A HYPERSPECTRAL APPROACH TO UNDERSTAND THE ASSOCIATION BETWEEN PSM (AS MEASURED BY INSAR DATA) AND VEGETATION ASSEMBLAGE FOR A SCOTTISH PEATLAND

Rachel Zoë Walker
Student Number: 20402170

Nottingham Geospatial Institute, University of Nottingham, UK

Dissertation submitted to The University of Nottingham in partial fulfilment of the degree of Master of Research in Geospatial Systems

August 2022

Abstract

Peatlands are vital ecosystems that store up to a third of terrestrial carbon despite covering 3 % of the land surface; it is therefore important to improve our understanding of these landscapes to enable continuous carbon sequestration as the climate changes. AVIRIS-NG hyperspectral data has the potential to add detail to current understanding from analysis of Peatland Surface Motion (PSM) from InSAR data. PSM is directly linked to vegetation assemblage, erosion rates and land use. Therefore four sites were chosen; one near natural, two undergoing restoration (starting a decade apart) and one that has been eroded. Machine learning was used to predict Plant Functional Types (PFTs) at each site using fieldwork, satellite imagery and expert knowledge. Exploratory analysis demonstrated that the random forest classifier was better at predicting PFTs in the Flow Country than SVM analysis (using either linear or RBF kernels). The fieldwork was mainly focused on the first restoration site as this site overlapped most with the others and ten PFTs were determined for this location, with an additional three added from fieldwork at the erosion site. Train-test data was created for these 13 PFTs and random forest classifiers applied to the data, the first restoration site underwent additional analysis using the fieldwork-focused specific train-test data to classify the data. The fieldwork-focused classification was the most successful with a mean accuracy score of 0.789, with the other mean accuracies ranging from 0.722-0.728, demonstrating the benefits of conducting fieldwork. Within this analysis, the whole dataset was utilised as well as smaller spectral ranges to determine whether all hyperspectral bands (post pre-processing) need to be used; it was found that the outcomes using the whole dataset were more accurate than the smaller spectral ranges. Additionally, the data was transformed with the original wavelengths, first and second derivatives and continuum removal used to classify the data, with the original and derivative outcomes proving more accurate than continuum removal. Supervised machine learning was much more successful at locating PFTs than the unsupervised k-means cluster analysis; it was concluded that k-means is unsuitable to predict PFT locations. The Peatland Surface Motion (PSM) data was analysed in conjunction with the PFT predictions for the first restoration site using a range of machine learning classification techniques (logistic regression, decision tree, random forest and SVM). Outcomes suggest that there is potential to use the hyperspectral analysis to increase understanding based on PSM outputs, however, further refinement of methods is required to achieve this.

Acknowledgements

My thanks go to EPSRC for funding the Geospatial Systems Centre for Doctoral Training studentship, giving me the opportunity to pursue postgraduate research.

Thank you to my project supervisors; Professor Doreen Boyd and Doctor David Large for their support and insight throughout the dissertation. Also, thanks to Andrew Bradley for his support in using the InSAR data, Stephen Grebbe for advice regarding hyperspectral data and Roxane Anderson for suggesting sites and providing detailed site descriptions in the early stage of the project when fieldwork was not possible.

Thanks the RSPB for giving me permission to conduct fieldwork and to Sean Ince for training me to use the GNSS equipment. Thanks also go to Colin Brodie-Smith for accompanying me to the Flow Country and assisting with field data collection.

0.1 Declaration

Declaration: "I hereby certify that this work is my own, except where otherwise acknowledged, and that it has not been submitted previously for a degree at this, or any other university."

RACHEL ZOE WALKER
Date: Friday 26th August 2022

Contents

Abstract **i**

Acknowledgements **iii**
0.1 Declaration iii

List of Figures **vii**

List of Tables **x**

List of Abbreviations **xi**

1 Introduction **1**
1.1 Blanket Bogs 1
1.2 Earth Observation 3
1.2.1 Hyperspectral Data 3
1.2.2 InSAR Data and Peat Surface Motion 4
1.3 Motivations and Research Purpose 5

2 Literature Review **7**
2.1 Remote Sensing Background 7
2.2 Remote Sensing and Peatlands 8
2.3 Flow Country Review 11
2.3.1 PSM Analysis 12
2.3.2 Hyperspectral Analysis in the Flow Country 13
2.4 Machine Learning 14
2.4.1 Supervised Machine Learning 14
2.4.2 Unsupervised Machine Learning 15
2.5 Gaps in the Literature 16

3 Aim and objectives **17**
3.1 Aim 17
3.2 Objectives 17

4 Methodology **18**
4.1 Data Sources and Software 18
4.2 Site Selection 20
4.3 Data Preparation for Machine Learning 22

4.3.1 Pre-processing 22
4.3.2 Data Transformations 22
4.3.3 Spectral Ranges 23
4.4 Machine Learning for PFT Prediction 23
4.4.1 Supervised Learning 23
4.4.2 Unsupervised Learning 29
4.4.3 Field Sites 29
4.5 Peatland Surface Motion and PFT Comparison 31

5 Results **34**
5.1 PFT Classifications 34
5.1.1 Random Forest PFT Predictions 34
5.1.2 K-means Cluster Outputs 38
5.2 PSM Relationship With PFTs 40

6 Discussion **43**
6.1 Machine Learning PFT predictions 43
6.1.1 Random Forest PFT Predictions 43
6.1.2 K-means Clustering 49
6.2 Using PSM to Predict PFTs 49
6.3 Applications 51
6.4 Future Research 52
6.4.1 PFT-focused 52
6.4.2 PFT and PSM 53
6.4.3 Wider Reaching Research 53

