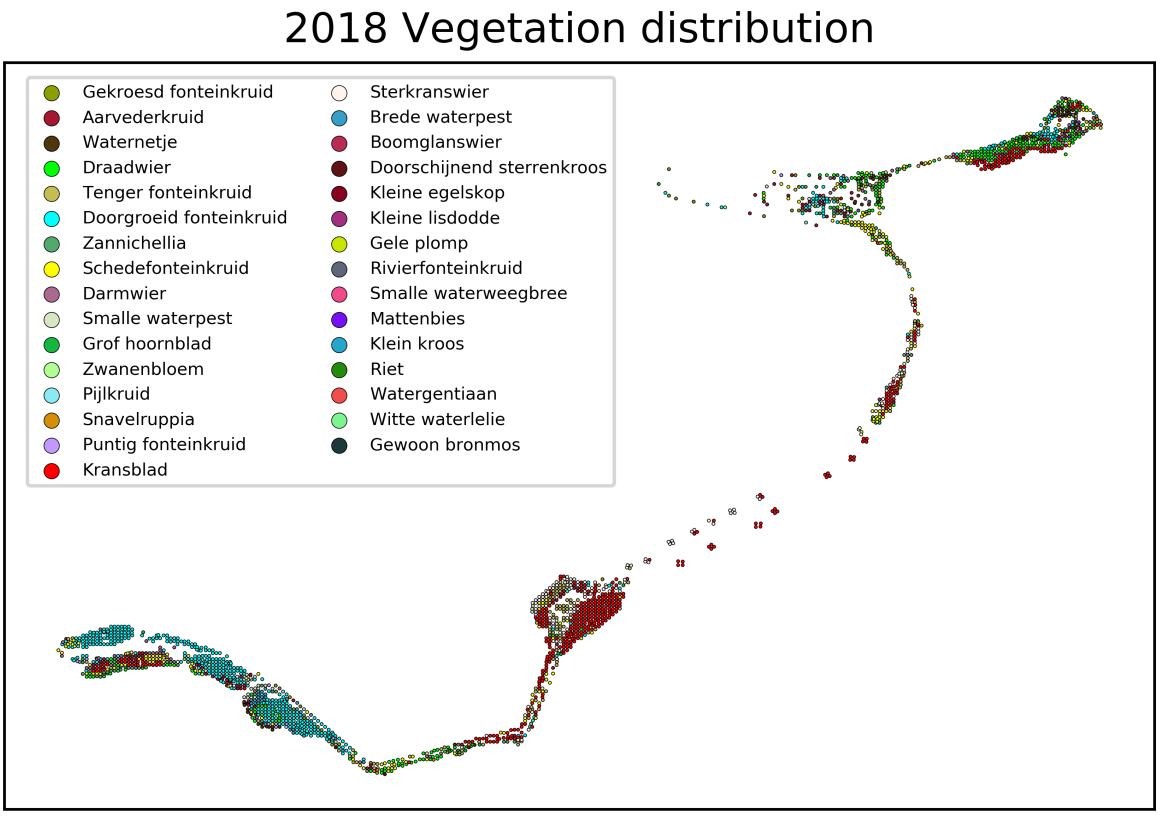


Figure 3.2: The red blocks represent the study plots. Each study plot cover 1 × 1 square meter. The black lines represent the grid where these plots were carried out. All these plots have a fixed distance of 200 m with their neighbor plots.



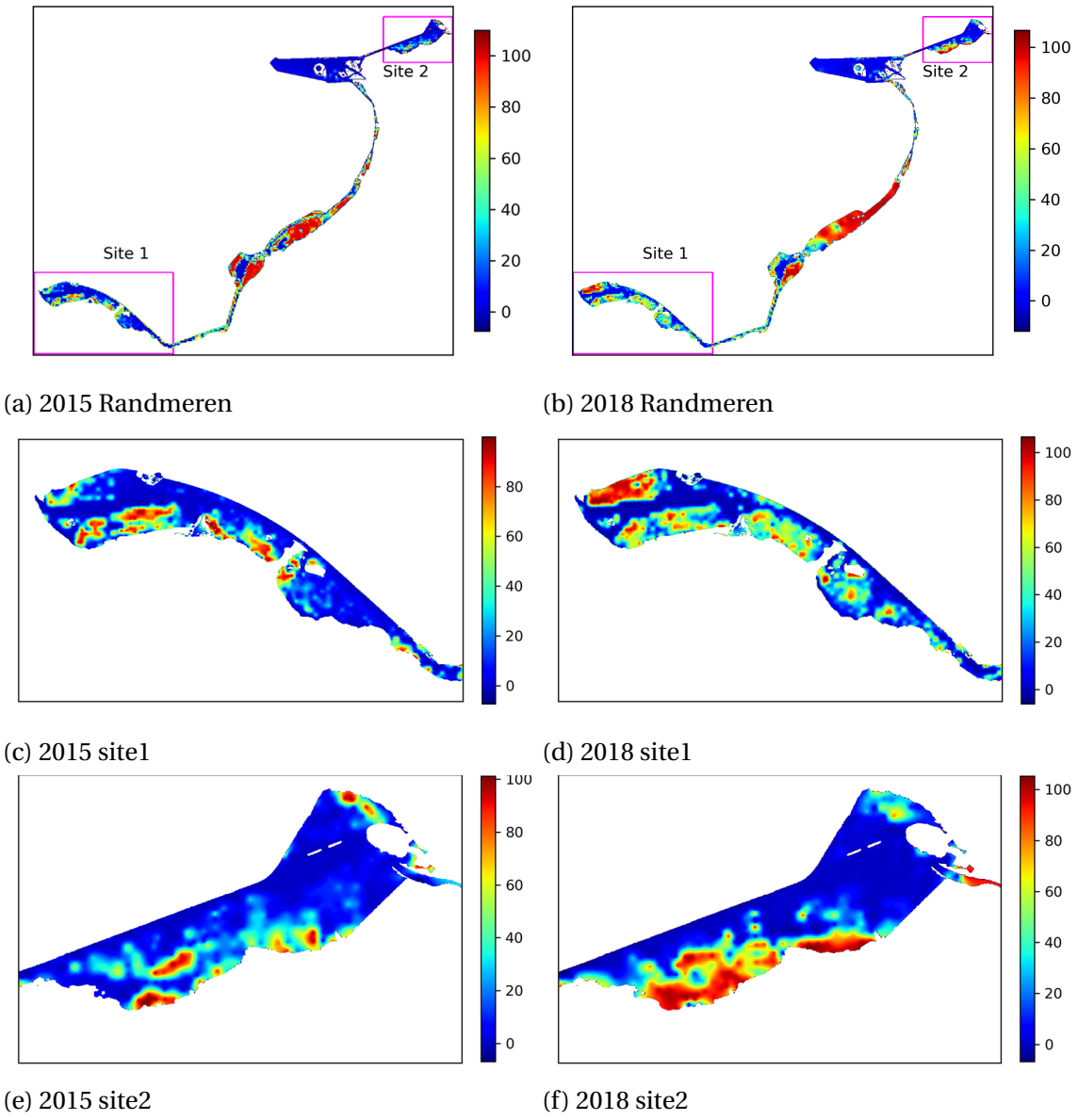
(a) The total cover of each plot ranges from 0 to 100 percent. For veluwemeer region, only edge area were recorded, with knowledge that inner area were also 100 percent covered by the same species.



(b) For mixed plots, species with maximum cover were assigned. The blue, red, green and yellow area represent the dominant species: Doorgroeid fonteinkruid, Kransblad, Draadwier and Schedefonteinkruid.

Figure 3.3: Field observations

A comparison of the field observations in 2015 and 2018 is shown in figure 3.4. In order to view changes clearly, a linear interpolation was utilized. Vegetation cover has extended severely in these three years, especially in two end of Randmeren. Site 1 and site 2 were dominant by Doorgroeid fonteinkruid and Kransblad respectively. They are submerged plants that grow up through water column. Explosive growth of these two species may interfere with the movement of boats and impede the flow of water.



(a) 2015 Randmeren

(b) 2018 Randmeren

(c) 2015 site1

(d) 2018 site1

(e) 2015 site2

(f) 2018 site2

Figure 3.4: Comparison of total vegetation cover in 2015 and 2018

3.3. Remote sensing data

The data was collected on two platforms: satellite and airplane. For satellite images, the product of Sentinel-2 satellites was selected as it was not only of high quality and covered the whole Randmeren, but also freely available. Especially when airborne images were only collected from two test locations and not sufficient for large area study, Sentinel-2 data gave the chance to provide a rough view of the efficiency of remote sensing technologies on aquatic vegetation mapping. Airborne data collection and processing were done by DRO Remote Sensing, a Dutch company that is specialised in acquisition and processing of aerial photos using drones and from manned aircraft. Eemmeer and Veluwemeer were chose as test locations for this pilot, due to the cost, time and access constraints to flying over large open area. Two types of airborne images was collected: hyperspectral images and multispectral images. The file size of both two types of airborne images is very large, which makes storage and processing more difficult than satellite images. The extent of sentinel-2 and two types of airborne images is demonstrated in figure 3.5. Tables below provide an overview of the data collected (table 3.1), the sensor characteristics used (table 3.2) with a general summary of the advantages and disadvantages of the characteristics of these data sources. The images from each source have their own characteristics and data processing methods, which will be described in detail below.

Platforms	Datasets	Camera	Time of collection	Collected area	Number of bands	Coordinate systems
Satellite	Sentinel-2 images	MSI	2018-07-02 2018-08-06	Randmeren	13	WGS84/ UTM zone 31N EPSG: 32631
Airplane	Multispectral images	CanonEOS-100D SONY ICX445	2018-08-03	Parts of the Eemmeer and Veluwemeer	4	WGS84/ UTM zone 31N EPSG: 32631
	Hyperspectral images	Sony IMX_273	2018-08-03	Small parts of the Eemmeer and Veluwemeer	288	WGS84 EPSG: 4326

Table 3.1: Basic information regarding the data sources per platform

Camera	Wavelength	Resolution	Advantages	Disadvantages
MSI	13 bands: 443-2190nm	10m,20m,60m	Large range of wavelengths	Low resolution
CanonEOS-100D	3 bands:RGB	0.2 - 0.5m	High resolution	Limited amount of bands
SONY ICX445	1band: Near infrared	0.5 - 1.3m		
Sony IMX_273	288 bands: 449-944 nm	1.2 - 2m	Having a higher level of spectral detail	Not suitable for large areas

Table 3.2: Detailed information about the cameras used

3.3.1. Sentinel-2 images

The Sentinel-2 mission comprises a constellation of two polar-orbiting satellites placed in the same sun-synchronous orbit, phased at 180° to each other, launched as part of the European Commis-

Figure 3.5: The yellow frames show the extent of hyperspectral images. 'hyper 1' represents the hyperspectral image that was collected in Eemmeer and 'hyper 2' represents the hyperspectral image that was collected in Veluwemeer. The red frames show the extent of multispectral images. 'multi 1' represents the multispectral image that was collected in Eemmeer and 'multi 2' represents the multispectral image that was collected in Veluwemeer. The background image was clipped from the True Colour Image (TCI) built from the B02 (Blue), B03 (Green), and B04 (Red) Bands of Sentinel-2 data, which covered the whole region of interest.

sion's Copernicus program on June 23, 2015. The mission provides a global coverage of the Earth's land surface (within the latitude range of -56° to +83°) every 10 days with one satellite and 5 days with 2 satellites, which allows to reach a 5-day revisit time and makes the data of great use in ongoing studies. The satellites are equipped with a Multispectral Instrument (MSI) of a 290 km wide swath. The Sentinel-2 mission offers 13 spectral bands from visible light, via near-infrared (NIR) to short wave-infrared (SWIR) with different resolutions varying from 10 to 60 meters on the ground as shown in table 3.3. The Sentinel-2 data are georeferenced in UTM/WGS84 projection. The launch of Sentinel-2 satellites aims at land management (inland waterways and coastal areas), agricultural production and forestry, and monitoring of natural disasters.

The Sentinel-2 data were acquired from July 2 and Aug 6, 2018. Data from August were only used

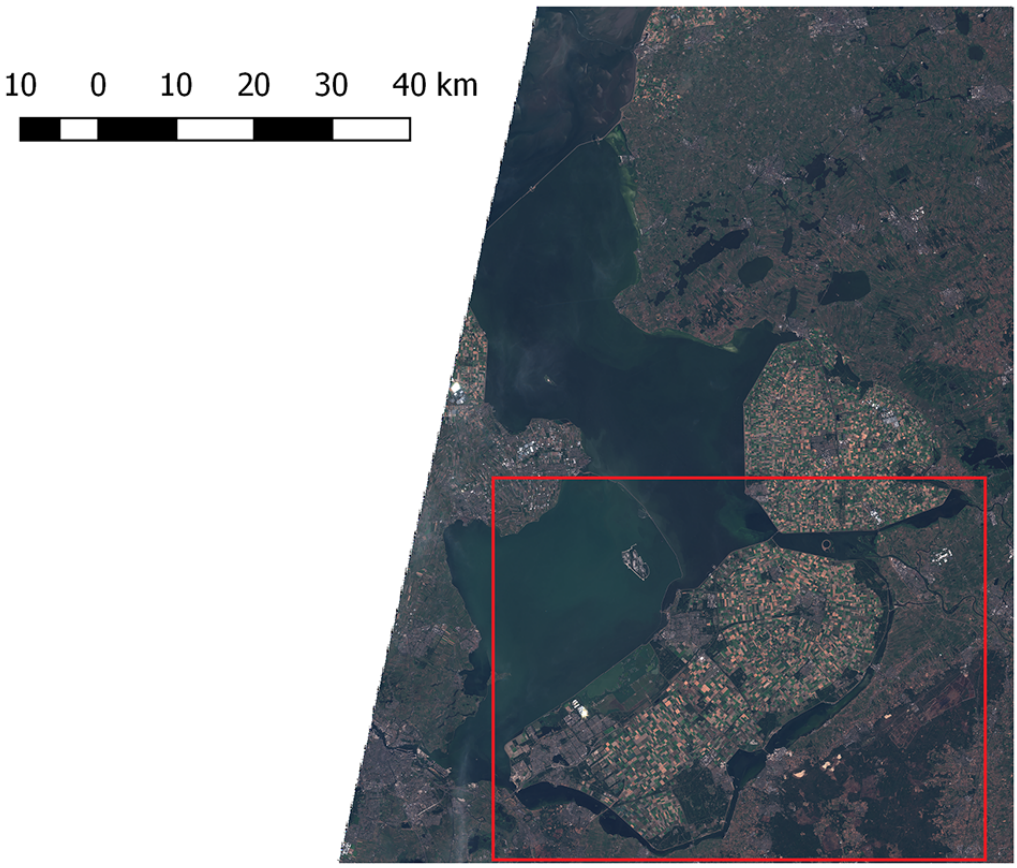
for testing purpose. This is to estimate the applicability of our approach on future days because data from the future cannot be used in the training process of our model. The sentinel-2 data were downloaded from EarthExplorer for free. The Sentinel-2 data here had been processed in Level-1C. Level-1C processing includes radiometric and geometric corrections along with ortho-rectification to generate highly accurate georeferenced products. Figure 3.6a shows the example of True Color Image (TCI) of Sentinel-2 data and the area in the red frame was the region of interest in this study. All the 13 bands were clipped into this area and resized to 10m resolution of 4800×6000 dimensions(figure 3.6b).

Band name	Resolution (m)	Central wavelength(nm)	Band width(nm)	Purpose
B01	60	443	20	Aerosol detection
B02	10	490	65	Blue
B03	10	560	35	Green
B04	10	665	30	Red
B05	20	705	15	Vegetation classification
B06	20	740	15	Vegetation classification
B07	20	783	20	Vegetation classification
B08	10	842	115	Near infrared
B08A	20	865	20	Vegetation classification
B09	60	945	20	Water vapour
B10	60	1375	30	Cirrus
B11	20	1610	90	Snow / ice / cloud discrimination
B12	20	2190	180	Snow / ice / cloud discrimination

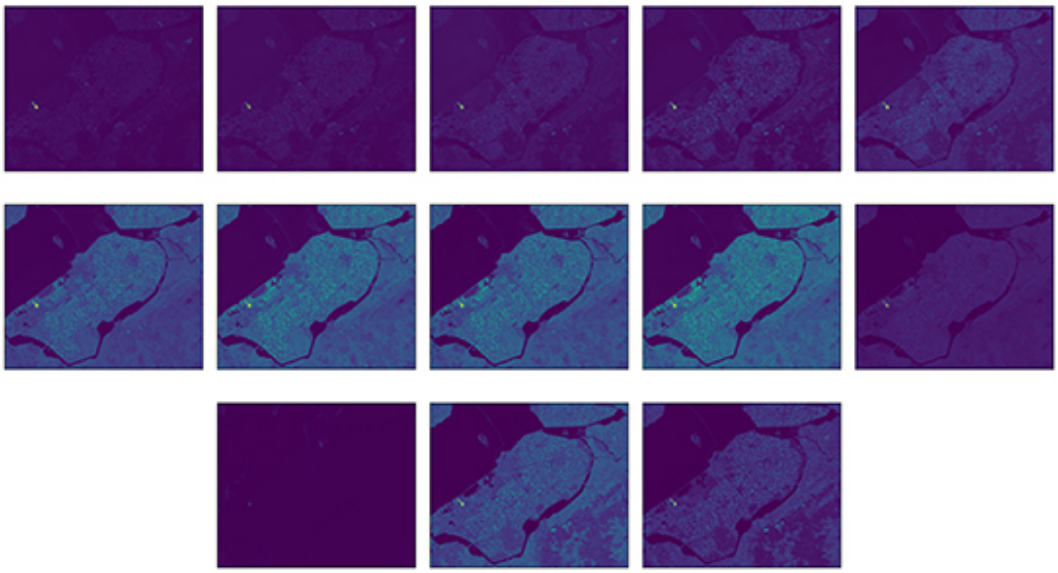
Table 3.3: Overview of 13 bands information in Sentinel-2 data

3.3.2. Airborne multispectral images

To capture the multispectral images, three separate cameras were used simultaneously to capture the multispectral images: an RGB camera and two separate cameras for the NIR 735 and NIR 850 bands. The RGB images were taken with a Canon EOS-100D camera with a 24 mm lens.



(a) Sentinel-2 images cover very large area, the region used in this study is indicated in the red frame and other areas were ignored.



(b) Sentinel-2 data: all the 13 bands have been clipped into the same region.

Figure 3.6: Sentinel-2 images

The images were captured with a fixed exposure time of 1ms and automatic ISO and F number adjustment. The NIR 735 and NIR 850 images were taken with a SONY ICX445 camera with a 6 mm c-mount lens and an F8 aperture setting. The images are georeferenced in UTM/WGS84 coordinate system by time synchronization with GPS data. Separate overlapping images were stitched together with the help of georeferencing. The raw RGB images were then binned 3x3 to lower the resolution and increase signal to noise ratio. GPS positions were synchronised with capture time and imported for mapping with Pix4D. Pix4d processed the orthomosaic maps and finally exported them as Geo TIFF images. The 12 bit raw NIR images were exported to 16 bit gray-scale tiff images together with their GPS positions to be imported for mapping with Pix4D. Pix4D processed the orthomosaic maps and finally exported them as Geo TIFF images.

Airborne multispectral images in Eemmeer and Veluwemeer are shown in figure 3.8. The blank areas inside the multispectral images in Eemmeer was caused due to difficulties in stitching NIR735 image with super dark area or large difference in signal between the banks and the water. Table 3.4 gives information about resolution and dimensions in detail. As we can see, the resolution varies as it depends on the altitude that was flown and the flight height was about 300 m for the Eemmeer and about 500 m for the Veluwemeer. The resolution of the images is on average about 0.22×0.22 m for the RGB bands in the Eemmeer and 0.5×0.5 m in the Veluwe due to the flight height, which differed for the Eemmeer and the Veluwemeer. For the 735 nm, a resolution of approximately 0.4×0.4 m was achieved in the Eemmeer and 1.3×1.3 m in the Veluwemeer. Because the airplane flew higher above Veluwemeer, it was also easier to pitch the images and therefore the 850nm band could also be stitched.