7 Conclusions **54**

Bibliography **57**

A Appendix A: Fieldwork Risk Assessment **67**

B Appendix B: Data collection sheet **68**

C Appendix C: Histograms and correlation matrix **69**

D Appendix D: Complete table of highest outcomes **72**

E Appendix E: Cross Lochs maps from intial analysis **73**

F Appendix F: Data Management Plan **75**

G Appendix G: Gantt Chart **79**

List of Figures

1.1 The Flow Country is located in northeast Scotland (Cameron 2019) . 3

1.2 Spectral signatures of mosses (He et al. 2011) 4

1.3 Comparison of InSAR and ground PSM data for low lying peat at the Munsary Site. Drought caused issues, meaning that there was increased variation from summer 2018, however, prior to that the accuracy of the InSAR was affected by microtopography (Marshall et al. 2022) 5

4.1 Pushbroom imaging mode, where FPA stands for focal plan array Jia et al. 2020 18

4.2 The locations of the AVIRIS-NG Europe 2021 campaign sites (ARES 2021) 19

4.3 The flight paths used to collect the hyperspectral data, July 2021 ((NASA JPL 2022) 20

4.4 Four 1km^2 site locations within the Flow Country. Cross Lochs (A) is a near-natural site (control site) with central coordinates of -3.937388, 58.386329, there are two restoration sites (B and C) at different stages of transition from plantation to peat with central coordinates of -3.975073, 58.385097 and -3.973670, 58.411235, and an erosion site (D) with area of bare peat and cuttings with central coordinates of -3.879604,58.469920. 21

4.5 Reflectance spectra of green, stressed and drying vegetation (EO College 2021) 24

4.6 Summary of machine learning methods used for each objective . . . 24

4.7 Spectral library for Cross Lochs developed from expert descriptors and satellite imagery to generate mean spectral signatures for each PFT 25

4.8 Scatter plot for the two dominant PCA bands regarding the train-test data for the exploratory analysis of Cross Lochs 26

4.9 Supervised learning PFT identification accuracies for exploratory analysis of Cross Lochs. The machine learning classifications were run for a range of train-test ratio, with better outcomes for the random forest classifier than SVM classification 27

4.10 Annotated satellite image; blue annotations were from expert description of the area with yellow additions post fieldwork 28

4.11 Spectral library for the first restoration site based on data collected on fieldwork with signatures of each PFT identified 29

4.12 Clusters of areas containing different potential PFTs based on analysis of satellite imagery and expert descriptions with random sites selected for data collection in the field 30
4.13 The field sites where data was ultimately collected from 31
4.14 Locating sites using the GNSS receiver in the first restoration site, with the quadrat then build centred around the receiver 32
4.15 Species identification and proportioning in a quadrat at the erosion site 32

5.1 The highest outcomes for each site and for each train-test ration, maximum depth and internal cross validation 34
5.2 A summary of the highest five random forest classification outcomes for each iteration through the first restoration site using the fieldwork-focused train-test data 35
5.3 Random forest classification map of predicted PFTs for the first restoration site with 10 classes using fieldwork-based train-test data (accuracy 0.770.05 36
5.4 Random forest classification map of predicted PFTs for the four sites using the whole train-test dataset which was applied across the sites 37
5.5 K-means clustering for the four sites with 'k' detemined by the number of PFTs 40
5.6 Mean accuracy of difference machine learning classifications when using PSM to predict PFTs. 'Fw' relates to the predictions determined from the fieldwork-focused train-test data and 'all' refers to predictions determined using the full train-test dataset 41
5.7 K-means clustering, with ten clusters, for a range of PSM attributes in the first restoration site 42

6.1 Precipitation (at Forsinain) and temperature (at Wick airport) each month from August 2020 to July 2022 (Met Office 2022, SEPA 2022) . 45
6.2 Scatter plot for the two dominant PCA bands regarding the fieldwork-focused train-test data. 47

B.1 Data collection sheet for fieldwork, July 2022. The subsites were numbered and ordered prior to data collection, the GPS number was the record number on the GNSS receiver and the start time was included so that relevant photos could be linked. Species percentages were recorded with space for additional species as required and other site notes. 68

C.1 Velocity histogram for the 'dead grass mix' PFT 69

C.2 Velocity histogram for the 'grass *Sphagnum*' PFT 70
C.3 Example correlation matrix, iterating through the PFTs (random points chosen from larger classes when a smaller class is the used as the basis of the matrix - smallest class removed with each iteration and determines the number of points for correlation) 71

E.1 K-means map with 6 clusters; cluster names based on location and similarities with the random forest maps. 73
E.2 Random forest map using the original spectra, focusing on the SWIR bands with 75% training data and 25% testing. 73
E.3 Random forest map using the first derivative, using all 358 bands with 70% training data and 30% testing. 74

G.1 Finalised Gantt chart regarding project management. The submission deadlines involve: the literature review poster, project proposal, data management plan, interim report and dissertation submission. Additionally, the first of each month sees the update of the data management plan, files backup and GitHub check. 79

List of Tables

2.1 Alternative uses of hyperspectral data to analyse peatlands 10

4.1 A summary of the hyperspectral and InSAR data 19

4.2 The full sepctral range and five sub-ranges of the electromagnetic spectrum used to analyse the reflectance of green vegetation (Erudel et al. 2017) 23

5.1 Highest ten outputs for the first restoration site using the fieldwork focused train-test data 35

5.2 Confusion matrices for the first restoration site predictions compared to the full train-test dataset, with an overall accuracy of 0.72. 38

5.3 Confusion matrices for the first restoration site predictions compared to the site-specific train-test dataset (from fieldwork) with an overall accuracy of 0.86 38

5.4 Confusion matrices for the predictions from all four sites compared to the full train-test dataset, with an overall accuracy of 0.75 39

5.5 The comparison of the classifications derived from the confusion matrices (Tables 5.2 and 5.3, generated from the predictions from the first restoration site using the focused train-test dataset and the full train-test dataset. z = 5.196 39

5.6 The comparison of the classifications derived from the highest two confusion matrices for the first restoration site with a testing size of 0.25 and maximum depth of 5, generated from the predictions from the first restoration site using the focused train-test dataset. z = 2.708 39

A.1 Faculty of Engineering risk assessment, completed for fieldwork in the Flow Country 67

D.1 Full table of all the first restoration site top outcomes (highest five for each iteration) using the focused train-test dataset 72

List of Abbreviations

PFT	**P**lant **F**unctional **T**ype
PSM	**P**eat **S**urface **M**otion
InSAR	**In**terferometric **S**ynthetic **A**perture **R**adar
CO_2	**C**arbon di**o**xide
NIR	**N**ear **I**nfra**r**ed
SWIR	**S**hort**w**ave **I**nfra**r**ed
PCA	**P**rincipal **C**omponent **A**nalysis
SVM	**S**upport **V**ector **M**achine
RBF	**R**adial **B**asis **F**unction
UAV	**U**nmanned **A**erial **V**ehicle
GIS	**G**eographical **I**nformation **S**ystems
GDAL	**G**eospatial **D**ata **A**bstraction **L**ibrary
AVIRIS-NG	**A**irborne **V**isible **I**nfra**R**ed **I**maging **S**pectroscoper - **N**ext **G**eneration
JPL	**J**et **P**ropulsion **L**aboratory
SHAC	**S**AR and **H**ypersepectral **A**irborne **C**ampaign
LiDAR	**Li**ght **D**etection **A**nd **R**anging
CASI	**C**ompact **A**irborne **S**pectographic **I**mager
DOC	**D**issolved **O**rganic **C**arbon
GPP	**G**ross **P**rimary **P**roductivity
NDVI	**N**ormalised **D**ifference **V**egetation **I**ndex
PRI	**P**hotochemical **R**eflectance **I**ndex
TP	**T**rue **P**ositive
FP	**F**alse **P**ositive
TN	**T**rue **N**egative
FN	**F**alse **N**egative
EPSRC	**E**ngineering and **P**hysical **S**ciences **R**esearch **C**ouncil
RSPB	**R**oyal **S**ociety for the **P**rotection of **B**irds

Introduction

1.1 Blanket Bogs

A blanket bog is a distinctive biome where peat covers the entire landscape, only broken where there are steep slopes (Gallego-Sala and Prentice 2013). They are ombrotrophic peatlands that develop in landscapes with a high number of rainfall days, low temperature range (maximum 15°C (Gallego-Sala and Prentice 2013)), low water pH and shallow slope gradients (Lindsay et al. 1988). These factors cause the peat to be anoxic and have slow rates of decomposition, enabling the build up of organic matter, resulting in low nutrient availability and plant productivity exceeding decomposition rates (Gallego-Sala and Prentice 2013, Hambley 2016, Levy and Gray 2015). The peatland is composed of two layers; the acrotelm (active layer) and catotelm (inert layer) (Lindsay et al. 1988, Marsden and Ebmeier 2012). The catotelm is a saturated layer comprised of compressed peat which is protected from external influences by the acrotelm, which is subject to changes in water table fluctuations, atmospheric exchange of gases and moisture, and is affected by vegetation, which helps to bind this layer together (Lindsay et al. 1988). These landscapes are typically composed of *Sphagnum* mosses and vascular plants (such as Calluna Vulgaris) with a range of microtopes including pools, hollows and hummocks (Gallego-Sala and Prentice 2013).

These landscapes provide a range of ecosystem services including climate regulation (they sequester and store carbon), water regulation and habitat provision for a range of wildlife, especially upland breeding birds (Bonn et al. 2014). Carbon sequestration only occurs in healthy peatlands (with low decomposition rates), with up to 2.8 tCO_2/ha/yr added to the peat (Lunt et al. 2019). As a result peatlands store about 30 % of terrestrial carbon, despite only covering 3 % of the land surface and are now recognised as vital ecosystems (Lunt et al. 2019, Marsden and Ebmeier 2012, Ratcliffe et al. 2018). Therefore, it is essential to monitor and restore these landscapes, as they are increasingly vulnerable to erosion and oxidation, because of human activity and climatic changes which enhance microbial activity, increasing the release of carbon (Gallego-Sala and Prentice 2013, Hambley 2016, Humpenöder et al. 2020, Marsden and Ebmeier 2012). As a result, the Kyoto Protocol recognises the importance of peatlands under Article 3.4, with wetland drainage and rewetting a recent addition (Bonn et al. 2014). Vegetation assemblage is a key area to study as

Sphagnum mosses are much more resistant to decay than grasses and sedges, meaning *Sphagnum*-based peat sequesters more carbon (Marsden and Ebmeier 2012). It is, therefore, important to research peatlands, especially factors, such as peatland management and vegetation assemblage, which affect the direction of carbon flow (Hambley 2016, Marsden and Ebmeier 2012).