3.3.3. Airborne hyperspectral images

The hyperspectral camera selects the middle scan line from a 2D image using an entrance slit in the camera. This thin strip of light is broken down in the camera by a prism on a spectral axis. 500 scan lines are collected with a shutter speed of 60 images per second. This creates a hyperspectral data set containing 2 spatial axes and 1 spectral axis. Data of these 500 scan lines were then split into 288 bins ranging from 449 nm to 944 nm as 288 bands in total. The resolution of these images is, depending on the flown height, between 1×1 and 2×2 meters, as shown in table 3.5. The hyperspectral images were collected for small parts of the Veluwemeer and Eemmeer to see whether the vegetation types could be better distinguished with these images than with the multispectral cameras. Figure 3.9 shows an example of this data in a gray-scale image. The image is somewhat distorted, which is the characteristic of these images because a line-scanner hyperspectral camera has been used, and is caused by the pitch and roll of the aircraft. The hyperspectral data is not presented as radiometric output. To normalize the hyperspectral data the data had been processed to relative spectral reflection with the use of a white reference. The Relative Spectral Reflection is

Locations		Wavelength	Spatial resolution	Dimensions
Eemmeer	Eemmeer west	RGB	0.2296m	X:57067 Y:29665 Bands: 3
		735nm	0.56873m	X:16398 Y:12493 Bands: 1
	Eemmeer east	RGB	0.22527m	X:47224 Y:18433 Bands: 3
		735nm	0.27955m	X:37278 Y:15672 Bands: 1
Veluvemeer		RGB	0.5m	X:20899 Y:20363 Bands: 3
		730nm	1.30673m	X:8468 Y:9362 Bands: 1
		850nm	1.30016m	X:8305 Y:9045 Bands: 1

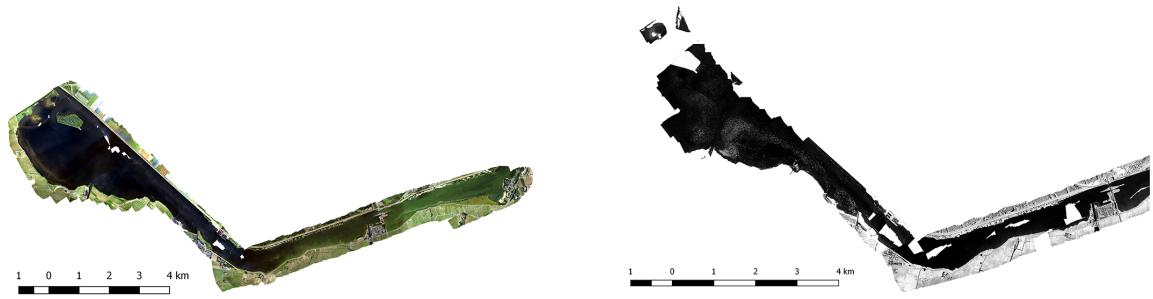
Table 3.4: Overview of airborne multispectral images(the resolution differs as it depends on the flight height.)

processed as follows:

$$Relative\ Spectral\ Reflection = \frac{(RawData - DarkLevel)}{(WhiteRefData - DarkLevel)} \tag{3.1}$$

Locations	Wavelength	Spatial resolution	Dimensions
Eemmeer	449-944nm	1.97124m	X: 1079 Y: 943 Bands: 288
Veluvemeer	449-944nm	1.23193m	X:1801 Y:998 Bands: 288

Table 3.5: Overview of airborne hyperspectral images(the resolution differs as it depends on the flight height.)

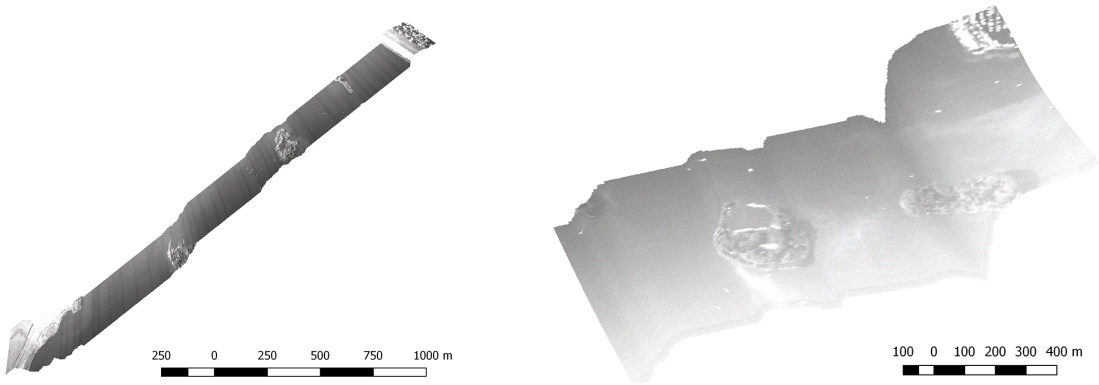


(a) RGB bands (b) NIR735 band

Figure 3.7: Airborne multispectral images in Eemmeer of total 4 bands

(a) RGB bands (b) NIR730 band (c) NIR840 band

Figure 3.8: Airborne multispectral images in Veluwemeer of total 5 bands



(a) Hyperspectral images of 288 bands in Eemmeer (b) Hyperspectral images of 288 bands in Veluwemeer

Figure 3.9: Airborne hyperspectral images in Eemmeer and Veluwemeer

3.4. Methods

This section described our works in two parts. The first part is to compare the basic differences between Sentinel-2 data and airborne data. This part will give hints to the research question that whether Sentinel-2 data can be an alternative to airborne data. Detailed approach is discussed in section 3.4.1. The second part, which is also the main part of our work, describes classification approaches in details in section 3.4.2-3.4.8. This part aims to answer the research questions that whether we can extract unique properties from Sentinel-2 data and what classification results Sentinel-2 data can yield, meanwhile, the contributions of spectral and texture features will also be analyzed.

Part I

3.4.1. Remote sensing data comparison

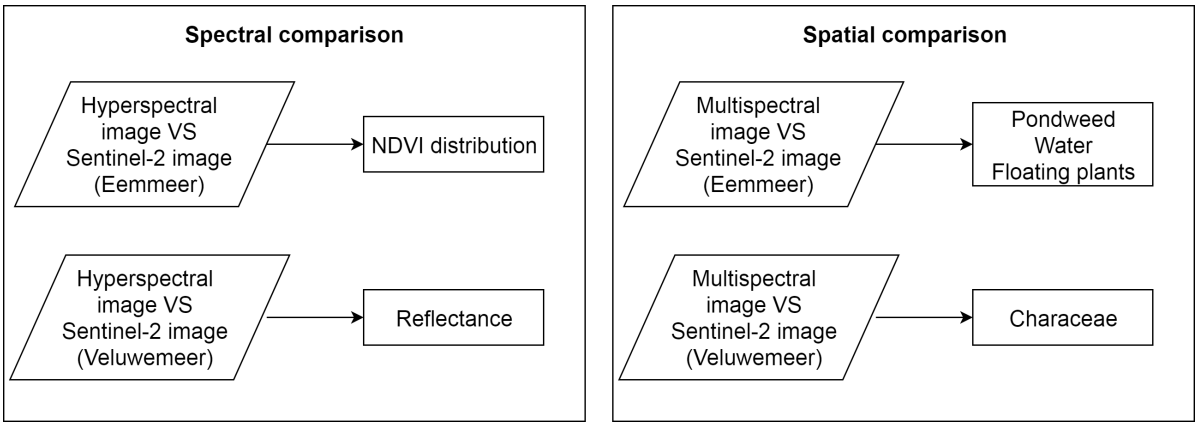
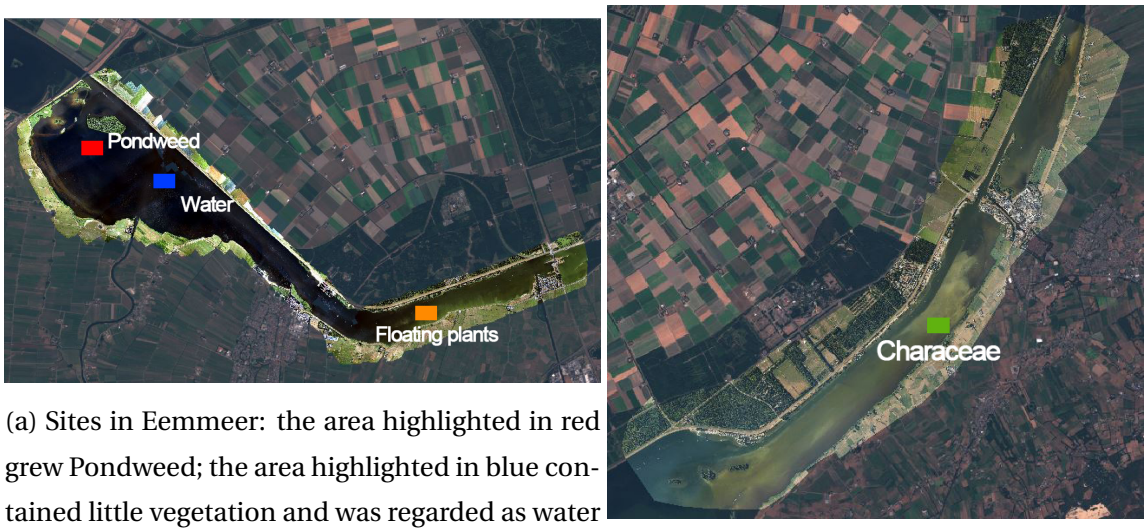


Figure 3.10: Part I workflow of remote sensing data comparison

As general information of remote sensing data introduced in section 3.3, we have known that airborne multispectral data are of highest spatial resolution while airborne hyperspectral data are of highest spectral resolution. Sentinel-2 data, however, are of moderate quality in both aspects. This section aims to investigate whether these properties may lead to large differences on the performance. Therefore, we compared spectral differences between Sentinel-2 and airborne hyperspectral data, and spatial differences between Sentinel-2 and airborne multispectral data.

Figure 3.10 is the workflow of data comparison. The regions covered by airborne multispectral data contain several classes, including water body, Pondweed, floating plants and Characeae. Four sites were selected for spatial comparison in both Sentinel-2 data and airborne multispectral data.(figure 3.11). As water, aquatic plants and island in veluwemeer were easier to be visually discriminated using

hyperspectral data, a site in veluwemeer was selected for reflectance comparison with Sentinel-2 data(figure 3.12). Hyperspectral data in Eemmeer covered a smaller region than that in Veluwemeer, the whole region that covered by hyperspectral data was selected for NDVI distribution comparison with Sentinel-2 data.



(a) Sites in Eemmeer: the area highlighted in red grew Pondweed; the area highlighted in blue contained little vegetation and was regarded as water body; the area highlighted in orange grew floating plants.

(b) Sites in Veluwemeer: the area highlighted in green grew Characeae.

Figure 3.11: Sites for spatial comparison: the background image is Sentinel-2 TCI while the front image is airborne multispectral RGB image. The same sites were clipped from both Sentinel-2 data and airborne multispectral data.

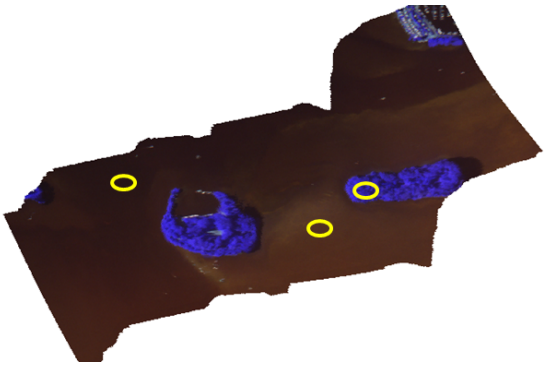


Figure 3.12: Sites for spectral reflectance comparison: this is a false color hyperspectral image. The areas in these three yellow circles were water, island and aquatic vegetation respectively. The same sites were clipped from both Sentinel-2 data and airborne hyperspectral data.

Part II

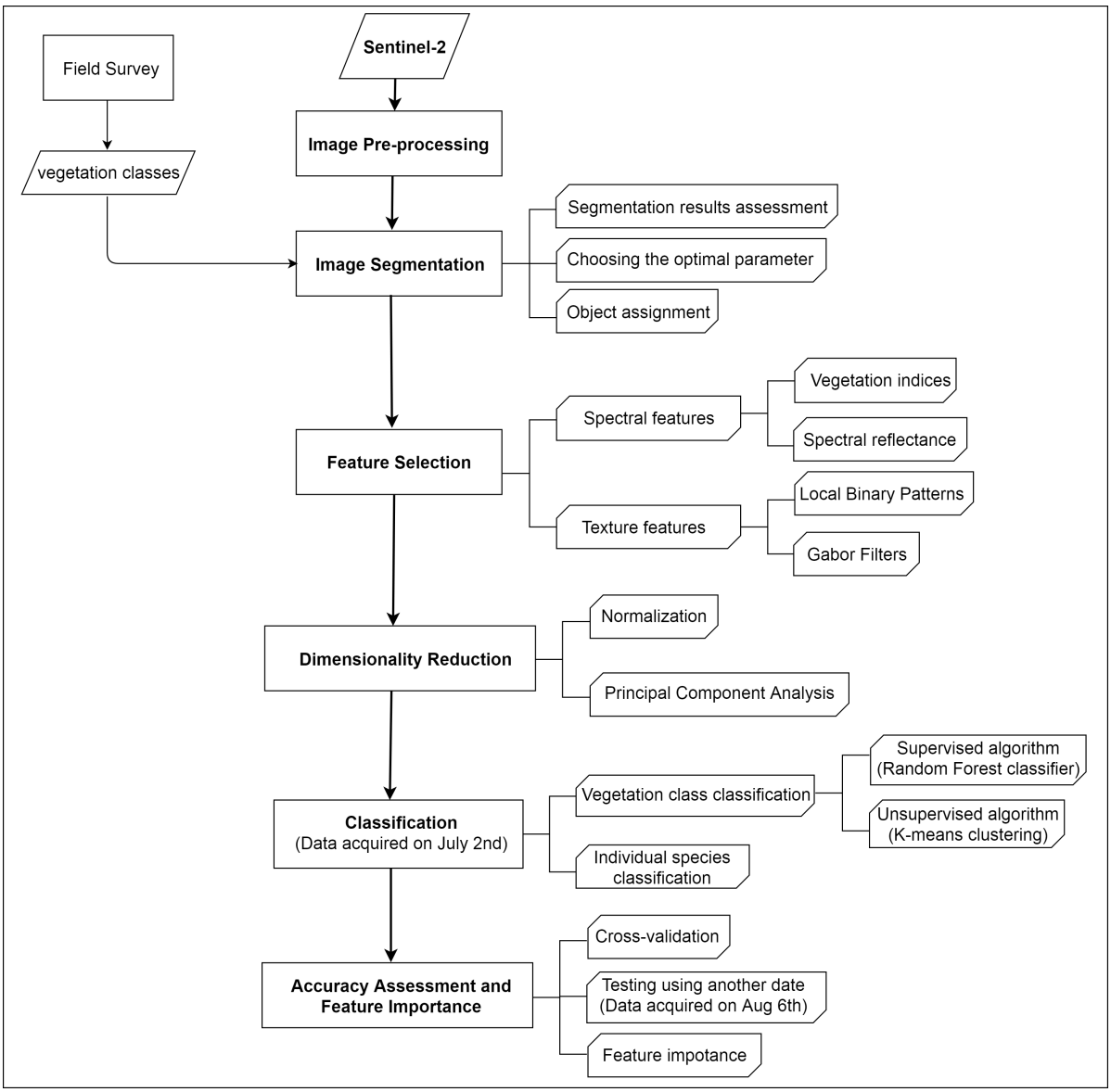


Figure 3.13: Part II workflow of aquatic plant species classification

The overview of the methodology and workflow are presented in figure 3.13. The remote sensing data we used were Sentinel-2 images, as airborne images only cover small parts of Randmeren and are insufficient to differentiate all categories. The main methodological procedures are as follows: (1) Process field data; (2) pre-process Sentinel-2 images by using the SCP plugin in QGIS; (3) partition the Sentinel-2 images into meaningful segments by using GRASS GIS toolbox in QGIS, select the optimal parameter and assign labels with the help of field samples; (3) extract spectral features including reflectance and vegetation indices, and texture features including Local Binary Pattern and Gabor filters; (4) normalize and reduce dimensions of reflectance and texture feature vectors

respectively; (5) perform two classification tasks separately including vegetation class classification and individual species detection, where in the first task both a supervised (Random Forest classifier) and an unsupervised (K-means clustering) machine learning methods were applied and compared; (6) assess the results using 5-fold cross validation, test pre-trained RF model using another date, compare predicted map with map derived from segmentation using confusion matrix and analysis the feature importance.

3.4.2. Field data processing

Figure 3.14 showed the occurrence of all recorded species. The occurrence was calculated by adding percent of cover of each species in all plots. As we can see, Randmeren was mainly dominated by a few species and most of the species have a very tiny cover which is insufficient to be differentiated based on remote sensing data. Thus, study plots with less than 5% total vegetation cover were regarded as water plots, with a total vegetation cover between 5% and 50% were regarded as low density plants while study plots with more than 50% total vegetation cover were regarded as the species with maximum cover. According to this rule, study plots were assigned by 20 specific species, water and low density plants, as listed in table 3.6. Based on morphological properties and amount of samples of these species, they are divided into six classes:

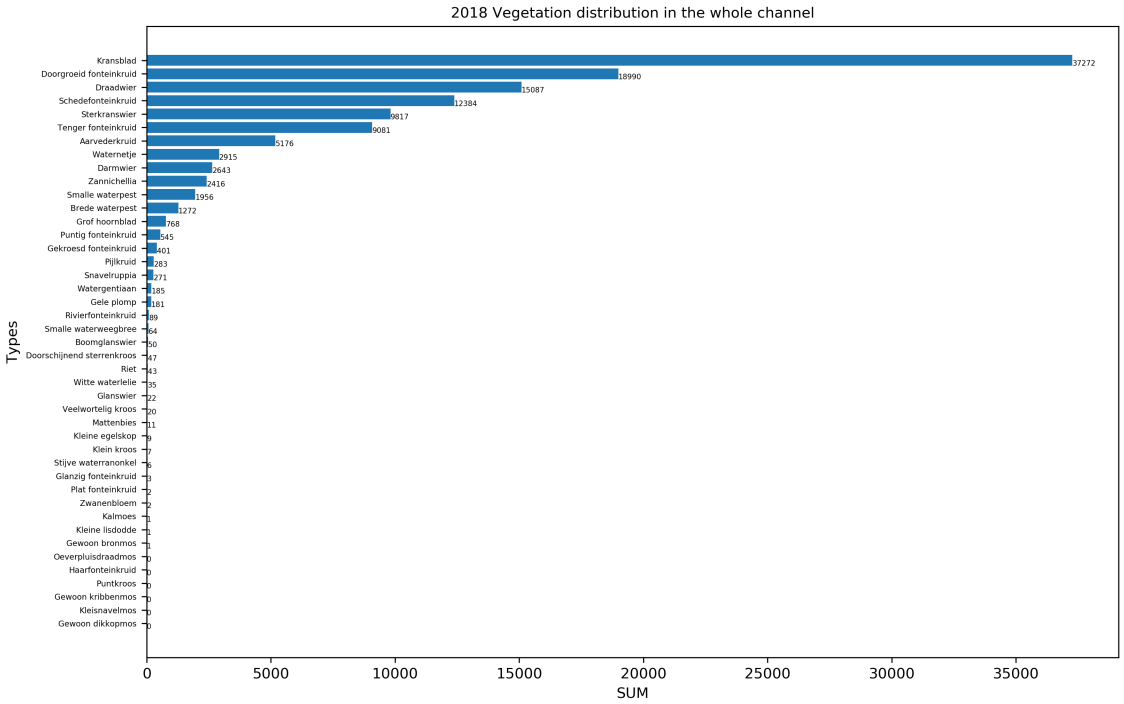


Figure 3.14: The x axis is the frequency of aquatic plant species and was calculated by the sum of cover in all plots and then rounded to the nearest integer. The y axis is the plant species ranking from lowest frequency to highest frequency.

1. Water: of less than 5% total vegetation cover;
2. Low density plants: of total vegetation cover between 5% and 50%;
3. Floating plants: mainly consist of Darmwier and Draadwier from algae family;
4. Characeae family: mainly consist of Kransblad and Sterkranswier;
5. Submerged plants: mainly consist of Hydrocharitaceae and Pondweed families.
6. Pondweed family: mainly consist of Doorgroeid fonteinkruid.

Study plots of mixed plant communities were assigned only by the species with maximum cover. In order to estimate the purity of each plot, average percent of the assigned species in total vegetation cover of all plots were calculated, as explained in equation 3.2. The higher the purity is, the higher the assigned species is concentrated in one plot, and the higher accuracy the plot is assigned with. The purity ranged from 24% to 97% with the average of 70%.

$$Purity_{species} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{C_{assigned}}{C_{total}} \tag{3.2}$$

where $C_{assigned}$ represents the cover of assigned species in plot i, C_{total} represents the total vegetation cover in plot i, n is the number of plots that were assigned by the same species.

Kransblad in Characeae family and Doorgroeid fonteinkruid in Pondweed family are two species that cause most problems. Therefore, we also classified Kransblad within Characeae family and Doorgroeid fonteinkruid within Pondweed family respectively.

Species	Family	Number of samples	Purity	Groups	Total number of samples in groups
Total vegetation<=5%		1042		Water	1042
5%<Total vegetation<50%		577		Low density plants	577
Darmwier	Algae	28	0.62	Floating plants	219
Draadwier	Algae	186	0.71		
Gele plomp	Water lilies	2	0.71		
Pijlkruid	Water-plantains	1	0.5		
Watergentiaan	Water lilies	2	0.97		
Boomglanswier	Characeae family	1	0.4	Characeae family	509
Kransblad	Characeae family	410	0.88		
Sterkranswier	Characeae family	98	0.82		
Aarvederkruid	Haloragaceae	55	0.68	Submerged plants	136
Brede waterpest	Hydrocharitaceae	12	0.6		
Grof hoornblad	Hornwort	4	0.66		
Smalle waterpest	Hydrocharitaceae	12	0.8		
Snavelruppia	Ruppiaceae	2	0.95		
Waternetje	Algae	51	0.81		
Doorgroeid fonteinkruid	Pondweed family	190	0.84	Pondweed family	406
Gekroesd fonteinkruid	Pondweed family	1	0.24		
Puntig fonteinkruid	Pondweed family	4	0.63		
Schedefonteinkruid	Pondweed family	106	0.68		
Tenger fonteinkruid	Pondweed family	88	0.74		
Zannichellia	Pondweed family	17	0.73		

Table 3.6: Overview of field samples.

The coordinate system of these samples was converted from Amersfoort to UTM/WGS84. Thus these samples were in the same coordinate system as Sentinel-2 images. Within 13 bands of Sentinel-2 data, there are three resolutions, including 10 m, 20 m and 60 m. Bands that are larger than 10 m were resized to 10 m-resolution. Then we could locate these samples in Sentinel-2 images. The coverage of one field sample plot was 1 square meter and the coverage of one pixel in Sentinel-2 images was 10 square meters. One pixel could only contain at most one field sample, based on the knowledge that samples were collected on the grid of 200 square meters. Therefore, we assumed that the whole area of 10 square meters contained the same plant species as field samples indicated in individual pixels, and there were 2889 labeled pixels in total;

3.4.3. Image Preprocessing

Sentinel-2 product are already processed in Level-1C, including radiometric and geometric corrections using a Digital Elevation Model (DEM) when it is released, together with a metadata file containing details about the scene and bands. Semi-Automatic Classification (SCP) Plugin in QGIS offers further preprocessing of Sentinel2 images. SCP was applied to convert all bands from radiance to reflectance. In addition, Dark Object Subtraction (DOS) was also applied all 13 bands with the help of the information in the metadata file. The DOS is a simple empirical atmospheric correction method for satellite imagery, which assumes that reflectance from dark objects includes a substantial component of atmospheric scattering. It searches each band for the darkest pixel value. The scattering is removed by subtracting this value from every pixel in the band. This simple technique is effective for haze correction in multispectral data.

3.4.4. Image segmentation

Image segmentation identifies meaningful objects over the remote sensing image and is a critical procedure in the workflow of OBIA. The segmented objects are groups of adjacent pixels that describe particular regions. In case of aquatic environment it could be a region dominated by floating plants or a clear water body for example. In this project, segmentation was done by applying GRASS GIS. GRASS GIS is a Geographic Information System (GIS) software and is integrated in QGIS. *i.segment* is an segmentation tool in GRASS GIS and was applied in this study. We selected four bands from Sentinel-2 images, including blue, green, red and near infrared bands. In addition, a region growing algorithm was selected, which starts at random points with one-pixel objects and then merges them into bigger objects based on their similarity. The similarity between current object and its neighbors was calculated according to Euclidean distance. Smaller distance values indicate a closer match. During the processing, objects are merged only when the similarity is lower than a threshold. The threshold must to be between 0 and 1. A threshold of 0 would allow only one-pixel objects to be merged, while a threshold of 1 would allow everything to be merged. Here we set the threshold to 0.5. The minimum number of cells in a segment refers to the minimum number of

pixels in an object.

After the segmentation process was finished, each individual object was labeled by the maximum categories if there were pixels labeled by field samples located within the this object, while if there were no labeled pixels located within the object, the object would not be labeled. Below are two example objects after segmentation process (figure 3.15). Objects were labeled based on the assumption that all pixels in individual objects were from the same categories. Purity was used here to estimate homogeneity for each labeled object. The purity was calculated as explained in equation 3.3. In the first classification task, objects were labelled by six categories: water, low density plants, floating plants, characeae, submerged plants and pondweed. In the second task, objects labelled by characeae were then re-labelled by Kransblad and other characeae, while objects labelled by pondweed were then re-labelled by Doorgroeid fonteinkruid and other pondweed. Other objects were ignored in this problem species detection task.

$$Purity_{object} = \frac{n}{N} \tag{3.3}$$

where n is the number of pixels labeled by the maximum category and N is the total number of pixels labeled in all categories.

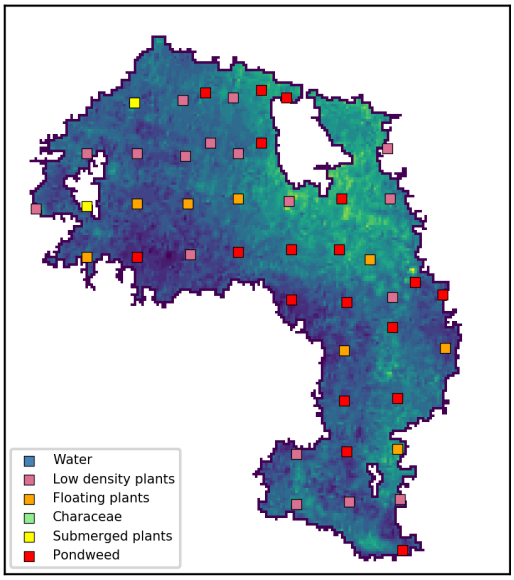
Different parameter settings can lead to segmentation results of different quality. The minimum number of pixels in an object indicates the minimum scale of vegetation distribution pattern and thus is necessary to choose the optimal value. Given three settings of the minimum number of pixels in an object, including 1 , 10 and 30, we compared the segmentation results in following aspects: (1) the number of objects that can be labeled, as the more labeled objects we have, the more objects can be used for validation; (2) average purity of each category, as the higher the purity is, the better it matches the assumption that all pixels in individual objects are from the same categories; (3) differences with field observations. Then the value with best performance was selected.

3.4.5. Feature selections

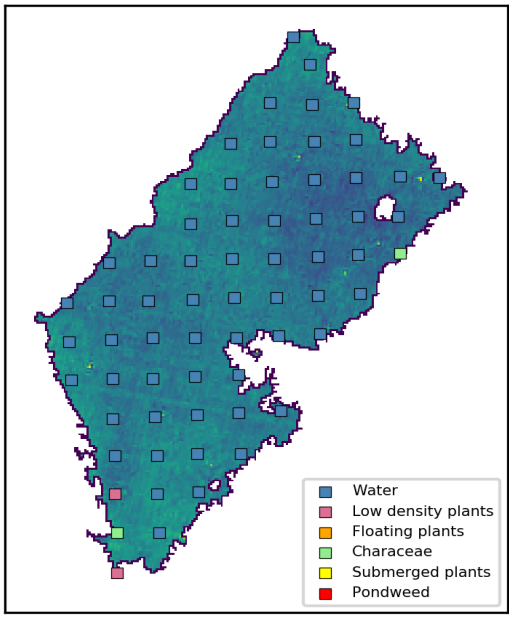
Sentinel-2 images contain only 13 bands and are of moderate spatial resolution. They are neither of high spectral resolution as airborne hyperspectral images nor of high spatial resolution as airborne multispectral images. Therefore, spectral features and texture features corresponding to these two characteristics respectively were investigated to see how well these features extracted from Sentinel-2 images perform in aquatic plant classification. Two types of spectral features were selected including reflectance and vegetation indices, and two types of texture features were selected including Local Binary Patterns and Gabor filters.

Spectral reflectance

Sentinel-2 images contain thirteen bands in visible and NIR and SWIR spectrum. Some of these spectral bands are likely to be closely correlated with others, which may end up with much redundant information. This redundant information can significantly reduce classification accuracy as



(a) A low-purity object: there are 45 pixels that are labeled by field samples, including 17 pixels labeled by low density plants, 8 pixels labeled by floating plants, 2 pixels labeled by submerged plants and 18 pixels labeled by pondweed family. Pixels labeled by pondweed family category counts the most. Therefore this object is labeled by pondweed family. The purity of this object is 0.4.



(b) A high-purity object: there are 65 pixels that are labeled by field samples, including 61 pixels labeled by water, 2 pixels labeled by low density plants and 2 pixels labeled by characeae. Pixels labeled by water category counts the most. Therefore this object is labeled by water. The purity of this object is 0.94.

Figure 3.15: Two examples of object assignment and their purity calculations

well as increase the computational resources required to carry out the classification process (Murray et al., 2010). Thus, a dimensionality reduction method was applied and will be introduced in the next section.

Vegetation indices

Vegetation Indices (VIs) are combinations of surface reflectance at two or more wavelengths. Each of VIs is designed to accentuate a particular vegetation property. More than 100 VIs have been published in scientific literature and implemented in a variety of applications using different satellite and airborne platforms (Jinru and Su, 2017). These indices can be grouped into many categories such as broadband greenness, narrowband greenness and canopy water content. The broadband greenness VIs are designed to measure the quality of photosynthetic material and vigor in vegetation, which is essential for understanding the state of vegetation. Broadband greenness VIs com-

pare reflectance peak in NIR range to reflectance in the red range, allowing sounding of the total amount of green vegetation in the column until the signal saturates at very high levels. The narrowband greenness VIs combine reflectance in red and NIR regions to sample red edge portion of the reflectance curve. This curve is caused by the trainsition from chlorophyll absorption and NIR leaf scattering. Measurements in red edge allows narrowband greenness VIs to be more sensitive to smaller changes in vegetation health than broadband greenness VIs. The canopy water content VIs measure the amount of water contained in the foliage canopy. These indices use reflectance in NIR and SWIR regions to take advantage of known absorption features of water and the penetration depth of light. In our study, nine indices were selected including Enhanced Vegetation Index (EVI), Green Normalized Difference Vegetation Index (GNDVI), Percentage Vegetation Index (PVI) and Normalized Difference Vegetation Index (NDVI) from broadness greenness VIs, Modified Chlorophyll Absorption Ratio Index (MCARI) from narrowband greenness VIs, Moisture Stress Index (MSI) and Normalized Difference Water Index (NDWI) from canopy water content VIs, and two Angle Index (Ma and Zhou, 2018). They were calculated as equations 3.4-3.12:

$$EVI = 2.5 \times \frac{NIR - RED}{(NIR + 6 \times RED - 7.5 \times BLUE) + 1} \tag{3.4}$$

$$GNDVI = \frac{NIR - GREEN}{NIR + GREEN} \tag{3.5}$$

$$PVI = \frac{NIR - a * RED - b}{\sqrt{1 + a^2}} \tag{3.6}$$

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{3.7}$$

$$MCARI = ((B5 - B4) - 0.2 \times (B5 - B3)) \times (\frac{B5}{B4}) \tag{3.8}$$

$$MSI = \frac{B11}{B8} \tag{3.9}$$

$$NDWI = \frac{NIR - SWIR}{NIR + SWIR} \tag{3.10}$$

$$AngleB3 = arccos \frac{\vec{B_2B_3} \cdot \vec{B_3B_4}}{|\vec{B_2B_3}| \times |\vec{B_3B_4}|} \tag{3.11}$$

$$AngleB4 = arccos \frac{\vec{B_3B_4} \cdot \vec{B_4B_5}}{|\vec{B_3B_4}| \times |\vec{B_4B_5}|} \tag{3.12}$$

where the parameters of PVI were set to be a=0.96916, b=0.084726 according to literature (Seo et al., 1998), NIR refers to the band in NIR spectrum region, B5 refers to band 5 in Sentinel-2 data, $\vec{B_2B_3}$ refers to the vector from band 2 to band 3 in wavelength-reflectance dimensions, and similar way for other variables.

As each of these indices is already a combination of spectral bands and has a particular purpose, we will not process it by demension reduction as other features.

Local Binary Patterns

Local Binary Patterns (LBP) is simple yet efficient texture descriptor. It is known by invariance to gray-scale changes and high discriminative power. For each pixel in an image, the LBP value is computed by considering an 8-neighbourhood around a center pixel and then comparing the value of this pixel with the eight neighbouring pixels. If the reflectance value of the neighbouring pixel is greater than the center pixel, then '1' is assigned for this location, otherwise a '0'. An 8-bit binary pattern is obtained by traversing the 8 neighbours clockwise or anti-clockwise, which is then converted to integer, as demonstrated in figure 3.16. Processing entire images similarly will then result in an LBP feature matrix.

Texture features are typically applied to each of bands in case of multispectral images. The disadvantage is the increased size of the feature vector and highly correlated texture features because of the redundancy between spectral bands. In particular, we computed the normalized difference vegetation index (NDVI) image as our target image and extracted texture features from it. we extracted LBP features using a sliding window of size 7×7. A histogram was computed for the center pixel in each sliding window to obtain a 256-dimensional feature vector, there were 2^8 combinations of binary codes in the case of 8-neighbourhood. For edge pixels or pixels around island, their sliding windows may contain pixels that have been masked. In this case, padding mean value in these locations was adopted, which has shown better performance padding zero (Liu et al., 2006). A sparse feature matrix will be obtained because each 256-dimensional feature vector is derived from a sliding window of only 49 pixels. Therefore it is necessary to reduce its dimensions.

Gabor filters

Gabor filters are linear filters used for texture analysis. An advantage of Gabor filters is their invariance to scale and rotation. Gabor filters encodes edges, where each is sensitive to a different orientation and scale, generates multiple responses for all textures in the image. It has been claimed that simple cells in the visual cortex of mammalian brains can be modeled by Gabor functions (G. Daugman, 1985). A 2D Gabor filter can be viewed as a sinusoidal wave plane of particular frequency and orientation, which is the product of the Fourier Transform and a Gaussian centered at the origin, as given in equation 3.13. The impulse response of a Gabor filter is obtained by convolving with target image, and the convolution function is given by equation 3.14. A strong response can be given using Gabor filters only if its direction matches with the direction of the edges in the target image.

$$g_{\theta,\lambda,\sigma,\gamma,\psi}(x,y) = exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) sin(2\pi \frac{x'}{\lambda} + \psi) \tag{3.13}$$

where

$$x' = (x - x_o)cos\theta + (y - y_o)sin\theta$$
$$y' = -(x - x_o)sin\theta + (y - yo)cos\theta$$

In this equation, $g_{\theta,\lambda,\sigma,\gamma,\psi}(x,y)$ represents the filter value in location (x,y) under the parameter setting $\theta, \lambda, \sigma, \gamma, \psi$. θ refers to normal orientation of the parallel lines of Gabor function, its value is

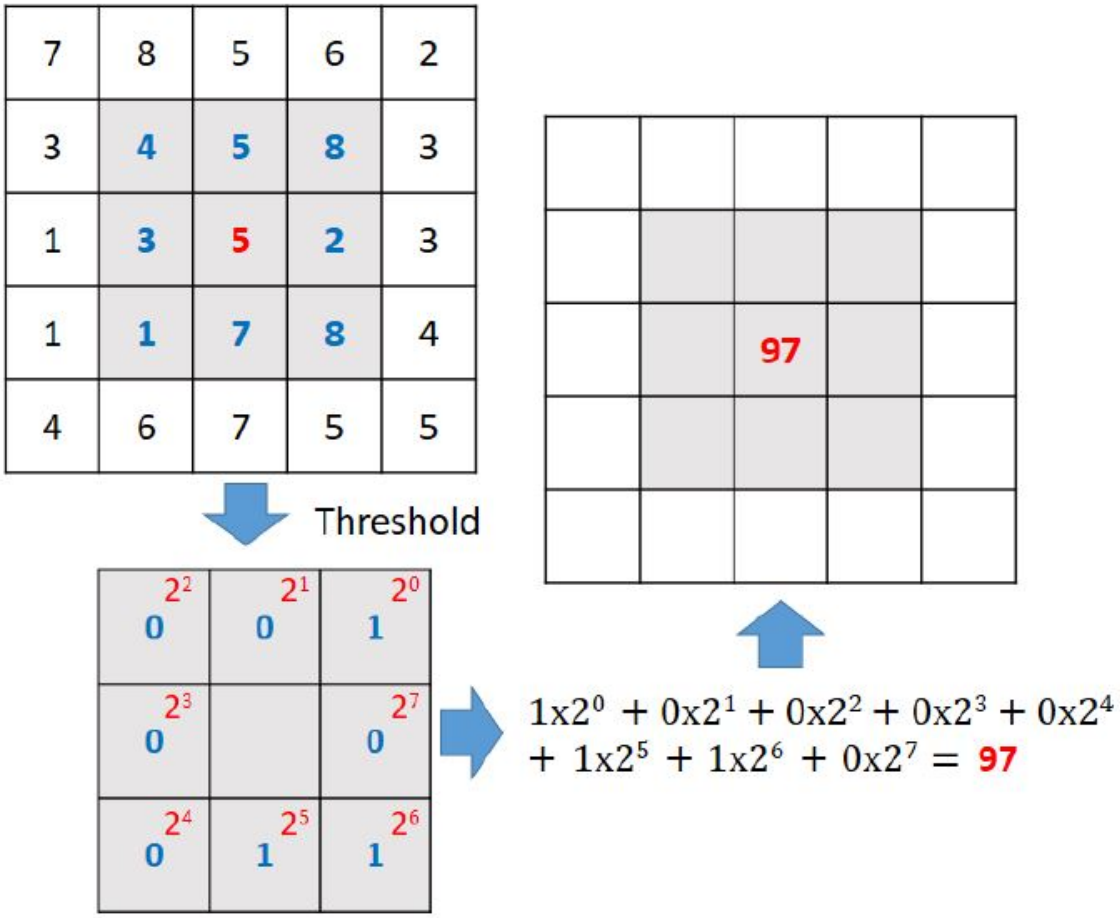


Figure 3.16: Computation of LBP value for a pixel (K Pai et al., 2017)

specified in degrees betwwen 0 and 360, λ refers to the wavelength of the sinusoidal factor, σ is standard deviation of the Gaussian function and it characterizes the spatial extent and bandwidth of the filter, γ is the spatial aspect ratio that determines the shape of the ellipse of the Gabor function and ψ is the phase offset and its value is in degrees between -180 and 180. (x_o, y_o) is the center pixel in the 2D Gabor filter kernel matrix.

$$G(x,y) = g * I(x,y) = \sum_s \sum_t I(x',y')g(s,t) \tag{3.14}$$

where

$$x' = x - (s_o - s)$$
$$y' = y - (t_o - t)$$

In this equation $G(x,y)$ is the Gabor feature value in location (x,y), g is the Gabor filter kernel computed from above equation, $I(x,y)$ is the original image that you want to extract features from. (s_o, t_o) is the center pixel in the 2D Gabor filter kernel matrix.

In particular, NDVI image was used as target image for Gabor feature extraction. We had four orientations, at 45-degree intervals, two wavelengths and two standard deviations in order to cover sufficient texture features, phase offset were set to 0 and spatial aspect is set to 1. Filters at different parameter combinations are shown in figure 3.17. Therefore 16-dimensional Gabor feature vectors were resulted from these combinations. As the Gabor filters usually cover excessively possible orientations and produce large feature size, this technique also requires dimensionality reduction.