About 20 % of Scotland is blanket bog, which equates to about 15 % of the global ecosystem (Marsden and Ebmeier 2012). The peatland of focus is the Flow Country, located in the northeast of the country (Figure 1.1); it is largest contiguous blanket bog in the UK covering over 4,000 km^2 and a proposed World Heritage Site (Alshammari et al. 2020, Alshammari et al. 2018, Lindsay et al. 1988, Marshall et al. 2022). This peatland started to form at the start of the Holocene (Hambley 2016, Levy and Gray 2015), ultimately extending to contain over four million tonnes of carbon. There is confidence that the Flow Country remains a carbon sink (Hambley et al. 2018, Levy and Gray 2015). However, due to human activity this peatland is increasingly at risk of becoming a carbon source, which would greatly increase Scotland's greenhouse gas emissions, potentially becoming responsible for up to 15 % of national emissions (Ferretto et al. 2021, Ferretto et al. 2019, Ratcliffe et al. 2018). This is further exacerbated by historical views, when peatlands were perceived as a wasted resource, which needed to be converted to agricultural land or woodland to increase their usefulness. As a result, some parts of the Flow Country were drained and forested prior to 1990, but more recently, there have been restoration projects to block drains and remove trees to improve their resilience to change and increase their carbon absorption potential (Alshammari et al. 2018, Hambley et al. 2018, Marsden and Ebmeier 2012). Restoration projects typically take between five and fifty years for degraded peatlands to return to bogs with a net carbon absorption (Hambley et al. 2018). Restoration began in 1998 and, with initial restoration sites showing rapid change toward bog-like conditions in the first six years, however, the rate stalled, especially in driers areas, in the following eight years (Hancock et al. 2018). Whereas, more recent restoration has shown faster rates of change (Marshall et al. 2021). The current restoration rate is 21,000 ha/yr and is driven by the Scottish government being committed to reduce carbon emissions by 80 % by 2050 (compared to the 1990-1995 baseline); something that will not be achieved if the peatland becomes a net carbon emitter (Ferretto et al. 2021).

FIGURE 1.1: The Flow Country is located in northeast Scotland (Cameron 2019)

1.2 Earth Observation

1.2.1 Hyperspectral Data

Hyperspectral data is a type of remotely sensed imagery which measures the reflected spectrum at wavelengths from 350-2,500 nm and a minimum of 21 channels (Kozma-Bognár and Berke 2010, He et al. 2011). The imagery can be collected using a range of technology, typically airborne (using planes or drones) or satellite, however, more recently there have also been developments regarding the collection of hyperspectral data using smart phones (Stuart et al. 2021). The hyperspectral imagery used in this study is AVIRIS-NG airborne data collected in July 2021. This data is three dimensional, with two spatial dimensions and one spectral (Hati et al. 2021, Jia et al. 2020, Luo et al. 2016), meaning it can provide details which are not visible to human eyes (in the infrared part of the electromagnetic spectrum). This enables better understanding of natural processes and the identification of specific aspects of the environment including land use and plant type (Figure 1.2) (GISG 2021, He et al. 2011, Zhong et al. 2018). This understanding is also due to the nature of the data collected; the wavelengths are narrow (0.01 μm) and continuous (Kale et al. 2017, Kozma-Bognár and Berke 2010), enabling spectral signatures to be identified and intricate information about the land surface (Kozma-Bognár and Berke 2010). This detail means it has the potential to be used to build on analysis from other data sources.

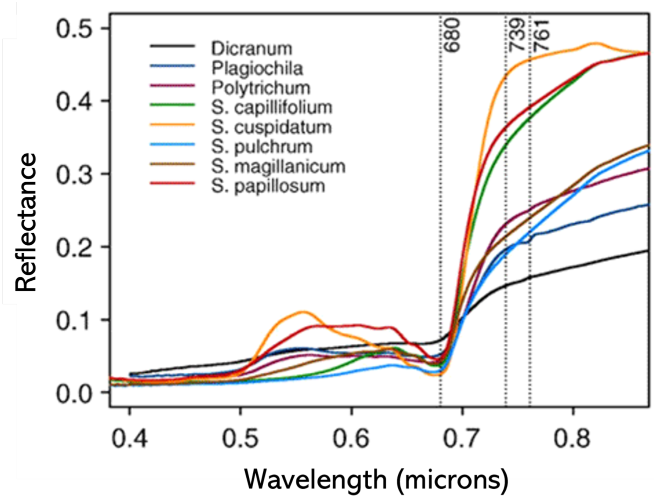


FIGURE 1.2: Spectral signatures of mosses (He et al. 2011)

1.2.2 InSAR Data and Peat Surface Motion

Interferometric Synthetic Aperture Radar (InSAR) is a radar technique that has been used to study the surface motion of peatlands (Alshammari et al. 2020, Alshammari et al. 2018, Bradley et al. 2021, Marshall et al. 2022, Zhou et al. 2019). As this technique uses radar, it is not limited by scale or cloud cover (Alshammari et al. 2018, Bradley et al. 2021) and as Sentinel-1 maps the globe every six days (previously 12 until the addition of a second satellite in 2016 ESA) there is the potential to perform detailed time-series analysis of Peat Surface Motion (PSM) using this data. This data has been used in preference to ground data as InSAR be used over much larger areas, is less expensive, and unaffected by weather conditions and accessibility issues (Alshammari et al. 2018 Bradley et al. 2021 Marshall et al. 2022), although this can come at the expense of accuracy (Marshall et al. 2022).

The research already undertaken focuses on PSM as this has been shown to be a good proxy for peatland condition linked to vegetation assemblage, water balance and erosion rates over the short-term, which helps to warn of long-term ecological changes and carbon losses (Alshammari et al. 2020, Kennedy and Price 2005, Marshall et al. 2022). Although PSM has been shown to be a representative indicator of peatland condition, it has been recognised that being used in combination with other Earth Observation techniques, a more holistic view of peatland condition can be developed (Marshall et al. 2021). In addition to this, InSAR data can be validated using ground data, however, uncertainty remains, especially in areas with pronounced variations in microtopography and stiffer peat (Marshall et al. 2022) (Figure 1.3) and, therefore, there is a need to compare outcomes to other data sources such as hyperspectral imagery.