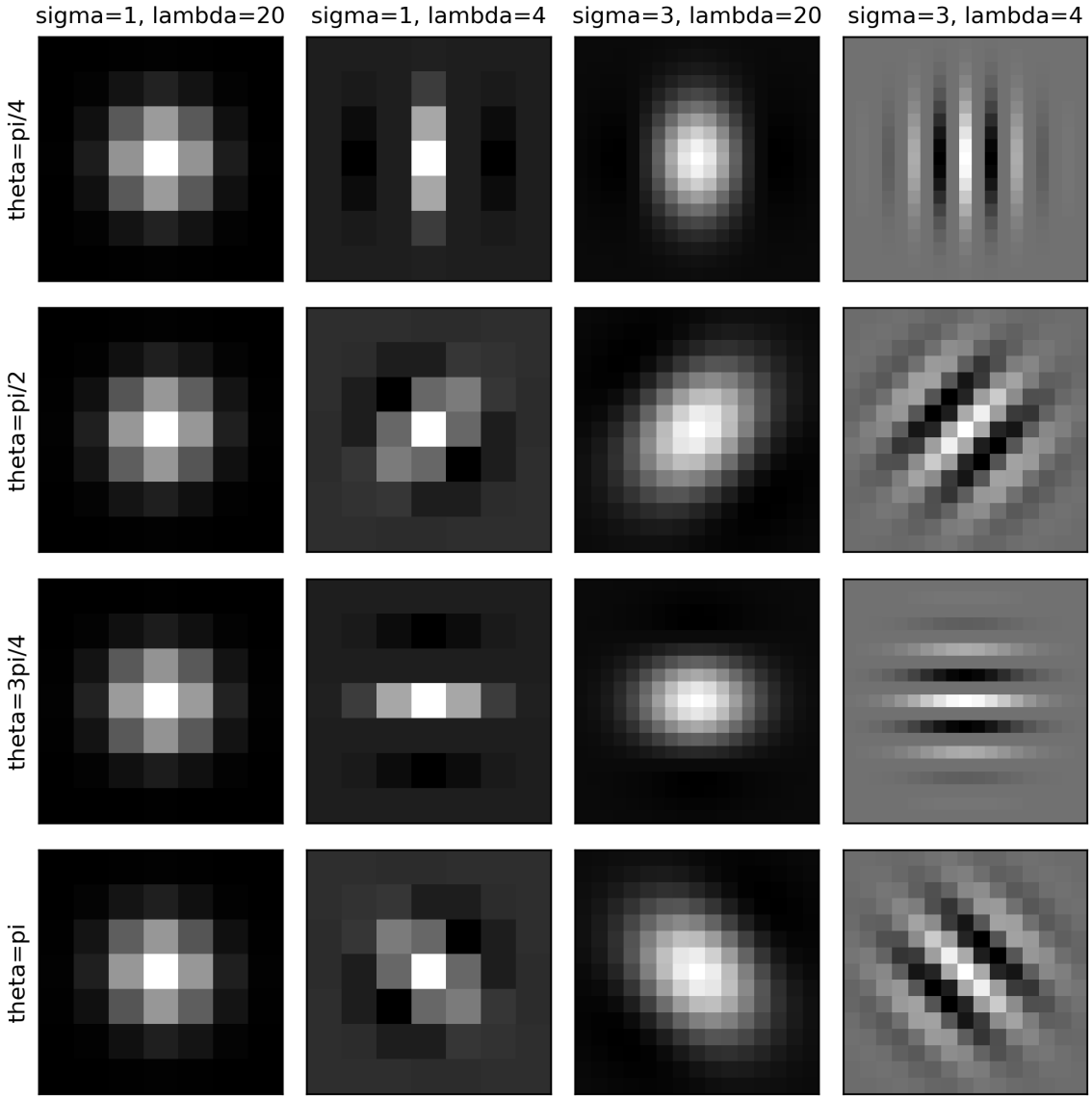


Figure 3.17: Gabor filter kernels: there are 16 combinations in total of four orientations including π,$\pi/2$,$3\pi/4$ and π, two wavelengths including 20 and 4, and two standard deviations including 1 and 3.

3.4.6. Principal component analysis

As we mentioned in feature selections section, there is often significant redundancy between spectral bands, and it is desirable to reduce the number of bands being analysed. Each 256-demensional LBP feature vector is extracted from a 7×7 sliding windowa, which will result in a highly sparse feature matrix, and the Gabor filters usually have excessively kernels in order to cover sufficient possible texture features. Thus, LBP and Gabor filters also require dimensionality reduction.

Principal Component Analysis (PCA) is a well-known technique used for dimensionality reduction. The goal of PCA is to find a new coordinate system in which all the dimensions are orthogonal and data are spread out across each dimension and hence independent. Variance is a measure of how spread the data set is, as given in equation 3.15 and covariance is measured between two dimensions, as given in equation 3.16. The goal of PCA is accomplished by following steps. First it calculates the covariance matrix (equation 3.17) of input features and then calculates eigen vectors and corresponding eigen values (equation 3.18). The first k eigen vectors are sorted according to their eigen values or variance in decreasing order. The new coordinate system is obtained by transforming the original n dimensional feature space into k dimensions which are called principal components (PCs). So that the first component has the highest variance, the second component the second-most variance, and so on. Therefore, the first few PCs containing most of the useful information, and the later PCs containing mostly noise which can be removed without significantly affecting the information content of the image.

$$var(x) = \frac{\sum(x_i - \bar{x})}{N} \tag{3.15}$$

$$cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N} \tag{3.16}$$

$$C_{i,j} = (cov(d_i, d_j)) \tag{3.17}$$

$$[Covariancematrix] \cdot [Eigenvector] = [eigenvalue] \cdot [Eigenvector] \tag{3.18}$$

where N is the number of samples in each dimension d and x_i and y_i refer to the i_{th} element in x and y dimensions respectively. The value in location (i, j) of covariance matrix is calculated by the covariance of the i_{th} dimension and the j_{th} dimension.

However, PCA is sensitive to the scale of the original variables. A disadvantages of the PCA mechanism is that original variables with larger ranges will dominate over those with small ranges, which will then lead to biased results. So it is critical to perform standardization prior to PCA, as given in equation 3.19.

$$z = \frac{x - \mu}{s} \tag{3.19}$$

where μ and s are the mean and standard deviation of the samples respectively. z is the standard score.

In this work, we first standardized feature vectors including reflectance, LBP and Gabor features of training samples and then applied PCA to reduce the dimensionality. To make it fair, spectral features and texture features had the same dimensions in total. The number of PCs that were selected for each type of features is shown in table 3.7. Reflectance PCs, vegetation indices, LBP PCs and Gabor PCs were then concatenated together to produce the final feature vector with dimension 28. The decision to set the length of final vector as 28 is based on variance while it should also to make sure that the number of spectral features and texture features are the same, in order for further comparison of their importance. The scaler model and PCA model were then applied for testing samples.

	Spectral features		Texture features		Total
	Reflectance	VIs	LBP	Gabor filters	
Before	13	9	256	16	294
After	5	9	7	7	28

Table 3.7: Number of PCs for each type of features. VIs represents vegetation indices and its number stays unchanged because PCA was not applied to it.

3.4.7. Aquatic plant species classification

Training data

The segmentation method is described in detail in section 3.4.4. The segmentation results were used to assume the homogeneity of individual segments and label pixels for training and validation. Segments that represent the same species in different regions may have different features. For example, characeae grows in both deep water and shallow water. Some water bodies are covered by algae while others are crystal clear. Therefore, the training data should cover all segments because of the various local environments. In this project, pixels in the segments that were labeled by field samples were used for training, in which process 20% pixels were randomly taken out for testing. Pixels in the segments that were not labelled by field samples were only used for predicting.

Random Forest Classifier

Random Forest (RF) are a combination of tree classifiers in which each tree contributes one vote, and the final classification label is obtained by maximum voting. This ensemble classification is often more accurate than any one from the ensemble (Breiman, 2001). RF is relatively robust to training dataset noise and does not overfit due to the Law of Large Numbers (Zhou et al., 2018). As a result, the RF classifier is widely used in processing remote sensing data. A random forest is comprised of a set of decision trees, each of which is trained on a randomly selected subsets of components of the feature vector by using a random vector θ_k for the k-th tree and resulting in a classifier $h(x,\theta_k)$, where x is an input vector. The random vector is independent of the prior

random vectors θ_1, ..., θ_{k-1}. Each tree grows by using the best split of a random subset of input features or predictive variables in the division of every node (Ho, 1998; Rodriguez-Galiano et al., 2012). Gini impurity is calculated to decide the variable to split a node. Meanwhile, for each tree of the forest, the sum of the Gini decrease is accumulated every time that variable is chosen to split a node. Therefore, an average , obtained by dividing the sum by the number of trees, is then equivalent to the variable importance. In our study, the importance of different types of features was analyzed based on this logic. Although weakening the strength of very single tree, it reduces the correlation between the trees and the generalization error (Rodriguez-Galiano et al., 2012). The predicted category of an observation is calculated based on the majority vote of the trees in the RF model, as defined in equation 3.20.

$$H(x) = argmax_{\gamma} \sum_{i=0}^{k} I(h_i(X,\theta_i) = \gamma) \tag{3.20}$$

where $I(\cdot)$ is the indicator function, $h(\cdot)$ is a single decision tree, and γ is the output variable, $argmax_{\gamma}$ denotes the γ value when maximizing $\sum_{i=0}^{k} I(h_i(X,\theta_i) = \gamma)$.

K-means clustering

Labels are an essential ingredient to supervised algorithms. However, conducting a field survey not only costs time but money. This experiment aims to answer the research question that whether field data is necessary in remote sensing techniques. Unsupervised algorithms only utilized input vectors without referring to known or labelled outcomes. Therefore, in this project, we also want to investigate the performance of an unsupervised algorithm. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. A pre-defined k refers to the number of centroids in the dataset. A point $x^{(i)}$ from set $x^{(1)},...,x^{(m)}$is allocated to a particular cluster and labelled by $c^{(i)}$ if it is closer to that cluster's centroid than any other centroid as euqation 3.21 and the centroid is updated every time a new point is added. The K-means algorithm starts by initializing cluster centroids $\mu_i, \mu_2, ..., \mu_k$ randomly, then performs iterative calculations to optimize the positions of the centroid as equation 3.22. Repeat the process until centroids have stabilized or it reaches the maximum number of iterations. For every point i, set

$$c^{(i)} := argmin_j ||x^{(i)} - \mu_j||^2 \tag{3.21}$$

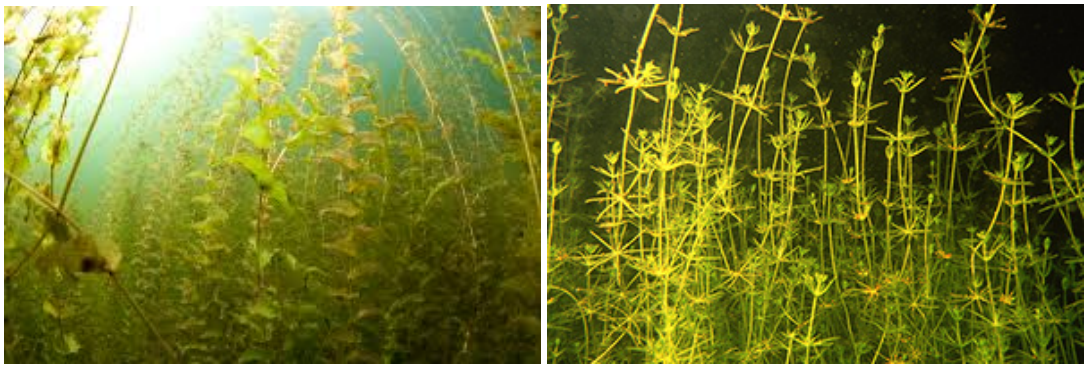
For each centroid j, set

$$\mu^{(j)} := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} \tag{3.22}$$

Two classification tasks

The first and main task is vegetation class classification. Vegetation classes included water, low density plants, floating plants, Characeae, submerged plants, and Pondweed. In this task , we conducted RF classifier and K-means clustering respectively.

The second task is individual species classification, which is to classify DF and other species within Pondweed family and classify KB and other species within Characeae family. As there are only two dominant species in Characeae, which are KB and Sterkranswier. It can be also regarded as classifying between KB and Sterkranswier. DF and KB are two plants that grow the whole water column. Their distributions are also expanded in recent years. This grow pattern pose a serious problem for boaters and others using the water for recreation. Boat propellers can very easily get stuck in the plant's stalks. Therefore, it is necessary to detect these two species particularly. In this task, pixels that were labelled by Characeae and Pondweed were taken out respectively. The same feature selection method as the first task. RF classifier was then applied to conduct the classification.



(a) Doorgroeid fonteinkruid (b) Kransblad

Figure 3.18: Two problem plant species

3.4.8. Assessment and feature importance

Cross validation

Cross-validation is a re-sampling procedure used in this project to evaluate RF classifier on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. This parameter was set to 5 in this case and this procedure can also be called 5-fold cross validation. 5-fold cross validation starts by shuffling the dataset randomly, then splitting the dataset into 5 groups. A stratified criteria was adopted in this work that each fold has the same proportion of observations with a given categorical value. For each group, it is considered as a test set and remaining groups are considered as a training set to fit the RF model. This model will then be evaluated on the test set. The accuracy will be retained and the model will be discarded. This procedure generally results in a less biased or less optimistic estimate. The overall accuracy is the average of the five evaluations.

Confusion matrix

Classification accuracy alone can be misleading especially when the dataset is imbalanced and have more than two classes in the dataset. Confusion Matrix is a performance measurement for machine learning classification. A confusion matrix is a specific layout that allows the visulization of the performance of an algorithm. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class. This can then give a better idea of which class can be easily classified and which class is mixed with others. Two confusion matrix will be calculated: one is calculated by the number of pixels classified into each class, by which we will know the exact number of correctly classified pixels; and another is normalized confusion matrix, by which we will know the percentage of accurately corrected pixels in each class.

In the classification process, four types of features were extracted including spectral reflectance, VIs, LBP and Gabor Filters. Feature importance can give insight into the performance of these features.

Testing on another date

In practical applications, pre-trained model is usually applied to new data acquired under different conditions, such as different solar irradiance, different cloud cover, species changes and so on. It is necessary to discuss the applicability of this approach. Thus, Sentinel-2 data from Aug 6 were used for evaluation of pre-trained model. We chose this day because it is close to July 2 and of low cloud cover. We assumed that in the period of July to Aug, the state of aquatic plants remain unchanged. The same region was clipped in QGIS and the same feature selection strategy was adopted. Then pre-trained RF model was applied.

Baseline

As aquatic vegetation classification studies in the literature were conducted under different conditions, it is impossible to find a study to compare with. Therefore, majority classifier is used in this research as the baseline. The majority classifier simply classifies all samples into the class with maximum occurrence. This classification results serve as the baseline for other classification approach. If proposed approach performs worse than this baseline, it can be concluded that the proposed approach does not work in this classification task.

4

Results

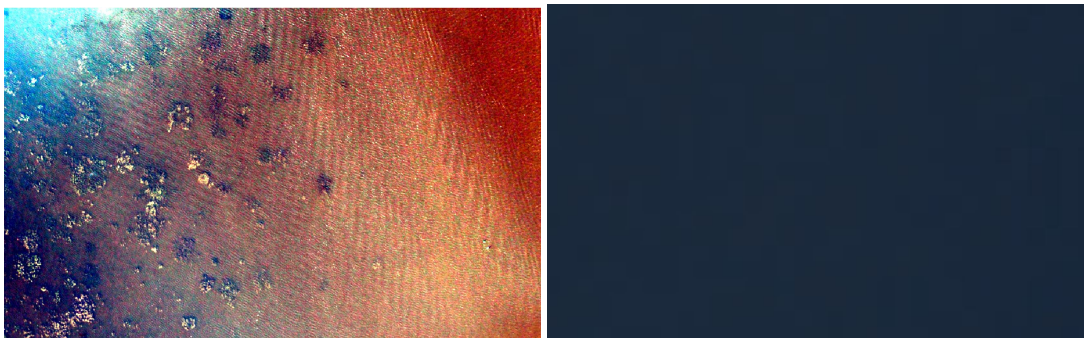
4.1. Remote sensing data comparison results

Figure 4.1 shows images of four classes clipped in airborne multispectral image and Sentinel-2 image. The water and submerged Sentinel-2 images are almost black for human visual system. This is because that the Pondweed and soil are all in the water bottom. Reflectance from water bottom is hardly to be captured by Sentine-2 imagery because of water depth and orbit altitude, while for airborne imagery more details can be captured because of much lower flight height. For floating plants and Characeae classes, Sentinel-2 images have a greenish layer above. This layer can be seen because vegetation grew close to the water surface, which makes them distinct from Pondweed and water classes. This comparison result indicated that airborne multispectral images might be more easier for human visual system to directly discriminate different classes than Sentinel-2 images.

The reflectance of water, island and aquatic vegetation classes in Sentinel-2 images matched that in hyperspectral images (figure 4.2a). In general, land vegetation had higher reflectance in NIR region, while water body and aquatic vegetation have lower reflectance in NIR region. Aquatic vegetation have higher reflectance in visible region while water body had lower. These patterns can be found from both hyperspectral and Sentinel-2 spectrum. From figure 4.2b, Sentinel-2 images have similar NDVI distribution as that in hyperspectral images, though there exists a tiny shift. In general, the NDVI of land vegetation is greater than zero, while NDVI of water body is smaller than zero. Two main parts, which refer to land vegetation and water body respectively, can be easily seen in both NDVI distributions. Also, there are two peaks in the above-zero part in both hyperspectral and Sentinel-2 data NDVI distribution, which is because that island and land vegetation may have NDVI distribution. From the spectral comparison, Sentinel-2 data can relatively show accurate spectral properties of different classes.

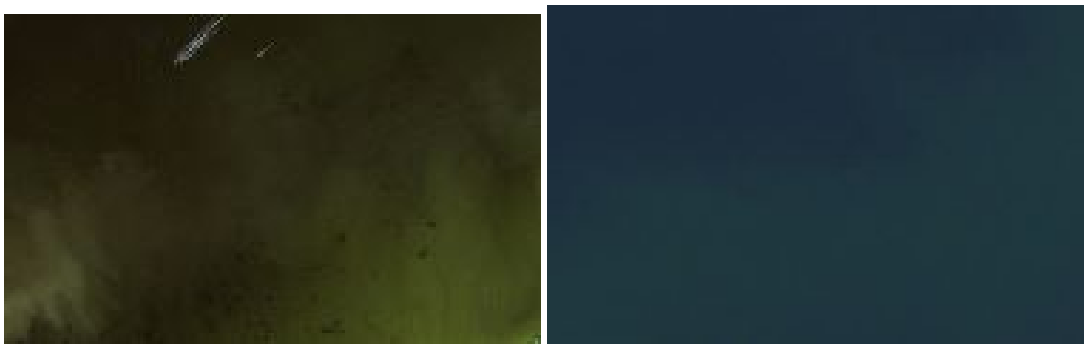
(a) Pondweed in airborne image

(b) Pondweed in Sentinel-2 image



(c) Water in airborne image

(d) Water in Sentinel-2 image



(e) Floating plants in airborne image

(f) Floating plants in Sentinel-2 image

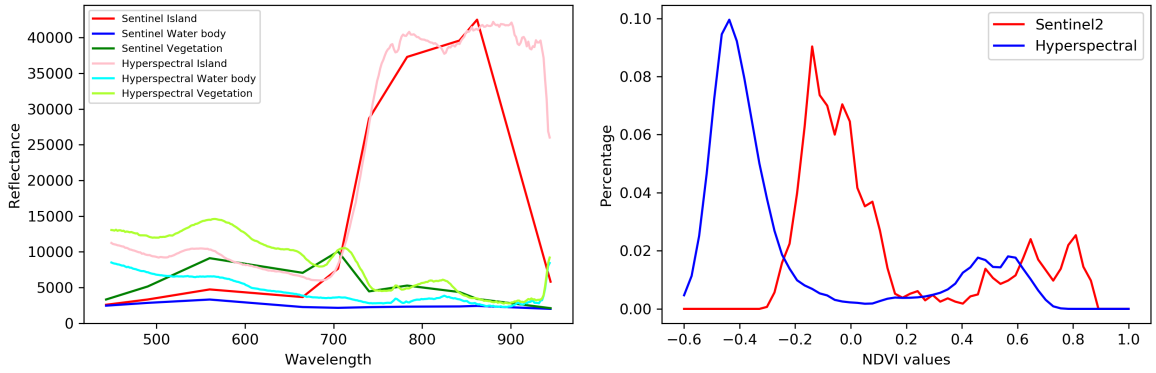
(g) Characeae in airborne image

(h) Characeae in Sentinel-2 image

Figure 4.1: Spatial comparison results

From all the remote sensing comparison results, Sentinel-2 data can be an alternative to airborne

hyperspectral data, and narrow bands collection is therefore not necessary. In spatial aspect, however, Sentinel-2 data are insufficient to directly discriminate classes by human visual system and cannot replace airborne multispectral data. Whether this spatial drawback can impact on classification performance will be revealed in following sections.



(a) Reflectance of three classes

(b) NDVI distribution

Figure 4.2: Spectral comparison results

4.2. Segmentation results

In this study, three minimum number of cells in a segment or an object settings including 1, 10 and 30 were compared, in the aspects of total number of labelled objects, average purity, and comparison with field observations. From table 4.1 and table 4.2, segmentation result with parameter 10 had most number of labelled objects and the highest purity or highest homogeneity in an object among these three parameter settings. This also indicated that the segmentation result matched our assumption that all pixels in individual objects were from the same categories.

Minimum number of cells in a segment	1	10	30
Water	38	137	144
Low density plants	37	125	117
Floating plants	11	35	37
Characeae	23	99	88
Submerged plants	8	37	29
Pondweed	25	90	92
Total number of segments	142	523	507

Table 4.1: Segmentation comparison in aspect of number of labeled objects

There were total 2297 objects, out of which there were 523 objects that can be labelled by field samples. Blank areas were unlabelled objects and we had no knowledge which category they were from.

Minimum number of cells in a segment	1	10	30
Water	0.84	0.90	0.90
Low density plants	0.78	0.89	0.86
Floating plants	0.62	0.79	0.77
Characeae	0.86	0.92	0.91
Submerged plants	0.88	0.84	0.80
Pondweed	0.86	0.88	0.90
Average purity	0.81	0.87	0.85

Table 4.2: Segmentation comparison in aspect of average purity

Pixels from these areas were not used for training and assessment. We also selected two sites from the segmented map to compare with field observations. In site 1 (figure 4.4), we compared the segmentation results with interpolated ground truth image. The interpolated ground truth image was total vegetation cover map obtained by linearly interpolating data among field samples. Blue area indicated water and red area indicated high density vegetation. Segmentation with parameter 30 failed to show the water area in this site. Therefore, segmentation with parameter 10 outperformed that with parameter 30. In site 2 (figure 4.5), we compared the segmentation results with field observations in that region. This region was mixed with all categories, however, the segmentation with parameter 1 showed large cover of water, which was inconsistent with field observations.

After analysis of the appropriate parameter, labelled pixels in the segmented image were used to extract feature vectors for classification. The number of pixels in each category was listed in table 4.3. The number of pixels that were used in problem plant species classification task was showed in table 4.4.

Water	Low density plants	Floating plants	Characeae	Submerged plants	Pondweed
564230	137953	87209	307633	42288	134159

Table 4.3: Number of pixels in each category

Pondweed		Characeae	
DF	Others	KB	Others
75353	58806	181983	125650

Table 4.4: The number of pixels in Pondweed and Characeae families

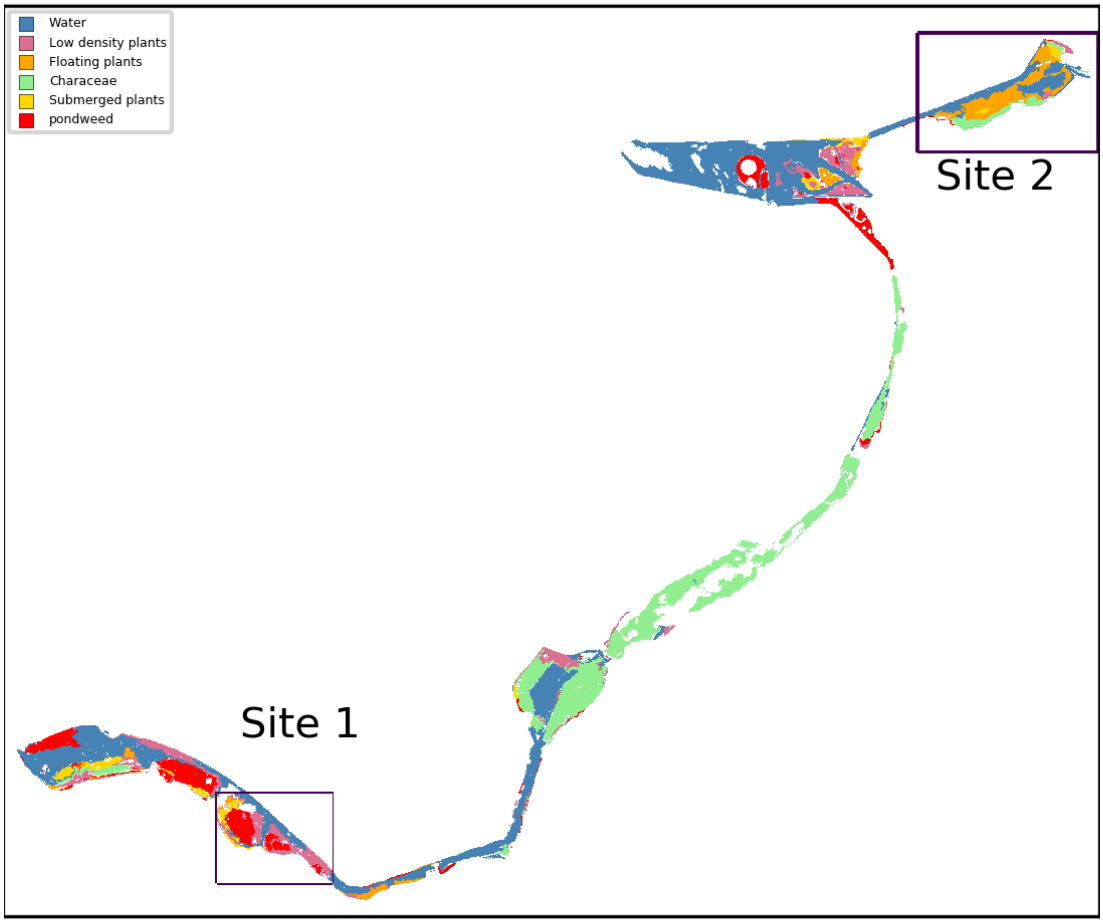


Figure 4.3: Segmentation results with minimum 10 cells in a segment

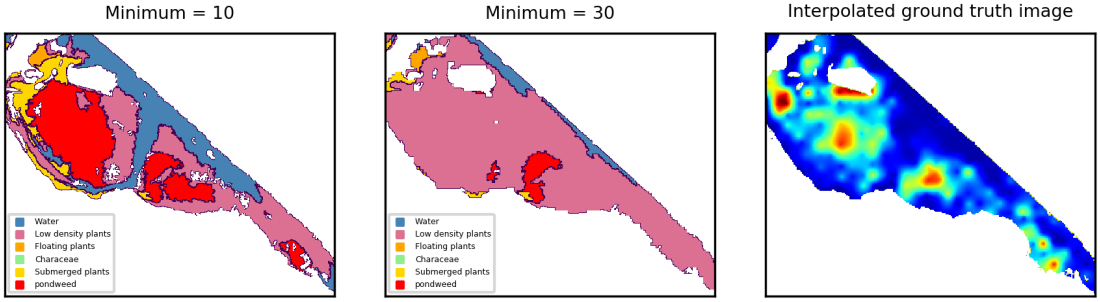


Figure 4.4: Comparison in aspect of total vegetation cover

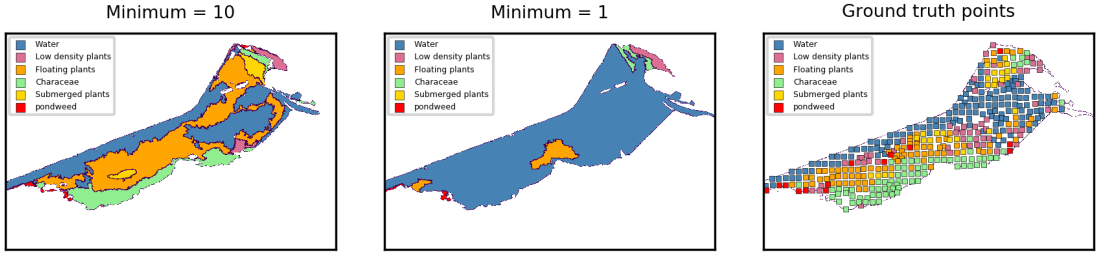


Figure 4.5: Comparison in aspect of vegetation distribution

4.3. Vegetation class classification results

In this section, we showed the results of supervised algorithm (RF classifier) and unsupervised classification (K-means clustering) respectively. In this classification task, there are six classes, including water, low density plants, floating plants, characeae, submerged plants and pondweed. The Seninel-2 data used in this section was obtained from July 2 2018.

4.3.1. Supervised classification using RF classifier

Table 4.5 shows the accuracy in 5-fold stratified cross validation, and the average accuracy is 87%. The confusion matrix (figure 4.6) was generated from all five test sets in the cross validation procedure. The left confusion matrix were obtained by counting the number of pixels that were classified into each class. The number of correctly predicted pixels of water class and characeae class are obviously higher than other classes. This is because that the total number of pixels in water class and characeae class are much higher than other classes. Thus, a normalized confusion matrix was also calculated as the right figure. This normalized confusion matrix demonstrated that the model had evenly performance on each class. All these pixels were then projected to an aquatic vegetation map (figure 4.7).

1	2	3	4	5	Average accuracy
86.7%	86.5%	86.4%	86.5%	86.6%	87%

Table 4.5: Vegetation class classification accuracy

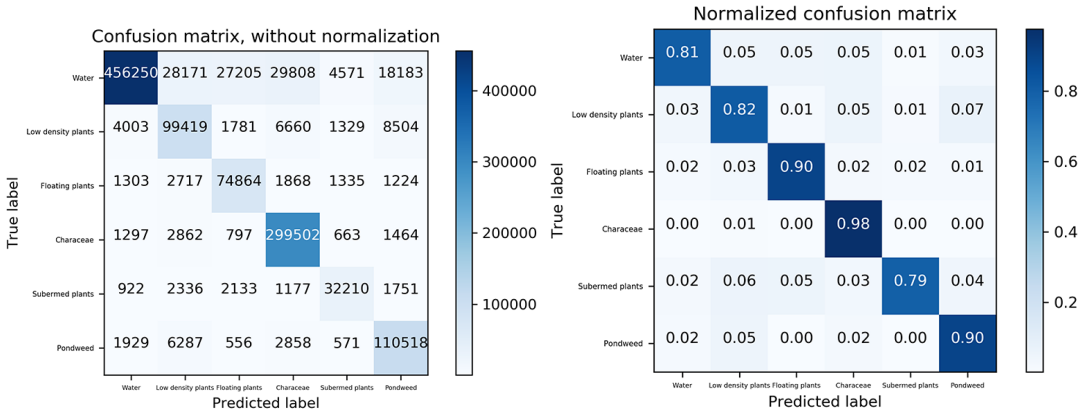


Figure 4.6: Confusion matrix of RF classifier. Each row represents the true label while each column represents the predicted label. The left figure shows the number of predicted pixels in each category, while the right figure is normalized confusion matrix and the values in the diagonal are classification accuracy of each category.

Figure 4.8 presents all features ordered by normalized feature importance. Out of total 28 features,

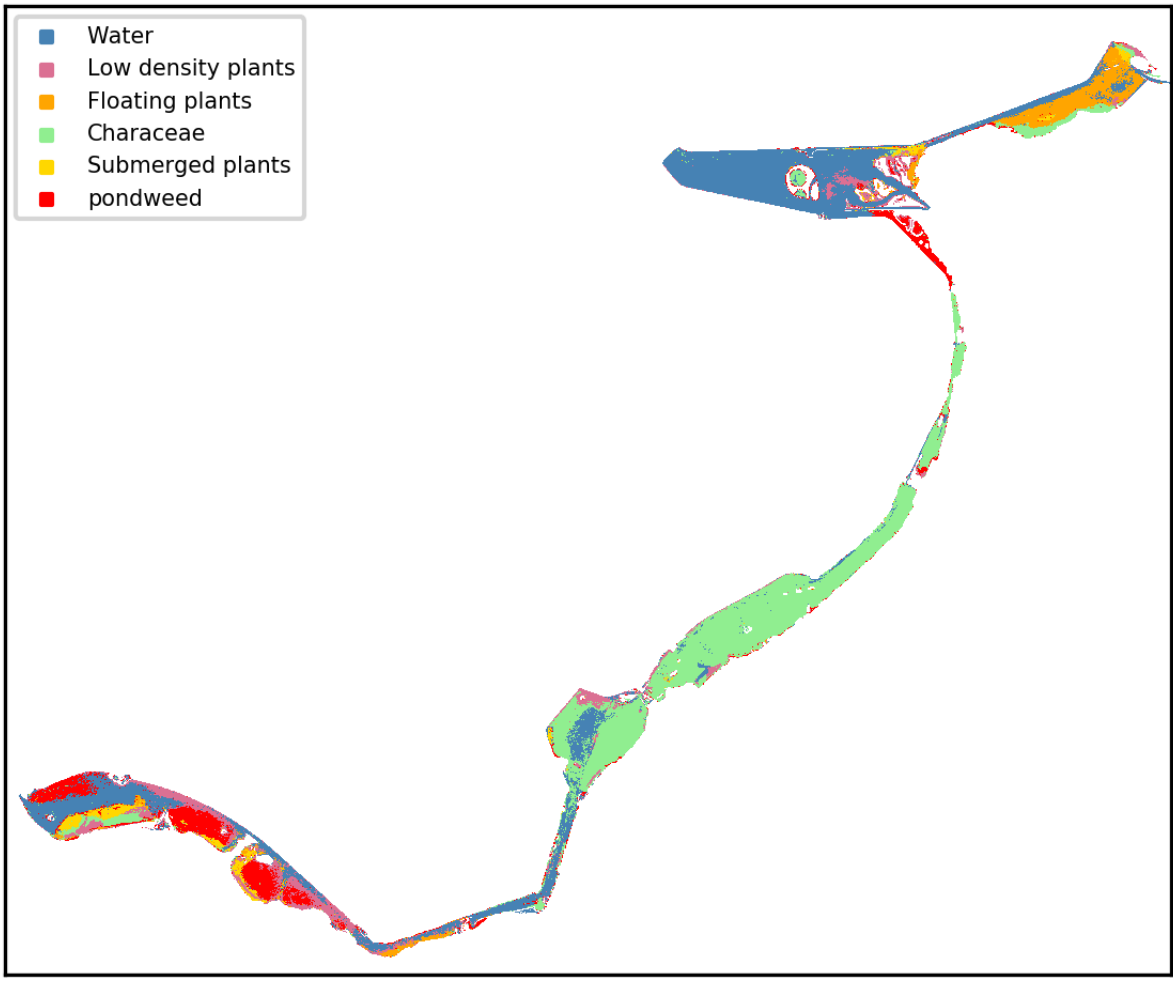


Figure 4.7: Predicted aquatic vegetation map using RF classifier (data from July)

top half features in this figure are spectral features, except one from Gabor filter component. The most important feature in this classification task is GNDVI (Green Normalized Difference Vegetation Index). This index is more sensitive to chlorophyll concentration than NDVI. Texture features, however, showed less significance in this task.

4.3.2. Unsupervised classification using k-means clustering

The overall accuracy is 46%. The overall accuracy was calculated using equation 4.1.

$$Overall\ accuracy = \frac{Total\ number\ of\ correctly\ classified\ pixels}{Total\ number\ of\ all\ pixels} \tag{4.1}$$

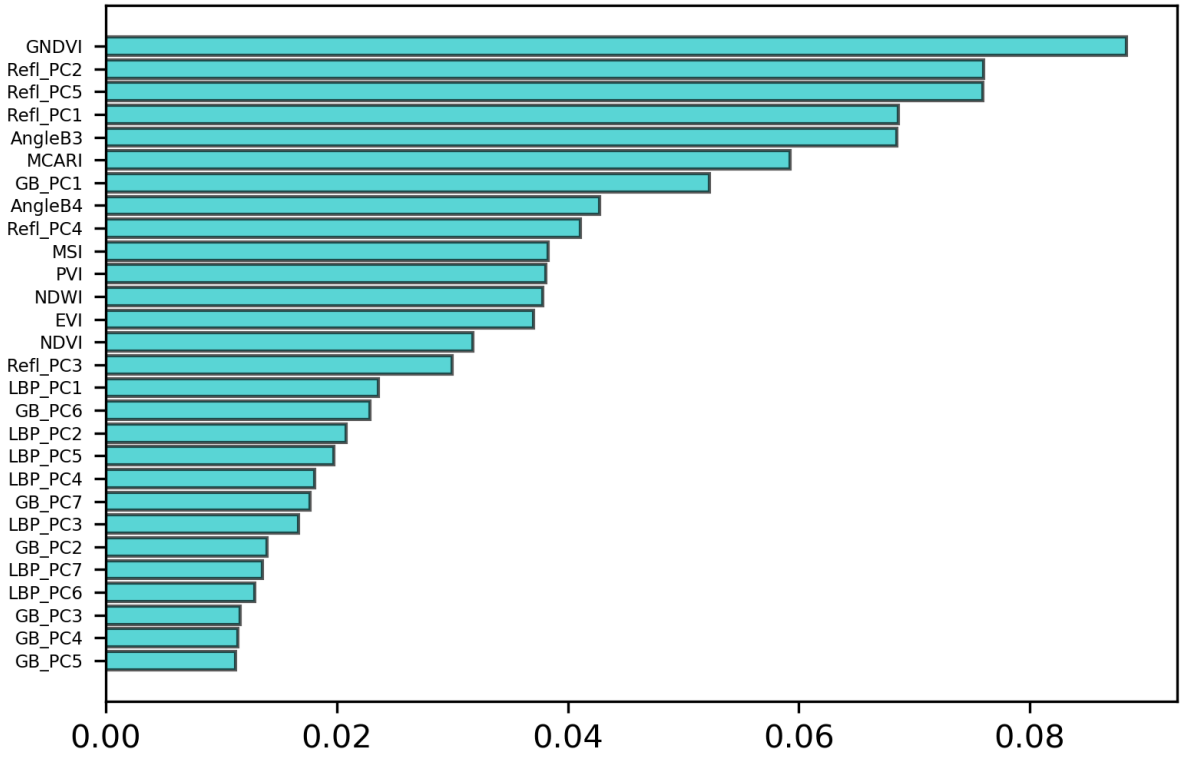


Figure 4.8: Normalized feature importance. Prefix 'Refl' refers to reflectance feature, prefix 'GB' refers to Gabor filters and suffix 'PC' refers to principal component.

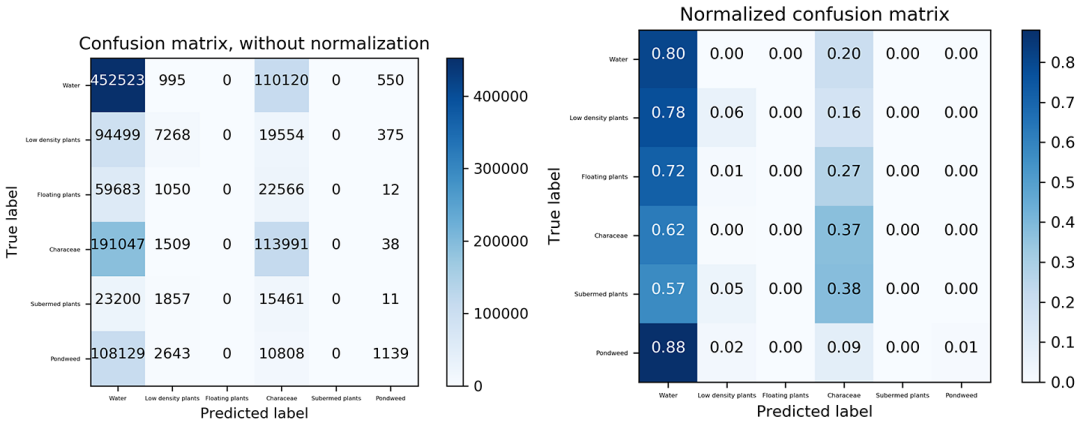


Figure 4.9: Confusion matrix of K-means clustering

K-means clustering classified pixels into only two classes, which were water and characeae, and failed to classify pixels from other four classes. As low accuracy of vegetation class classification, we want to see whether K-means clustering can at least classify between water and aquatic plants. Therefore, we grouped pixels from water and low density plants classes into water class and grouped others into aquatic plants class. Figure 4.11 shows the classification results. The overall accuracy is 55%. This result indicates that unsupervised algorithm is not suitable to classification in large water bodies. Field data is still necessary in remote sensing techniques.