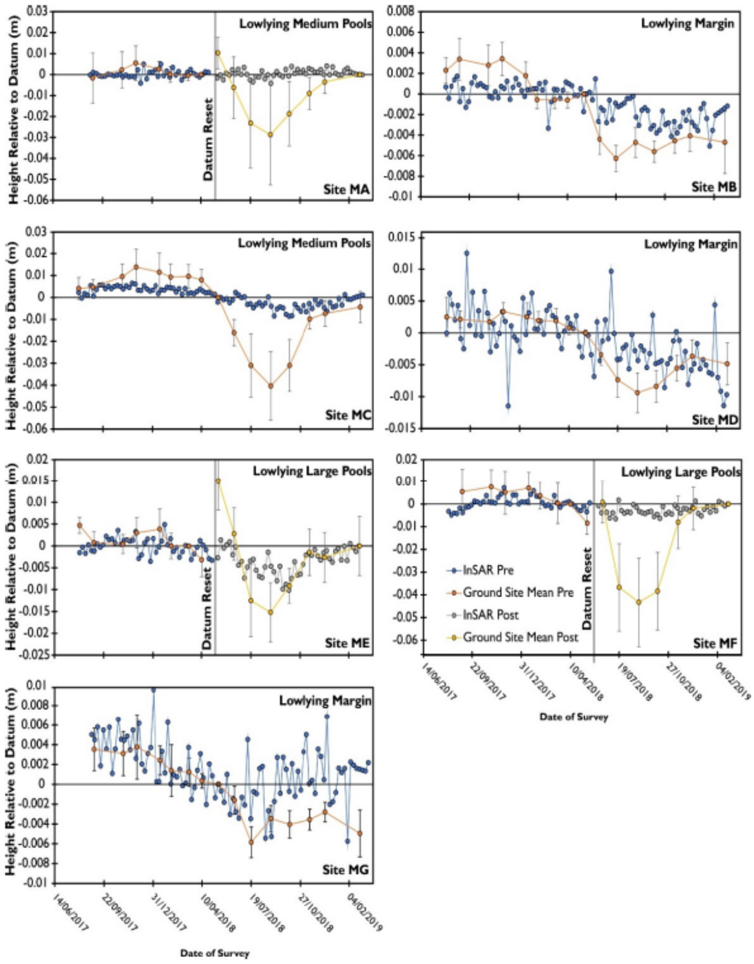


FIGURE 1.3: Comparison of InSAR and ground PSM data for low lying peat at the Munsary Site. Drought caused issues, meaning that there was increased variation from summer 2018, however, prior to that the accuracy of the InSAR was affected by microtopography (Marshall et al. 2022)

1.3 Motivations and Research Purpose

This research has the potential to improve understanding of peatland health and therefore policies which will aid carbon sequestration and storage. This has wider societal benefits as degraded peatlands contribute disproportionately to greenhouse gas emissions (Bonn et al. 2014) and healthy peats have the highest storage capacity per unit area of all terrestrial ecosystems (Lunt et al. 2019, Marsden and Ebmeier 2012, Ratcliffe et al. 2018).

The Centre for Doctoral Training themes that this project links to are 'spatial analysis and modelling' and 'visualisation and decision support'. The first through analysing and modelling a peatland environment and the second by comparing outcomes to InSAR data relating to peatland health; this research could be used to support decision making in the Flow Country regarding peatland restoration. This study falls under the EPSRC programme of 'Living with Environmental Change', developing more environmentally harmonious strategies to restore peatland in the Flow Country and links to EPSRC's priorities regarding curiosity-driven discovery and decarbonising society. This research also relates to the 13$^{\text{th}}$ (Climate Action) and 15$^{\text{th}}$ (Life on Land) Sustainable Development Goals and the Geospatial Commission's aim to improve social and environmental outcomes.

The purpose of this research is to determine whether hyperspectral data can be used to improve understanding of PSM data in the Flow Country.

Literature Review

The academic fields of Geography/Space Science (remote sensing journals), Science (conservation and vegetation science journals) and Computer Science (machine learning) provide the main foundation of this research, with methods coming from machine learning mathematical models and GIS-based analysis.

2.1 Remote Sensing Background

Remote sensing has increased over the past few decades, expanding the amount of information available for analysis and used to improve understanding of a range of interrelated social, economic and environmental systems (Chi et al. 2016). As a result of increased Earth observation, the amount of new data has grown to the point where in 2016, 90 % of the data in the world had been generated within the previous two years with the data being more diverse and efficiently gathered than previously (Chi et al. 2016). The data collected has a range of different spatial, temporal and spectral resolutions, giving different insights into the world such as refugee settlement mapping (Quinn et al. 2018), the impact of urbanisation on ecological environments Shao et al. 2020 and Arctic sea ice extent (Kumar et al. 2020).

Remotely sensed data is increasingly being used in nature conservation management of a range of biomes, including peatlands (Kopeć et al. 2020). This includes the mapping of Plant Functional Types (PFTs) which were historically based on the properties of different plants, but increasingly focused on the plants' response to environmental conditions, with LiDAR and spectrometers being key to new observations (Ustin and Gamon 2010). Many datasets are openly available for for ecosystem monitoring, with a range of resolutions. These can be used to classify landscapes according to vegetation type, soil moisture, fire detection and a variety of vegetation indices (Kerr and Ostrovsky 2003).

Specific landscapes such as woodland and farmland have been the focus of various hyperspectral studies. For instance, Spiraea tomentosa (steeplebush) threatens peatland plant communities and coniferous woodlands in Central Europe (Kopeć et al. 2020). Kopeć et al. (2020) focused their study on a Polish forest which is protected as a Natura 2000 site. They were able to demonstrate that the use of hyperspectral data alone could be used to accurately map the invasive species in both summer and autumn when more than 70 % of a polygon contained steeplebush, with better results in autumn when the invasive species could be more easily identified based on reflectance when 30-70 % of the polygon contained the plant. Additionally, Hati et al. (2021) demonstrated that AVIRIS-NG data is more accurate than other data sources (Landsat 8 OLI, Sentinel-2 and Hyperion) when identifying species due to the low sampling interval and high spatial resolution. Whereas a hyperspectral study in northwest Indiana, USA, used hyperspectral data to classify extended morphological profiles for 16 different land uses, predominantly farm-based (Anand et al. 2021). They achieved this using machine learning and tested a range of methods including Support Vector Machines (SVM), random forest and decision trees. They also compared outcomes using confusion matrices to assess the accuracy, finding that random forests produced the most accurate outcomes, with 80 % accuracy, with other accuracies ranging from 70-78 %. With high levels of accuracy, there is the potential to apply these machine learning methods to peatlands such as the Flow Country.

2.2 Remote Sensing and Peatlands

Zhou et al. (2019) used satellite data to demonstrate changes in PSM in abandoned peatlands that have not undergone restoration in Central Kalimantan, Indonesia. They used InSAR data collected between 2006 and 2010 with a spatial resolution of 80 m^2. They demonstrated that 47 % of the peat is subsiding at rates reaching more than 5 cm/yr. This data could then be used to investigate the impact of restoration and assess damages to peatlands due to drainage and fires (Zhou et al. 2019). Whereas, Zhou et al. (2013) found that Sumatran peatland height can decrease by 15 cm in areas with palm oil plantations, demonstrating the extent to which land use and human activity impacts PSM. Alternatively, a range of studies have used bioclimatic envelope models to predict future distributions of blanket bog globally (Gallego-Sala and Prentice 2013) and in Scotland (Ferretto et al. 2021). These studies used climatic data from Climate 2.2 (Gallego-Sala and Prentice 2013) or the Met Office (Ferretto et al. 2021) with both outcomes anticipating reductions in the size of current blanket bogs as temperatures increase (especially in the summer) and precipitation reduces. Similar outcomes were found in studies using eddy covariance analysis. Artz et al. (2021) used eddy covariance to demonstrate that periods with drought and higher temperatures double the rate of carbon loss in blanket bogs in the Cairngorms, especially in areas with eroded peat as they are less resilient to change. The current carbon flux can also be compared to past carbon measurements collected using peat cores (Lunt et al. 2019). Lunt et al. (2019) demonstrated that there has been a slight reduction in carbon sequestration in a blanket bog in southwest England by about 2 $tCO_2eq/ha/yr$ since 1850. They concluded that many studies may underestimate the sequestration potential for peatlands, and they may be more resilient than thought by many (Lunt et al. 2019). Other peatland studies have focused more on restoration (Humpenöder et al. 2020, Bonn et al. 2014, Parry et al. 2014) and the impact of burning (Garnett et al. 2000, Whitehead et al. 2021).

Hyperspectral data collected remotely can be used to monitor and analyse peatlands in a non-destructive manner (J. M. McMorrow et al. 2004). Hyperspectral data collected using Unmanned Aerial Vehicles (UAVs) over a Finnish peatland has the potential to significantly improve the efficiency of peat production (Honkavaara et al. 2016). This is because Honkavaara et al. (2016) were able to estimate moisture content more accurately when using the hyperspectral data than when using other data types. The use of miniaturised hyperspectral imagery sensors on small UAVs enabled enhanced processing and interpretation potential when compared to traditional pushbroom scanners (Honkavaara et al. 2016). This was due to the development of 3-dimensional geometry and multiple object views with the reflectance signatures. Once the images were collected, lab calibration corrections were applied to the images and reflectance data generated from transformed digital numbers to form hyperspectral image mosaics and the estimation of surface moisture. Whereas other Finnish hyperspectral studies collected field data from transects of peat cores, which were then analysed in the laboratory to classify peat and estimate humification (Granlund, Keinänen, et al. 2021, Granlund, Vesakoski, et al. 2021). In tropical peat swamp forests, hyperspectral data can detect changes in the biochemical and biophysical characteristics which impact tree crown segmentation (Nordin et al. 2019, Tochon et al. 2015). This is better achieved with hyperspectral data than other data types due to the high spatial and spectral resolution, enabling the identification of specific species despite the dense heterogeneous woodland. However, the addition of other data sources, such as LiDAR, were found to improve accuracy further (Nordin et al. 2019), demonstrating the benefits of including more than one data type in analysis.

HyMap images (from the SHAC) were used by J. M. McMorrow et al. (2004) to test candidate indices of peat composition for eroded blanket peat in the southern Pennines. Analysis used 35 sites of field data, lab work regarding peat properties in addition to the HyMap data. They found strong correlations between the Shortwave Infrared (SWIR) reflectance and transmission, along with strong positive correlations between cellulose absorption and transmission, ultimately suggesting that hyperspectral data has the potential to provide information on the composition of peatland surfaces across large areas. An alternative approach to map eroded peatland was used by Carless et al. (2019) where a combination of data were used, composed of airborne LiDAR, CASI (visible and Near Infrared (NIR) hyperspectral data), and aerial images, to identify and quantify areas with peatland degradation in Dartmoor National Park. They determined that using an amalgamation of data increased outcome accuracy in addition to being robust and cost effective (Carless et al. 2019). Their analysis was 94 % and 87 % accurate when identifying peat drains and peat cuttings respectively, however, much lower when identifying bare peat by digitised data compared to ground data (Carless et al. 2019).

Cole et al. (2014) focused on a range of species in functional groups (graminoids, bryophytes and shrubs), collecting data monthly from spring to autumn in 2009 and 2010. They assessed the spectra and a range of vegetation indicies to analyse changes over time, finding that the vegetation was most spectrally separable in July (when chlorophyll pigments and leaf area index are highest), but optimal time varies depending on the species. As a result of this Cole et al. (2014) concluded that there is not one suitable optimal recommended temporal window for monitoring the peatland. This study did, however, improve their understanding of the phenological cycle of peatland species. Cole et al. (2013) also used hyperspectral data in the Peak District, however, they focused on PFTs in areas undergoing restoration. They used SPSS software to extract information from the hyperspectral data (vegetation indices) and the cover type (PFTs) and determined that partial least squares regression models could be used to map PFTs over large areas of peatland to judge the effectiveness of restoration, but only with high spatial resolution data (0.7 m pixels). Other studies using hyperspectral data have focused on different attributes of peatlands (Table 2.1).