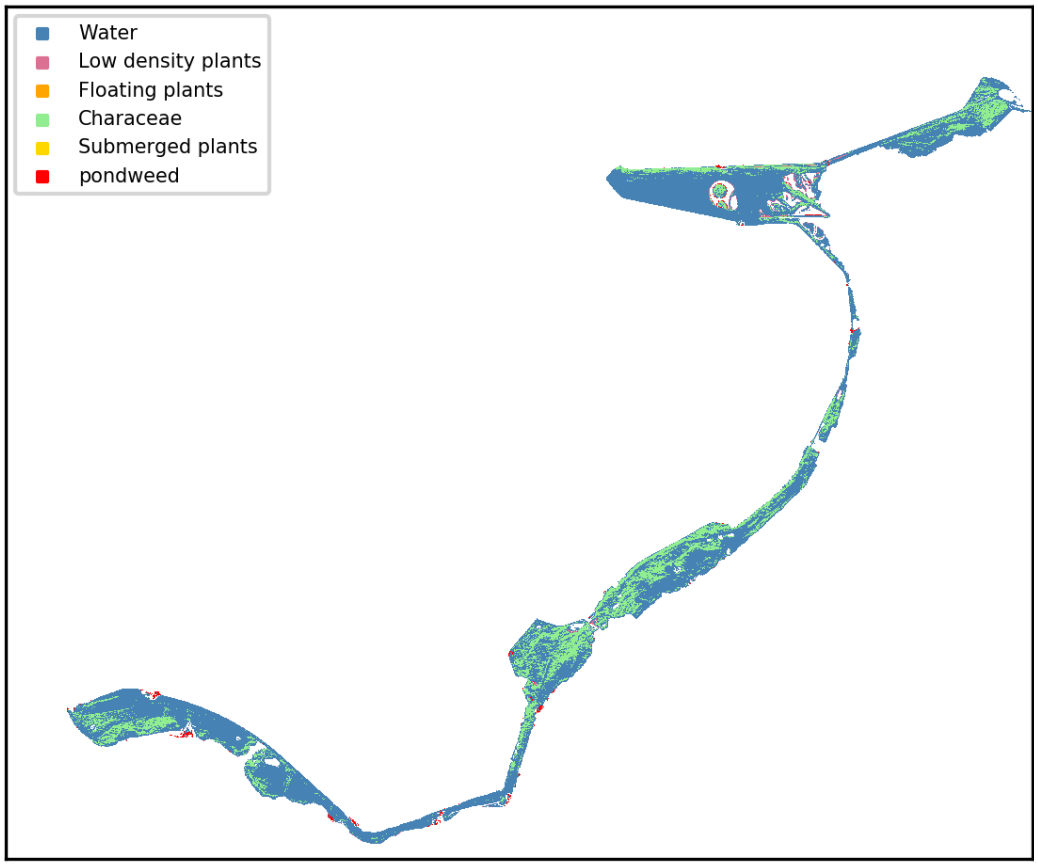


Figure 4.10: Predicted aquatic vegetation map using K-means clustering (data from July)

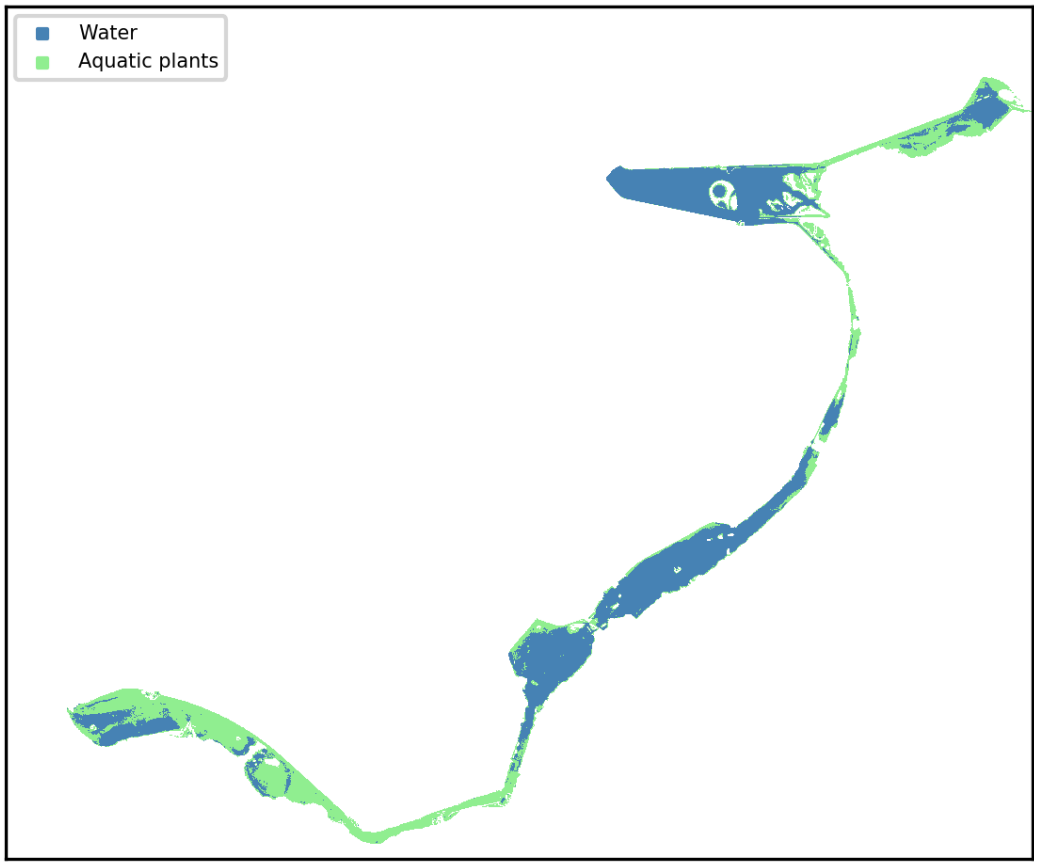


Figure 4.11: Water-vegetation classification using K-means clustering (data from July)

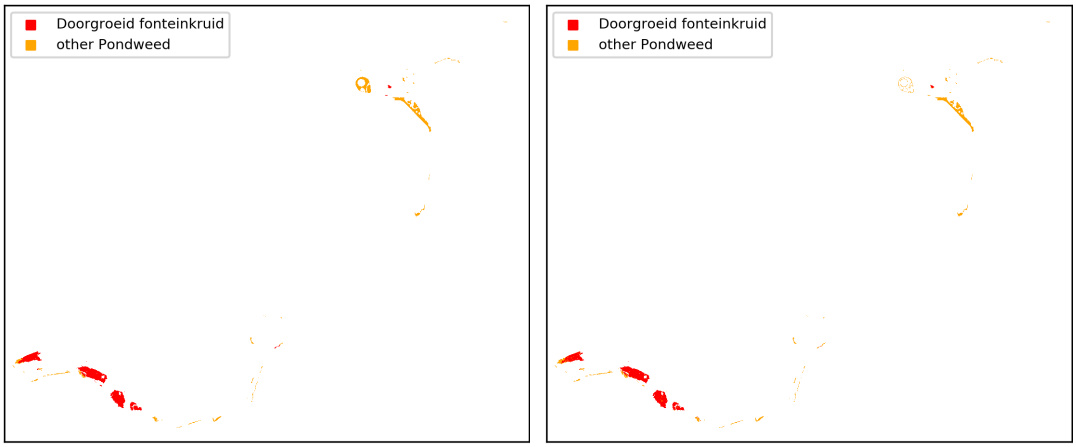
4.4. Individual species classification results

In the second classification task, two experiments were conducted. One was to detect DF from Pondweed family and another was to detect KB from Characeae family. The same approach as vegetation class classification was applied. Table 4.6 shows the accuracy in 5-fold stratified cross validation, and the average accuracy is 96% for DF classification and 95% for KB classification respectively.

	1	2	3	4	5	Average accuracy
DF	96.4%	96.3%	96.6%	96.8%	96.3%	96%
KB	95.5%	95.3%	95.4%	95.5%	95.3%	95%

Table 4.6: Individual species classification accuracy

We projected all pixels into the Pondweed map (figure 4.12). Figure 4.14a was obtained by relabelling objects using field samples, while figure 4.14b was our prediction.



(a) Target map (b) Predicted map

Figure 4.12: Doorgroeid fonteinkruid classification

Figure 4.13 presents all features ordered by normalized feature importance. Reflectance PC and the angle index of Band 3 are the top two most important features. Distributions of DF and other Pondweed on these two features were plotted in figure 4.14 using all labelled pixels. To neglect outliers, we only selected values in the range of 5% to 95% of the ordered feature values and then divided these values into 20 bins. We can see the first reflectance PC counts 28% importance. DF has much higher value than other Pondweed, and most DF distributed around -1.5 while other Pondweed distributed around -2.5 on this feature. The angle index of band 3 is the second important feature, and in this case, DF has relatively lower value than other Pondweed and they overlapped in the range of 0.7 to 0.9. Nevertheless, these two features show us very special spectral characteristics of DF comparing to other Pondweed.

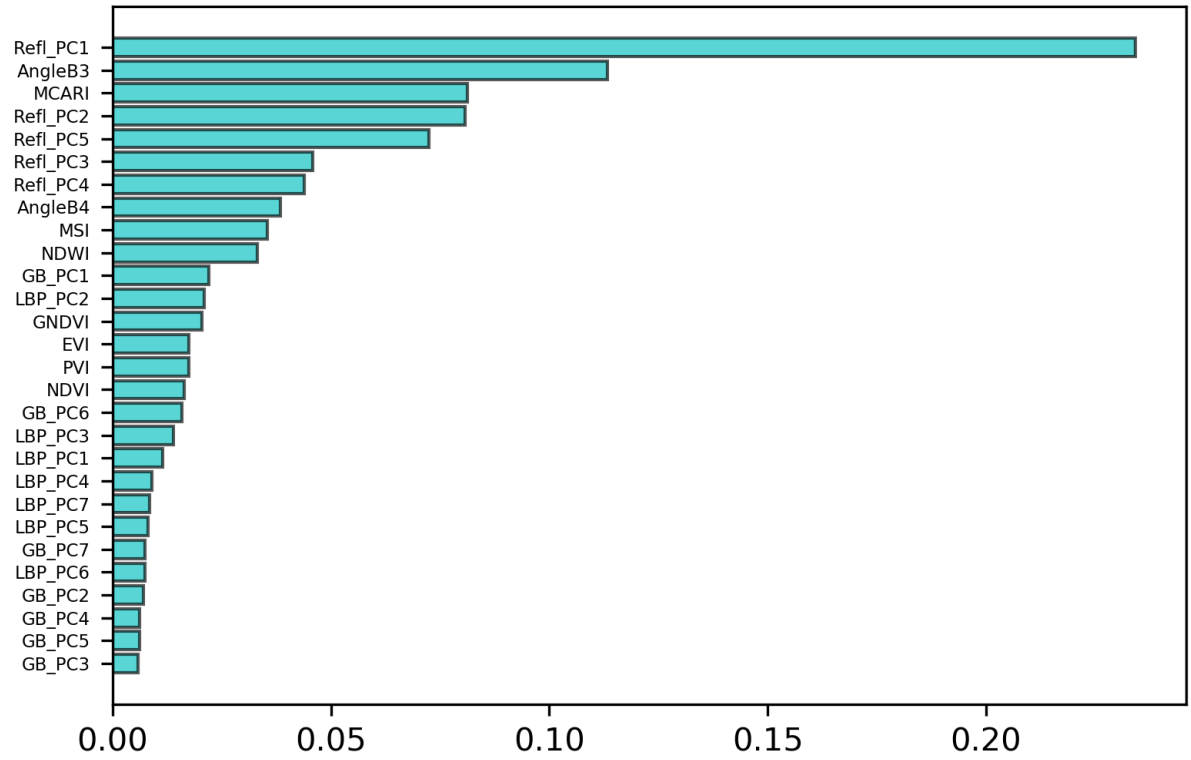
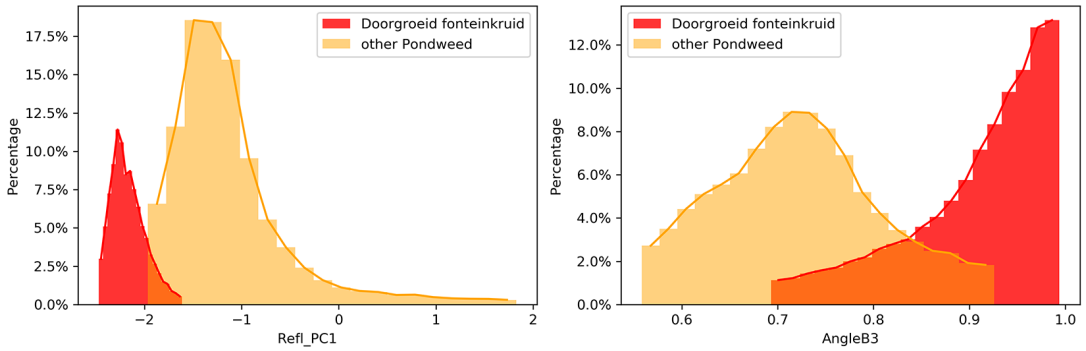


Figure 4.13: Normalized feature importance for DF classification



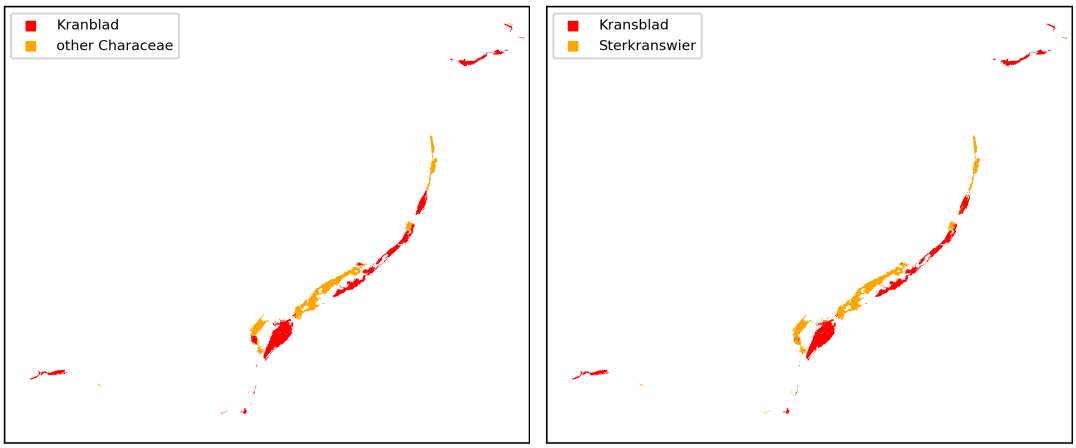
(a) Reflectance principal component (b) The angle index of band 3

Figure 4.14: Top two important features for DF classification

Characeae mainly grows in Veluwemeer. There are only two dominant plant species, Kransblad and Sterkranswier. Classifying Kransblad within Characeae in this case is equivalent to classify Kransbald and Sterkranswier. We projected all pixels into the Characeae map (figure 4.15). Figure 4.15a was obtained by re-labelling objects using field samples, while figure 4.15b was our prediction.

Figure 4.16 presents all features ordered by normalized feature importance. Also, reflectance PC and the angle index of Band 3 are the top two most important features. Distributions of KB and Sterkranswier on these two features were plotted in figure 4.17 using all labelled pixels. To neglect outliers,

we only selected values in the range of 5% to 95% of the ordered feature values and then divided these values into 20 bins. We can see in both feature distributions that KB covers a larger range and has two peaks. This is because that KB grows in more complex environment than Sterkranswier and these two peaks maybe from lower water-depth regions and deeper regions.



(a) Target map (b) Predicted map

Figure 4.15: Kransblad classification

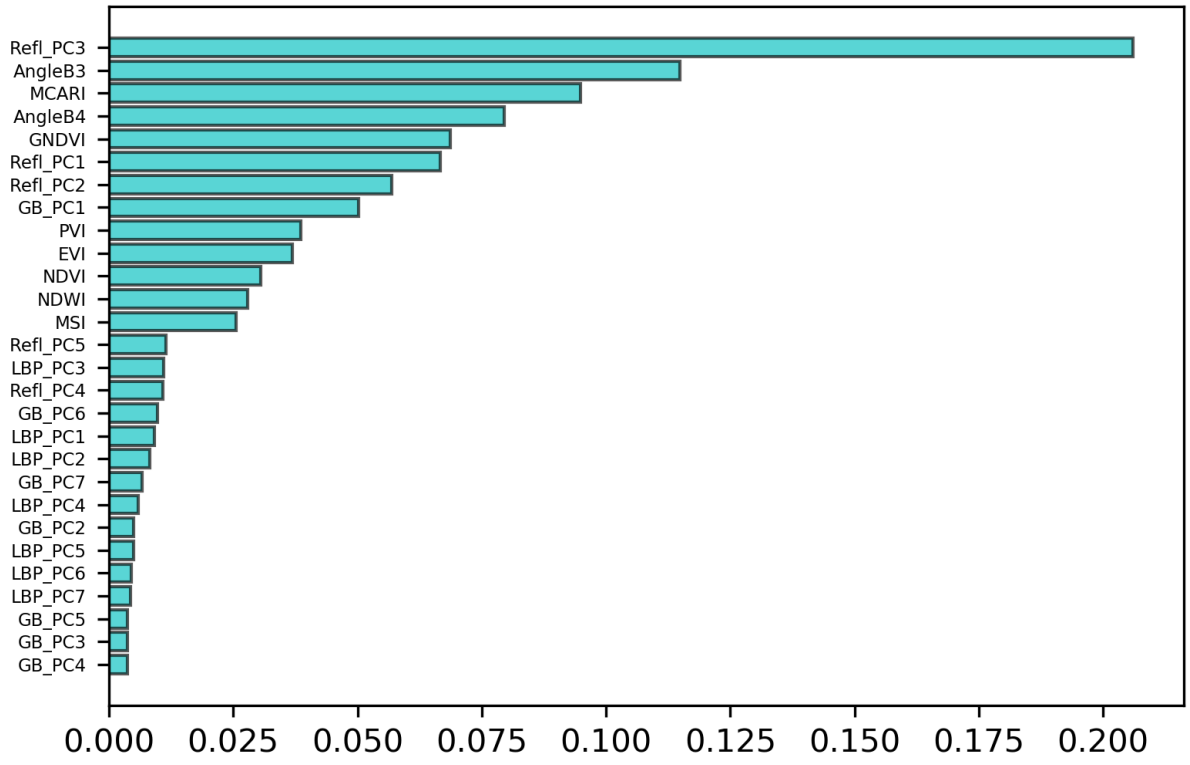
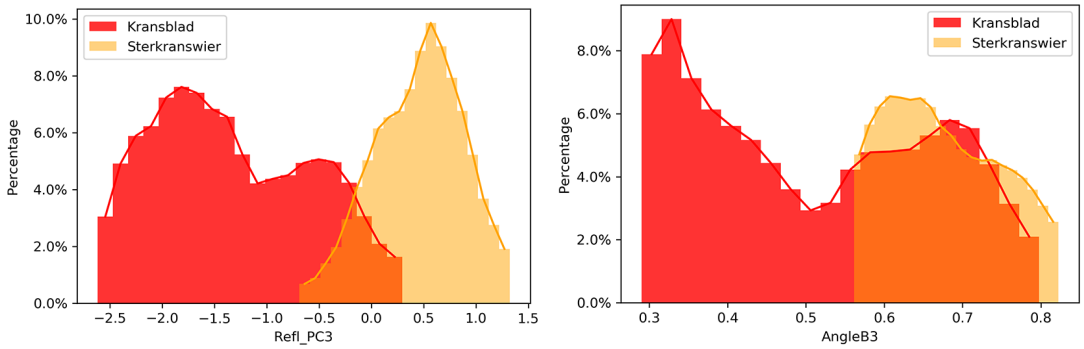


Figure 4.16: Normalized feature importance for KB classification



(a) Reflectance principal component (b) The angle index of band 3

Figure 4.17: Top two important features for KB classification

4.5. Test on another date

In order to test whether a pre-trained RF model can be applied to other dates, we selected Sentinel-2 data acquired from August 6, 2018. As we have no field data from August, the overall accuracy was calculated using field samples collected from June to July, based on the assumption that aquatic vegetation grows in similar state. The overall accuracy is 39%. From the confusion matrix (figure 4.19), the model had better performance on water, characeae and pondweed classes than other classes. However, this result is questionable which will be discussed in discussion chapter.

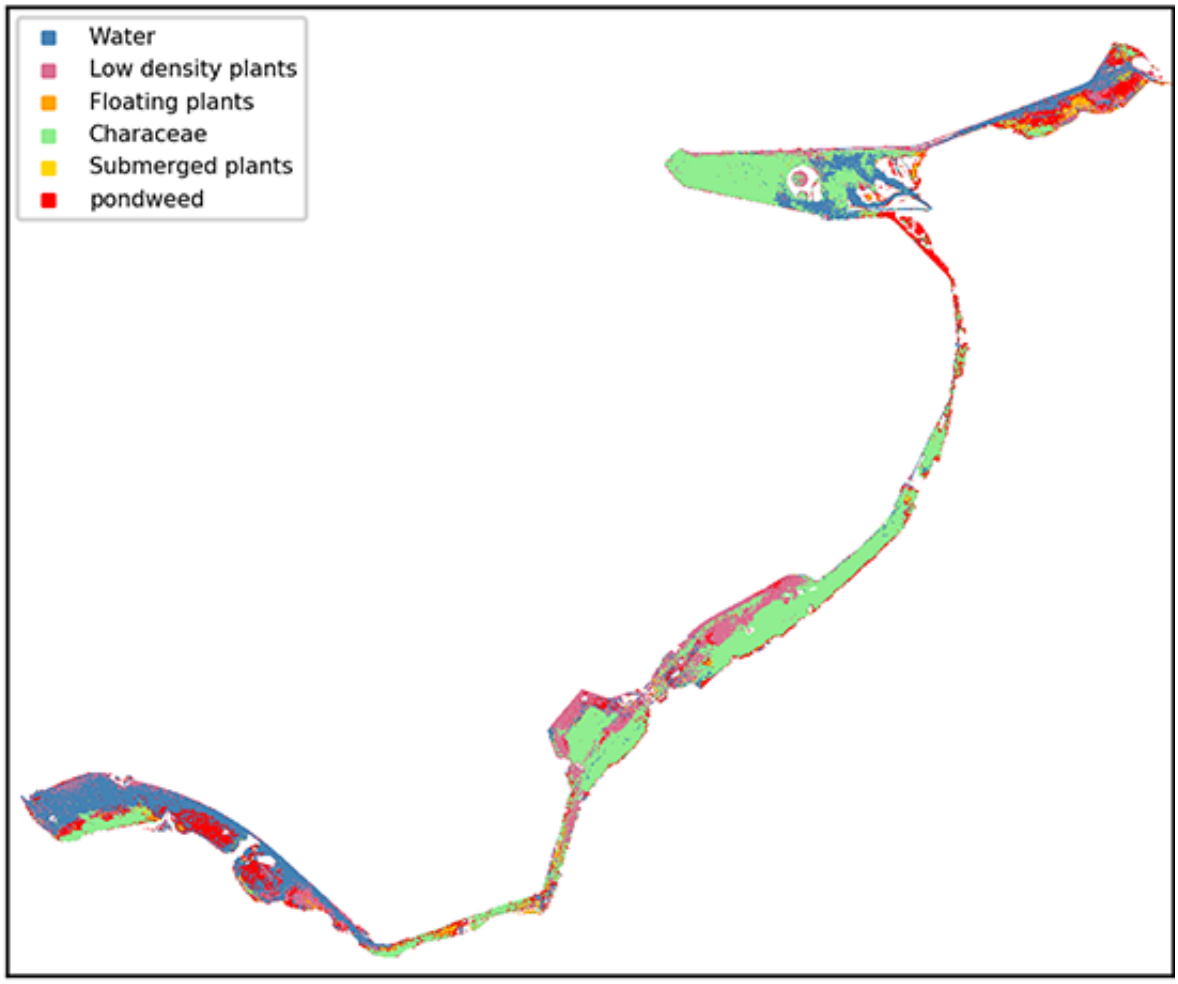


Figure 4.18: Predicted aquatic vegetation map using August 6 Sentinel-2 data

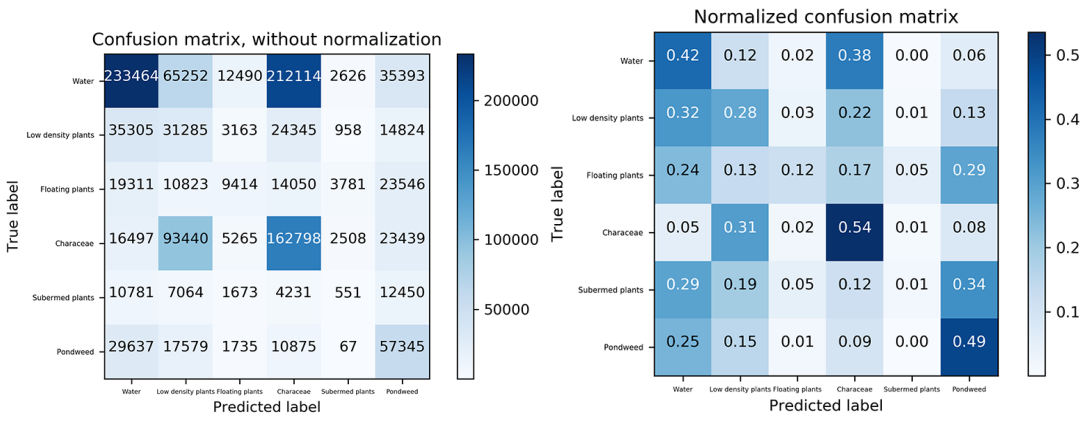


Figure 4.19: Confusion matrix using August 6 Sentinel-2 data

5

Discussion

Assumptions in our approach

Our approach is based on some assumptions. Firstly, the field samples were collected in mixed-species condition. We assumed that each field sample only contained one species that had maximum percentage. The purity ranged from 0.24 to 0.97, which indicated that features extracted from a pixel with low purity were not typical features from that particular species.

Secondly, the field samples were collected from 1 square meter while the highest resolution of Sentinel-2 images were 10×10 square meters. We assumed that a pixel in Sentinel-2 data had the same distribution with the field sample that was located in this area. This assumption may conflict with species with small distribution pattern or low richness.

Thirdly, we conducted segmentation in GRASS GIS toolbox and then labelled those objects with the class that had maximum number of field samples in this object. Then we assumed that all pixels in individual objects were from the same class. Although the average purity showed in the segmentation results was above 0.8 for each class, it remained questionable for large objects because of various local conditions.

Our last assumption is that aquatic vegetation from July to August remained unchanged, when testing pre-trained RF model using data from August. Based on this assumption, the classification accuracy was only 39%. We made a comparison between the predictions in July and August in Eemmeer. (figure 5.1). Vegetation cover in August was much less than in July, but there still clearly contained the pattern of two vegetation regions separated by water body.

In order to investigate the vegetation changes in the whole year, we processed Sentinel-2 data from Janurary to November in 2018 with cloud cover under 10%. The sum of NDVI in Eemmeer was calculated and regarded as total vegetation. From figure 5.2, there is an abrupt increase in May and drop in July. In principle, this change pattern agrees with the knowledge that aquatic vegetation

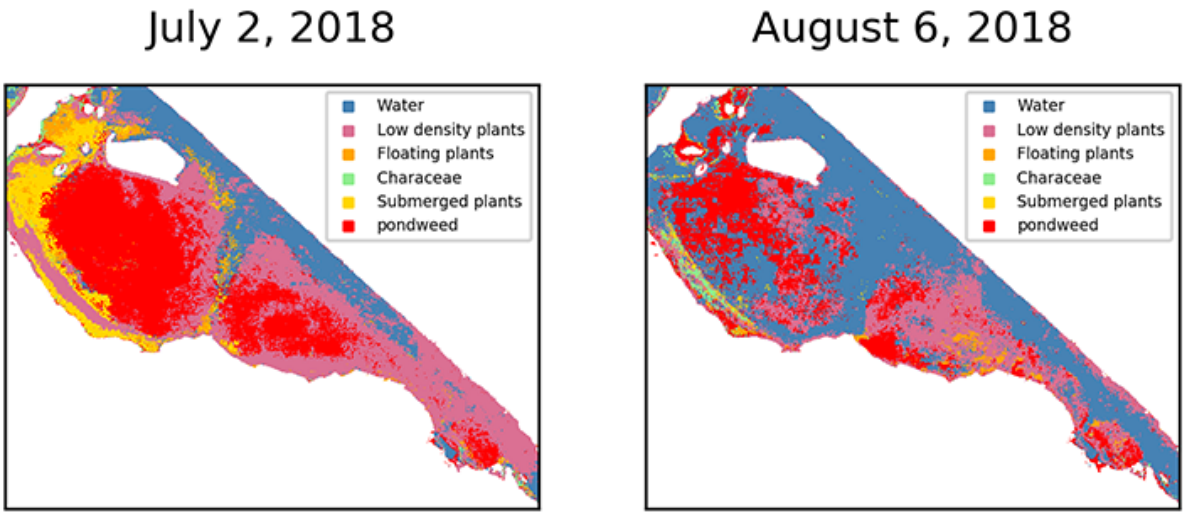


Figure 5.1: Predicted map in Eemmeer in 2018

usually explodes in summer. Therefore, the low classification accuracy could also be caused by inappropriate assumption, rather than bad RF model. In other words, it is not necessary to determine that the low classification accuracy means a bad prediction. Field data in August is needed in order for further assessment.

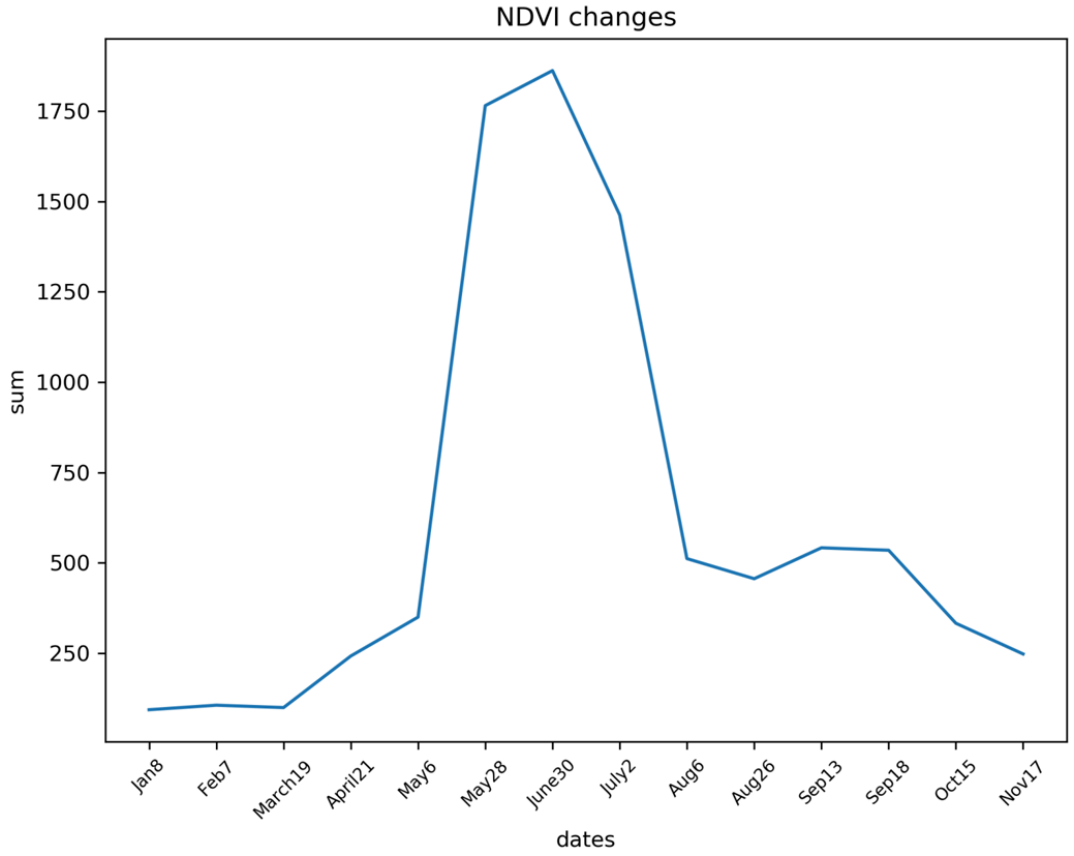


Figure 5.2: NDVI changes in Eemmeer in 2018

Image preprocessing

Satellite image preprocessing is a crucial step of this research. We adopted the simplest atmospheric correction method provided by the SCP plug-in in QGIS. However, data from different days are acquired under different conditions. For example, the cloud cover on July 2 was 0.0039% while that on August 6 was 1.5279%. There were also differences in solar irradiance (figure 5.3). As the atmospheric correction method utilizes the darkest pixel value in each band, different solar irradiance lead to different darkest pixel values in the same bands of different dates. To what extent these differences impact on final classification should be explored further. A training model only using data from one day maybe not easily applicable to other dates.

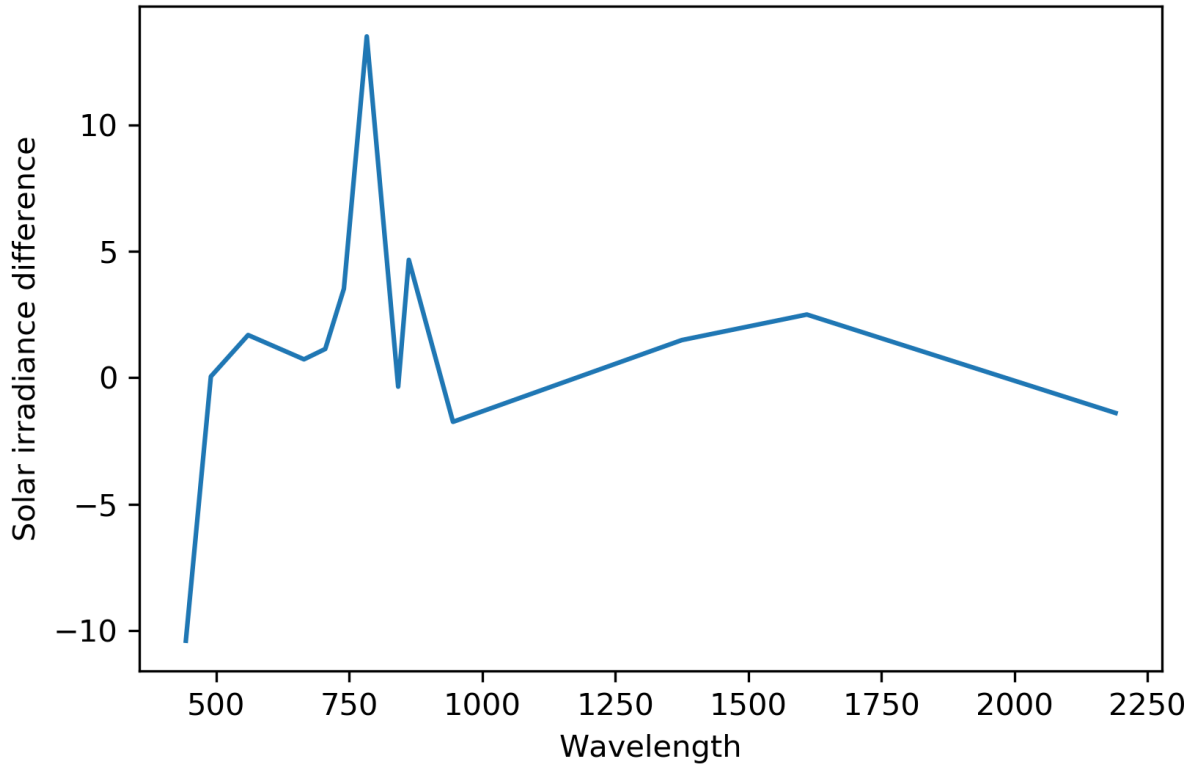


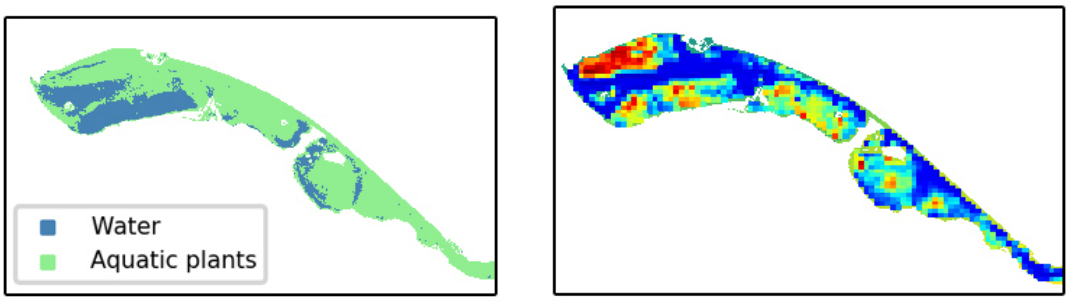
Figure 5.3: Solar irradiance differences of two dates. The metadata file recorded one solar irradiance value for each band. This plot was obtained by subtracting solar irradiance of August from July for each band

Unsupervised classification

We have performed two experiments using unsupervised machine learning approach. In these two experiments, we used the simple K-means clustering without using any pre-known information, such as large proportion of water class and tiny proportion of floating plants class. The only input was feature vectors and the number of clusters, which was set to 6 in the first experiment and 2 in the second experiment. However, vegetation class classification accuracy was 46% and the prediction was almost random.

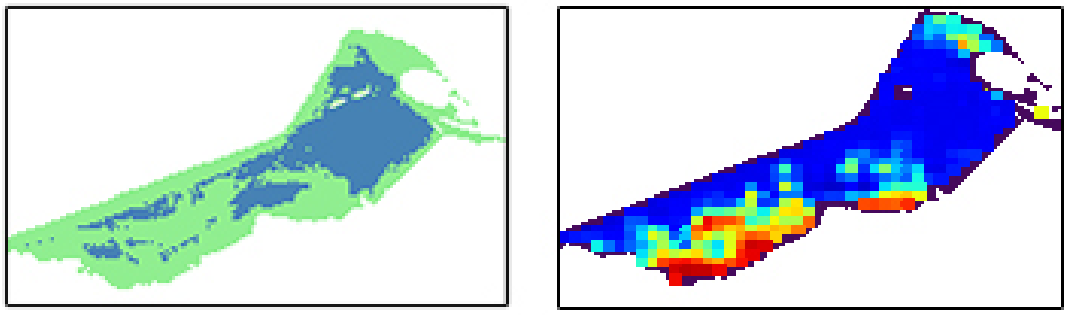
The water-vegetation classification accuracy was 55%. Although this was still not a satisfying result,

these two sites matched the field observations better (figure 5.4, 5.5), comparing to vegetation class classification experiments. In Veluwemeer, most pixels were classified as water, as in that region vegetation grew in deep water and most vegetation was brown colored Characeae, which was not easily detectable. Therefore, unsupervised classification method maybe not suitable for large water bodies, due to various water depth and complex location conditions.



(a) K-means results in site 1 (b) Interpolated ground truth in site 1

Figure 5.4: Comparison K-means results with field observations in site 1



(a) K-means results in site 2 (b) Interpolated ground truth in site 2

Figure 5.5: Comparison K-means results with field observations in site 2

Individual species

DF and KB are two species that cause most problems. Using Sentinel-2 data, our approach can extract unique reflectance patterns of DF, but failed to extract that of KB. The features extracted from KB were actually the combination of many factors due to complex local conditions. If new species in characeae family begins to grow in the neighborhood of KB, there will be difficulties in discriminating KB from other characeae species.

New species in the future

Our approach can only classify vegetation class using in the training process. It cannot predict new vegetation species that appear in the future. Therefore, it is necessary to update the model regularly.

Airborne images

Using high spectral resolution such as airborne hyperspectral images is unnecessary in these classification tasks, as Sentinel-2 can show similar reflectance patterns as hyperspectral images. Airborne multispectral images can show vegetation details in complex conditions, such as color, depth and whether there are algae above, while Sentinel-2 data cannot. Also, the details in the multispectral images can be seen directly by human visual system. This advantage can be beneficial for species with color that is similar to water bottom, such as brown. Most of VIs are designed based on green level, because of which brown species are not distinct in Sentinel-2 images. Therefore,

6

Conclusions

6.1. Classification results overview

Results presented in table 6.1 showed that supervised method has much higher accuracy in most tasks than that of baseline, and performs better in vegetation class classification than unsupervised method. The tasks of individual species classification were to classify DF and other pondweed within pondweed class, and to classify KB and other characeae within characeae class. The two tasks reached highest accuracy than others, which are 96% and 95% respectively. The test accuracy using data from August is 39%, using field data from July as validation data. The accuracy is lower than baseline, however, this accuracy is questionable due to dynamic vegetation changes during July and August. Field data from August is needed to get the true performance of pre-trained model. The unsupervised method has higher accuracy in water-aquatic plants classification task (55%) than vegetation class classification task (46%). Their accuracy are similar with baseline, which means this unsupervised approach failed to outperform the majority classifier.

Method	Task		Accuracy	Baseline
Supervised classification (RF classifier)	Vegetation class		87%	44%
	Individual species	DF	96%	56%
		KB	95%	59%
	Test on another date		39%	44%
Unsupervised classification (K-means clustering)	Vegetation class		46%	44%
	Water - aquatic plants		55%	55%

Table 6.1: Classification results overview

For supervised classification, one of the advantages is high accuracy, and another advantages is

that money cost can be decreased to a certain extent due to less field samples required. The disadvantage, however, is the need of field data, which means labor cost and error caused by vegetation changes are still unavoidable. On the other hand, new species and large changes cannot be detected using supervised classification, which means it is necessary to conduct regular field survey. For unsupervised classification, the biggest advantage is no field data required. The trade-off then is low accuracy. By weighing between the advantages and disadvantages, supervised classification using RF classifier is the best method.

6.2. Research questions

1. *What are the unique properties to discriminate aquatic plant species?*
2. *Can we extract those unique properties from Sentinel-2 images?*

In Sentinel-2 data aquatic plants is easily discriminated from water body using VIs such as GNDVI. GNDVI as the most important feature in vegetation class classification is showed in figure 6.1, water has a relatively low GNDVI, compared to aquatic plants, though there is also a peak around zero value, which may be caused by the reflectance from algae above water body. Characeae largely overlapped with water in this distribution, because Characeae grew in very deep region and the reflectance was attenuated by water column. Characeae (ranges from -0.4 to -0.1) and floating plants (ranges from -0.1 to 0.1) showed distinctive GNDVI distributions.