TABLE 2.1: Alternative uses of hyperspectral data to analyse peatlands

Peatland attribute(s)	Project description	Location	Reference
Hydrology and flood risk	Discusses a range of restoration projects aiming to keep more moisture on the bog	Kinder Scout, Peak District	Shuttleworth et al 2018
Water table depth and ecosystem exchange	Assesses the ability of satellite and airborne data to estimate peatland water table depth and ecosystem exchange using SWIR hyperspectral data over time.	Mer Bleue, Ottawa, Canada	Kalacska et al 2018

2.3 Flow Country Review

Various studies have been conducted in the Flow Country, however, there has been little hyperspectral analysis in this area.

One key method used in recent years to assess carbon accumulation rates is eddy covariance (Hambley et al. 2018, Levy and Gray 2015, Ratcliffe et al. 2018, Ratcliffe 2015). This technique uses peat cores collected in the field in conjunction with multi-sensor scanners which combines x-ray fluorescence and x-radiography to precisely date the peat in the cores at a spatial resolution of 0.2-2 mm (Ratcliffe et al. 2018). Outcomes of eddy covariance analysis have varied with Ratcliffe et al. (2018) estimating long-term accumulation rates of 15.4 $gC/m^2/yr$ for a site in which the Levy and Gray (2015) study suggested carbon was accumulating in the same place at a rate of 99.37 $gC/m^2/yr$. However, other sites were found to give comparable results (Ratcliffe et al. 2018), suggesting that additional research is required at the conflicting site. Other key findings are in line with other research linked to the Flow Country (and other peatlands) regarding carbon exchange, suggesting that the current carbon sink strength will change over time due to changes in ecological drivers, such as fires, and feedback loops linked to changes in ecology and hydrology (Ratcliffe et al. 2018). Hambley et al. (2018) increased the temporal resolution of eddy covariance data to assess seasonal variations with other studies missing spring 'greening' and autumn senescence, which increase rates of microbial activity and therefore carbon output. Similar to the other studies they demonstrated that the Flow Country is a net carbon sink, regarding CO_2, however, demonstrated that methane understanding needs to be improved, suggesting that emissions could increase following restoration. This could reduce the accuracy of carbon accumulation rates discussed in other eddy covariance studies including Levy and Gray (2015) and Ratcliffe et al. (2018), although Levy and Gray did recognise that methane emissions equate to the majority of non-CO_2 carbon losses.

An alternative method to measure carbon stocks uses a combination of large-scale remote sensing and spatial covariates, including topography and climate, rather than local scale field data (Aitkenhead and Coull 2020). This enabled mapping of carbon stores across the whole of Scotland and suggested peat covered a greater area than previously thought, linked to soil depth. This has potential to be used in conjunction with eddy covariance, but would require higher resolution data; spatially (the 100 m^2 pixels used could miss local variations; Aitkenhead and Coull 2020) and temporally.

Other studies in the Flow Country have focused on water chemistry, especially dissolved organic carbon (DOC; Muller and Tankéré-Muller 2012, Vinjili 2012). Muller and Tankéré-Muller (2012) focused on the impact of felling and seasonal changes on DOC and metals in streams within the catchment of the River Thurso. Whereas Vinjili (2012) focused on the River Dyke catchment assessing the impact of afforestation in addition to felling for restoration. Muller and Tankéré-Muller (2012) found that seasonal changes impacted all sites (near recently felled areas as well as more natural sites), with felled areas only affecting dissolved aluminium and manganese levels with high levels of leaching. The limited changes in DOC, iron and potassium were associated with a buffer zone between the felled forest and the water (Muller and Tankéré-Muller 2012). Similarly, Linden et al. (2015) found that climatic changes were much more important factors in DOC concentrations than land use change.

2.3.1 PSM Analysis

The InSAR studies have demonstrated changes in PSM over time (Alshammari et al. 2018, Alshammari et al. 2020, Bradley et al. 2021, Marshall et al. 2022). The PSM is likely to be underestimated (1-2 mm/yr), especially in the most dynamic parts of the peatland; however, if these areas are under drought conditions, underestimation can increase to 15-42 mm/yr (Marshall et al. 2022). As the least dynamic parts of the ecosystem are dominated by shrubs (dynamism decreases with height above the water table) and the more dynamic dominated by *Sphagnum* (Marshall et al. 2022), it is worth assessing the rate of transition between these communities. PSM has been shown to be a sensitive indicator of peatland function at a large scale due to a range of factors including mechanical deformation, vegetation composition, water level changes and land management, meaning it has the potential to give a holistic view of the landscape (Marshall et al. 2021). Surface motion is especially affected by accumulation which causes growth of peat, drainage, compression and decay of organic matter which cause irreversible subsidence and seasonal variation in water and gas storage which cause reversible peat deformation (Alshammari et al. 2020). These changes in PSM over time are directly related to precipitation rate, water level and vegetation composition (Marshall et al. 2021). Measurements need to be taken over long timescales to determine whether peat is accumulating, indicative of healthy, wetter, peat which sequesters carbon, or subsiding, characteristic of unhealthy, drier peat with carbon loss (Alshammari et al. 2020, Bradley et al. 2021).