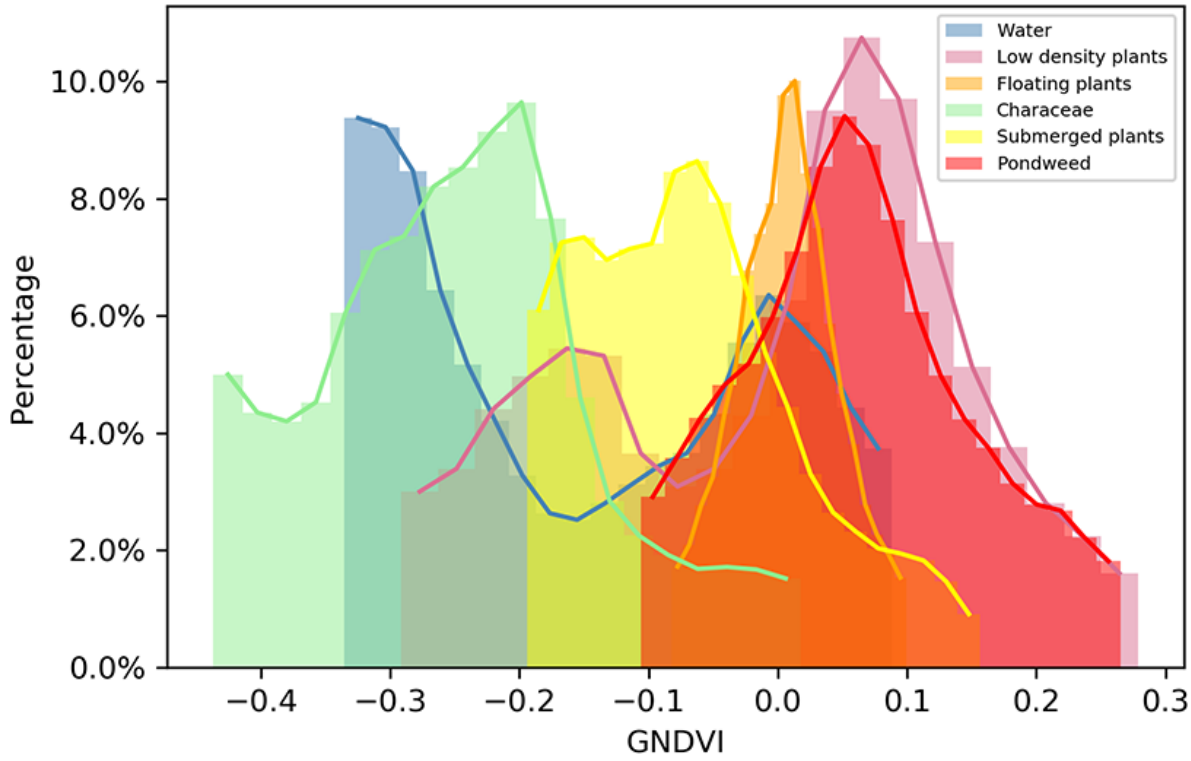


Figure 6.1: GNDVI distribution in six-category classification using RF classifier

3. *What classification result can Sentinel-2 images yield?*
When using supervised machine learning algorithms such as RF classifier, an average accuracy of 87% can be reached, with the help of 2889 field samples. All six vegetationn classes had relatively even classification accuracy. Doorgroeid fonteinkruid and Kransblad that caused main problems in Randmeren can be detected with accuracy 96% and 95% respectively.

An overall accuracy of 39% was obtained when applying RF model trained based on data from July to data from August. Water, Characeae and Pondweed got relatively higher accuracy than other categories, which were 42%, 54% and 49% respectively. However, this test result is questionable because there might be vegetation changes during July and August. In this case, field data from July may be not suitable to be used as validation data.

4. *Are field data necessary in remote sensing techniques?*
When using unsupervised machine learning algorithms such as K-means clustering, however, the average accuracy was only 46%. Most pixels were classified as water and Characeae. Even in classification between water and aquatic plants using K-means clustering, only 55% overall accuracy can be reached. From visually comparison with interpolated total vegetation cover ground truth map, the K-means results matched better in Eemmeer than other regions. The overall performance indicates that K-means clustering is not suitable for large water bodies. Therefore, field data are still needed.

5. *To what extent spectral and spatial resolution influence the classification result respectively?*
From table 6.2, vegetation indices contribute most in vegetation class classification task and KB classification task, while reflectance principal components contribute most to DF classification. For all classification tasks, the average importance of vegetation indices counts most, while that of LBP counts least. Spectral features contribute most in all classification tasks, which indicates the spectral resolution of Sentinel-2 data is sufficient for these classification tasks; while texture features contribute least due to coarse spatial resolution of Sentinel-2 data.

Features		Vegetation class	Individual species		Average importance
			DF	KB	
Spectral features	VIs	44%	37%	52%	44.3%
	Reflectance	29%	48%	35%	37.3%
Texture features	LBP	14%	7%	5%	8.7%
	Gabor filters	13%	8%	8%	9.7%

Table 6.2: Feature importance in each task

6. *Can Sentinel-2 data be an alternative to airborne data?*
In shallow water bodies, unique spectral features of each vegetation class can be extracted from Sentinel-2 data. In this case, Sentinel-2 data can be an alternative to airborne data. When in deep

water bodies like Veluwemeer, however, it is difficult to analysis its components because of attenuated reflectance. Airborne data with resolution 0.2×0.2 can give more details than Sentinel-2 data. Thus, Sentinel-2 data cannot replace airborne data in complex conditions.

7. *Can remote sensing techniques be used for the water bodies as a tool to achieve faster and more efficient monitoring of the vegetation?*
The average purity of segmentation result is 0.87. For high purity segments such as figure 3.15b, less field samples are required. Thus, labor cost will be decreased. Suppose that in each segment region, only one sixth field samples are collected, then the time will be decreased from two months to ten days. The error caused by vegetation changes during the field survey can also be minimized. For every month, there are available Sentinel-2 data from two or three separate dates. This means aquatic vegetation monitoring can be conducted every month, by using remote sensing. It is much faster than field survey every three years. Therefore, remote sensing techniques can be used for water bodies as a tool to achieve faster and more efficient monitoring of the vegetation.

6.3. Recommendations

Remote sensing data
Remote sensing techniques can be used as a tool for efficient aquatic vegetation monitoring. Using satellite Sentinel-2 data with cloud cover less than 10% is sufficient for vegetation class classification in Gooimeer, Eemmeer and Zwarte Meer. It is better to use Sentinel-2 data from different dates rather than a single day, in order to improve the robustness of the model. For other water bodies in Randmeren with complex conditions and deep regions, high spatial resolution airborne images are recommended, as more details can be observed directly by human visual system. This is a cost-effective way because free satellite images will be used for most regions and airborne images will only be taken for small parts of Randmeren.

Image preprocessing
Seninel-2 data preprocessing can be done using Semi-Automatic Classification (SCP) Plugin in QGIS. However, Sentinel-2 data acquired from different dates usually have different level of cloud cover, solar irradiance and other parameters. Whether these differences can be normalized in this process needs further investigation.

Image segmentation and field survey
Segmentation can be conducted in GRASS GIS, using region-based method and setting the minimum number of pixels in a segment as 10. As the resolution of Sentinel-2 images ranges from 10 m to 60 m. Images with resolution 10 m is selected for segmentation. The average purity of the segmentation result in our study is 0.87. This high purity means in the following years, less field samples are required. For example, as shown in figure 6.2, segmentation has grouped these pixels into one segment, which means all these pixels are labelled as the same class. In this segment, 65 samples were collected in which 61 were water class. Water class counts most, so all pixels are la-

belled as water class. Assume that only four samples are collected and they are from water class, the segment can still be labelled as water class. Therefore, for high purity segments, decreasing the number of field samples may not influence the final results.

Feature selection

Our study has shown that texture features contributed little to the classification, while vegetation indices contributed most. Therefore, it is recommended to select more vegetation indices and abandon texture features. Harris lists various vegetation indices with different purposes, which may helpful for discriminating aquatic plant species.

Classification

Unsupervised algorithms are not recommended for large complex water bodies. Although supervised algorithms require field data, higher accuracy can be obtained than unsupervised algorithms.

Vegetation dynamic changes

Vegetation species and their distribution may have large changes within three years (figure 3.4). Pre-trained model cannot detect new species. Therefore, it is necessary to take regular field survey.

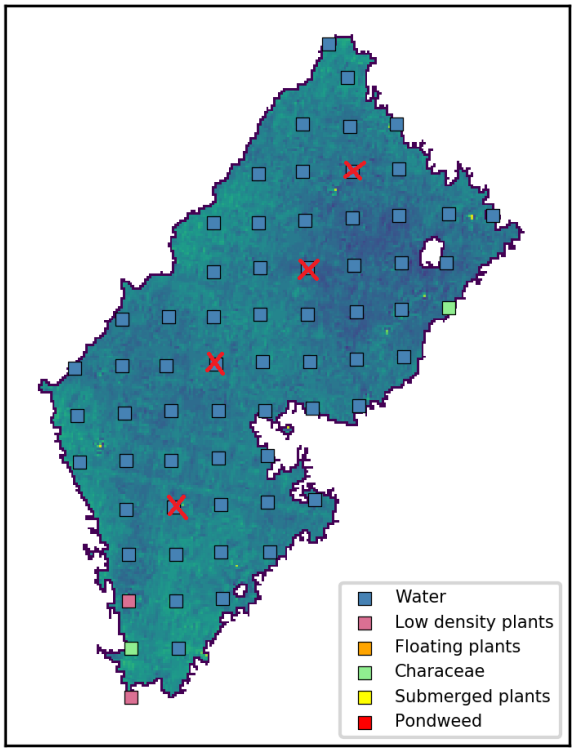


Figure 6.2: Future field collection example. The red crosses refer to possible remained field samples, other field samples may not be collected in the next field survey.

Bibliography

Adam, E., Mutanga, O., and Rugege, D. (2010). Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: A review. *Wetlands Ecology and Management*, 18:281–296.

Allan, J. (1990). 1 - sensors, platforms and applications; acquiring and managing remotely sensed data. In STEVEN, M. and CLARK, J., editors, *Applications of Remote Sensing in Agriculture*, pages 3 – 18. Butterworth-Heinemann.

Arabi, B., Salama, M., Bayat, B., and Verhoef, W. (2016). Monitoring of water quality in high turbid waters using coupled atmospheric-hydro-optical models and remote sensing observations (case study: The wadden sea). : Agu ocean science meeting at the ernest n. morial convention center.

Autrey, B. and Reif, M. (2011). Segmentation and object-oriented classification of wetlands in a karst florida landscape using multi-season landsat7 etm+ imagery. *International Journal of Remote Sensing - INT J REMOTE SENS*, 32:1471–1489.

Bartodziej, W. and Ludlow, J. (1997). Aquatic vegetation monitoring by natural resources agencies in the united states. *Lake and Reservoir Management*, 13(2):109–117.

Blaschke, T. (2010). Object based image analysis for remote sensing. isprs j photogramm remote sens. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65:2 – 16.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Buchroithner, M. F. (2000). *Remote Sensing for Environmental Data in Albania:A Strategy for Integrated Management*. Springer Netherlands.

C. Frohn, R., Reif, M., Lane, C., and Autrey, B. (2009). Satellite remote sensing of isolated wetlands using object-oriented classification of landsat-7 data. *Wetlands*, 29:931–941.

C Ritchie, J., Zimba, P., and H Everitt, J. (2003). Remote sensing techniques to assess water quality. *Photogrammetric Engineering Remote Sensing*, 69.

Carpenter, S. and Lodge, D. (1986). Effects of submersed macrophytes on ecosystem processes. *Aquatic Botany*, 26:341–370.

Chaki, J. (2012). Plant leaf recognition using gabor filter.

Chatterjee, S. (2017). An analysis of threats to marine biodiversity and aquatic ecosystems.

Chen, Q., Yu, R., Hao, Y., Wu, L., Zhang, W., Zhang, Q., and Bu, X. (2018). A new method for mapping aquatic vegetation especially underwater vegetation in lake ulansuhai using gf-1 satellite data. *Remote Sensing*, 10:1279.

Cheng, G. and Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:11–28.

Council, N. R. (1992). *Restoration of Aquatic Ecosystems: Science, Technology, and Public Policy*. The National Academies Press, Washington, DC.

Dar, N., Kumar Pandit, A., and Ganai, B. (2014). Factors affecting the distribution patterns of aquatic macrophytes. *Limnological Review*, 14.

Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.*, 40.

Diop, F. N. (2010). The monitoring of flora and aquatic vegetation.

Dronova, I. (2015). Object-based image analysis in wetland research: A review. *Remote Sensing*, 7:6380–6413.

Freedman, B. and Lacoul, P. (2006). Environmental influences on aquatic plants in freshwater ecosystems. *Environmental Reviews*, 14:89–136.

G. Daugman, J. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America. A, Optics and image science*, 2:1160–9.

Green, A. A., Berman, M., Switzer, P., and Craig, M. D. (1988). A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):65–74.

Hadjimitsis, D. G., Papadavid, G., Agapiou, A., Themistocleous, K., Hadjimitsis, M. G., Retalis, A., Michaelides, S., Chrysoulakis, N., Toulios, L., and Clayton, C. R. I. (2010). Atmospheric correction for satellite remotely sensed data intended for agricultural applications: impact on vegetation indices. *Natural Hazards and Earth System Sciences*, 10(1):89–95.

Haralick, R., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Trans Syst Man Cybern*, SMC-3:610–621.

Hauxwell, J., Knight, S., Wagner, K., Mikulyuk, A., Nault, M., and Chase, S. (2019). Recommended baseline monitoring of aquatic plants in wisconsin: Sampling design, field and laboratory procedures, data entry and analysis, and applications.

Ho, T. (1998). The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20:832 – 844.

Husson, E., Ecke, F., and Reese, H. (2016). Comparison of manual mapping and automated object-based image analysis of non-submerged aquatic vegetation from very-high-resolution uas images. 8(9):724.

Ibarrola, E., Marcello, J., and Gonzalo-Martin, C. (2017). Assessment of component selection strategies in hyperspectral imagery. *Entropy*, 19:666.

Jakubauskas, M., Peterson, D., D. Campbell, S., and Penny, D. (2012). Mapping and monitoring invasive aquatic plant obstructions in navigable waterways using satellite multispectral imagery.

Jindal, P. and Kumar, D. (2017). A review on dimensionality reduction techniques. *International Journal of Computer Applications*, 173:42–46.

Jinru, X. and Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, 2017:1–17.

K Pai, A., A K, K., and U, R. (2017). A novel crowd density estimation technique using local binary pattern and gabor features. pages 1–6.

Kiang, N. Y., Siefert, J., Govindjee, Blankenship, R. E., and Meadows, V. (2007). Spectral signatures of photosynthesis. 1. Review of Earth organisms. *Astrobiology*, 7:222.

Kim, M., Warner, T., Madden, M., and S. Atkinson, D. (2011). Multi-scale geobia with very high spatial resolution digital aerial imagery: Scale, texture and image objects. *International Journal of Remote Sensing*, 32:2825–2850.

Kobayashi, S. and Sanga-Ngoie, K. (2009). A comparative study of radiometric correction methods for optical remote sensing imagery: the irc vs. other image-based c-correction methods. *International Journal of Remote Sensing*, 30(2):285–314.

Kruse, F., Lefkoff, A., Boardman, J., Heidebrecht, K., Shapiro, A., Barloon, P., and Goetz, A. (1993). The spectral image processing system (sips)-interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, 44:145–163.

Lembi, C. A. (2009). Aquatic plant management: Identifying and managing aquatic vegetation.

Li, Z., Hayward, R., Zhang, J., Jin, H., and Walker, R. (2010). Evaluation of spectral and texture features for object-based vegetation species classification using support vector machines. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 38.

Liang, S., Li, X., and Wang, J. (2012). Chapter 1 - a systematic view of remote sensing. In *Advanced Remote Sensing*, pages 1 – 31. Academic Press, Boston.

Liu, Y., Zhang, D., Lu, G., and Ma, W.-Y. (2006). Study on texture feature extraction in region-based image retrieval system.

Luo, G., Chen, G., Tian, L., Qin, K., and Qian, S.-E. (2016). Minimum noise fraction versus principal component analysis as a preprocessing step for hyperspectral imagery denoising. *Canadian Journal of Remote Sensing*, 42(2):106–116.

Luo, J., Duan, H., Ma, R., Jin, X., Li, F., Hu, W., Shi, K., and Huang, W. (2017). Mapping species of submerged aquatic vegetation with multi-seasonal satellite images and considering life history information. *International Journal of Applied Earth Observation and Geoinformation*, 57:154–165.

Lyzenga, D. (1978). Passive remote-sensing techniques for mapping water depth and bottom features. *Applied optics*, 17:379–83.

Ma, Z. and Zhou, G. (2018). Canopy modeling of aquatic vegetation: Construction of submerged vegetation index. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-3:1253–1258.

Madsen, J. (1999). Point intercept and line intercept methods for aquatic plant management. page 16.

Madsen, J. D. and Wersal, R. M. (2012). A review of aquatic plant monitoring and assessment methods.

Maillard, P. (2005). Comparing texture analysis methods through classification.

Manessa, M., Haidar, M., Budhiman, S., Winarso, G., Kanno, A., Sagawa, T., and Sekine, M. (2016). Evaluating the performance of lyzenga's water column correction in case-1 coral reef water using a simulated wolrdview-2 imagery. *IOP Conference Series: Earth and Environmental Science*, 47.

Murray, H., Lucieer, A., and Williams, R. (2010). Texture-based classification of subantarctic vegetation communities on heard island. *International Journal of Applied Earth Observation and Geoinformation - INT J APPL EARTH OBS GEOINF*, 12:138–149.

Nalepa, J. and Kawulok, M. (2018). Selecting training sets for support vector machines: a review. *Artificial Intelligence Review*.

Pande-Chhetri, R., Abd-Elrahman, A., and Jacoby, C. (2014). Classification of submerged aquatic vegetation in black river using hyperspectral image analysis. *GEOMATICA*, 68:169–182.

Perivolioti, T., Mouratidis, A., Doxani, G., and Bobori, D. (2016). Monitoring the water quality of lake koronia using long time- series of multispectral satellite images.

Rodriguez-Galiano, V., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93 – 104.

Sagawa, T., Boisnier, E., Komatsu, T., Ben Mustapha, K., Hattour, A., Kosaka, N., and Miyazaki, S. (2010). Using bottom surface reflectance to map coastal marine areas: A new application method for lyzenga's model. *International Journal of Remote Sensing*, 31:3051–3064.

Satterwhite, M. B. and Henley, J. P. (1990). Hyperspectral signatures (400 to 2500 nm) of vegetation, minerals, soils, rocks, and cultural features: Laboratory and field measurements. page 478.

Seo, D.-J., Park, C.-H., Park, J.-H., and Tateishi, R. (1998). A search for the optimum combination of spatial resolution and vegetation indices. volume 3, pages 1729–1731.

S.K.Jain and V.P.Singh (2003). *Developments in water Science*, volume 51.

Sytsma, M. D. (2008). Introduction: workshop on submersed aquatic plant research priorities. *J. Aquat. Plant Manage*, 46:1–7.

Tian, Y. Q., Qian, Y., Zimmerman, M., Flint, S., and Waldron, M. C. (2010). Differentiating aquatic plant communities in a eutrophic river using hyperspectral and multispectral remote sensing.

Tsai, F. and Chou, M. (2006). Texture augmented analysis of high resolution satellite imagery in detecting invasive plant species. *Journal of the Chinese Institute of Engineers*, 29(4):581–592.

Tschirky, P., Hall, K., and Turcke, D. (2001). Wave attenuation by emergent wetland vegetation. volume 276, pages 865–877.

Wang, L., Sousa, W., and Gong, P. (2004). Integration of object-based and pixel-based classification for mapping mangroves with ikonos imagery. *International Journal of Remote Sensing - INT J REMOTE SENS*, 25.

Wang, M., Fei, X., Zhang, Y., Chen, Z., Wang, X., Yeu Tsou, J., Liu, D., and Lu, X. (2018). Assessing texture features to classify coastal wetland vegetation from high spatial resolution imagery using completed local binary patterns (clbp). *Remote Sensing*, 10:778.

Wood, E., Pidgeon, A., Radeloff, V., and S Keuler, N. (2012). Image texture as a remotely sensed measure of vegetation structure. *Remote Sensing of Environment*, 12:516–526.

Yang, S., Liao, X., and Borasy, U. (2012). A pedestrian detection method based on the hog-lbp feature and gentle adaboost. *International Journal of Advancements in Computing Technology*, 4:553–560.

Zhou, T., Li, Z., and Pan, J. (2018). Multi-feature classification of multi-sensor satellite imagery based on dual-polarimetric sentinel-1a, landsat-8 oli, and hyperion images for urban land-cover classification. *Sensors (Basel, Switzerland)*, 18.