Monitoring of a range of peatlands undergoing restoration enables improved understanding of the impact of re-wetting strategies and expectations regarding the potential for peatland stability and carbon absorption (Bradley et al. 2021). It also improves estimations of the distribution of peatland and their carbon inventories, potentially on a national scale, whilst helping to identify areas that are at risk of becoming unstable, eroded or affected by fire (Marshall et al. 2021). This can only be achieved by analysing PSM data over long time-scales, with analysis of 600 management blocks demonstrating a clear restoration trajectory over the past 20 years, especially since 2010, with changes in restoration techniques causing sites to recover up to eight years more quickly than earlier sites (Marshall et al. 2021). However, areas of woodland that have been felled recently record the highest rates of subsidence as they undergo a period of instability and change (Bradley et al. 2021, Marshall et al. 2021), with them becoming more stable over time, reaching equilibrium (Marshall et al. 2021).

Peatland monitoring has also lead to improved understanding of peatland dynamics throughout the year, with two maxima observed annually. The first occurs in areas with steeper gradients, at peatland margins, in areas with degraded peat and in communities dominated with shrubs (Bradley et al. 2021). Whereas the second maximum is associated with flatter areas with pool systems and *Sphagnum* dominated communities (Bradley et al. 2021). These maxima occur between August and November, whereas maximum subsidence occurs between April and June (Alshammari et al. 2018). The amplitude of the swelling and the average annual motion was strongly linked to which species dominated the landscape (Bradley et al. 2021).

2.3.2 Hyperspectral Analysis in the Flow Country

There has been very limited analysis in the Flow Country using hyperspectral imagery. Lees (2019) used a handheld SVC HR-1024 spectroradiometer to assess the relationship between spectral indices and the moisture content of Gross Primary Productivity (GPP) of the blanket bog vegetation species (especially) *Sphagnum* species under a range of conditions. They used the hyperspectral data to find the relationship between a range of vegetation indices, including Normalised Difference vegetation Index (NDVI) and Photochemical Reflectance Index (PRI) and GPP, and water indices and moisture content, in both in the field and in the laboratory. They found that all vegetation indices had significant relationships both in the laboratory and field, but the water indices only demonstrated significant relationships in the laboratory.

2.4 Machine Learning

2.4.1 Supervised Machine Learning

Supervised machine learning algorithms are used to predict patterns by categorising data from inputted information (Singh et al. 2016). All data is required to be split into training and testing data. Each class is determined from the combination of features and seeking patterns common to each class in the training data and validated using the testing data (Singh et al. 2016). Cross validation can be used to compare the accuracies of different machine learning outputs, however, outcomes can be misleading; overestimating the applicability of the model (Granlund, Keinänen, et al. 2021).

Decision tree classifiers are non-parametric and build a tree until the inputted threshold is reached, preventing any more splitting of leaf nodes (Pal 2005). They are easy to understand, flexible, robust to noise and require little data cleansing, however, they are prone to overfitting and can be unstable with small data variations producing very different trees (Pal 2005, Provost and Fawcett 2013). They can also be biased towards the more dominant classes (Provost and Fawcett 2013). Additionally, they can find it difficult to handle data with many dimensions (Pal 2005). Pradhan et al. (2014) found the decision tree classified land use from remotely sensed data more successfully than more conventional prediction methods such as the maximum likelihood classification, whereas the decision tree used in the Anand et al. (2021) study had the lowest accuracy of the supervised learning classifications applied.

Random forest classifiers are a combination of tree classifiers with each classifier generated using a random vector sampled independently from the input vector (Pal 2005, Provost and Fawcett 2013). Each tree ultimately casts a vote for the most popular class to determine the assigned class (Pal 2005). They are robust to outliers and scalable, however take time to build and can miss important interactions between features (Mather and Tso 2009, Provost and Fawcett 2013). The development of random forests removes the overfitting and bias issues of decision trees due to the combination approach (Provost and Fawcett 2013). Random forest classification was found to be the most accurate method (80 %) in the Anand et al. (2021) study based on farmland classification in the USA, with a similar accuracy of 83 % achieved in the Poland study (Kopeć et al. 2020). Whereas in the Erudel et al. (2017) study, accuracies ranged significantly from 65.10-81.60 % (when classifying vegetation types using different indices), with the highest accuracies when the training size was greater than or equal to 40 %. Generally, in the Erudel et al. (2017) study, the random forest outcomes were weakest, which is surprising when compared to other studies.

Support vector machines aim to determine the location of decision boundaries that produce optimal class separation (Provost and Fawcett 2013). They look to maximise the minimum distance from the hyperplane to the nearest sample point enabling them to be robust to highly dimensional data (Provost and Fawcett 2013). They can also deal with overfitting issues and minimise classification errors (Mather and Tso 2009). The key issues with SVMs are slow training speed and performance is heavily dependent on parameter choice (Provost and Fawcett 2013). The Anand et al. (2021) study found that SVMs were almost as accurate as random forests (accuracy of 78 %), whereas Erudel et al. (2017) generally found that SVM outcomes were better than random forest accuracies for both the linear and Radial Basis Function (RBF) kernels, with both kernel outcomes exceeding the random forest in 67 % of cases and one of them exceeding the random forest accuracy in the remaining 33 %. Whereas, the SVM accuracies were about 50 % in the more complex classification involving humification in the Granlund, Keinänen, et al. (2021) study, but higher when classifying peat types. This suggests that all three need to be assessed for the classification of PFTs in the Flow Country.

Logistic regression is a statistical model which fits a logistic curve to the dataset and can be updated easily with new data (Singh et al. 2016). Additionally, they are quick and the outputs have a probabilistic interpretation, they can be regularised to reduce the potential for overfitting (Provost and Fawcett 2013). However, they are often too simple and are unable to handle multinomial problems (Provost and Fawcett 2013). As a result, Erudel et al. (2017) applied a regularised logistic regression to the species classification with accuracies ranging from 69.42 to 83.55 % depending on the training size and data preparation prior to analysis.

2.4.2 Unsupervised Machine Learning

K-means clustering is very robust, fast and flexible, however, as it always searches for globular clusters can form poor clusters when they are not spherical (Provost and Fawcett 2013). Additionally, ‘k’ must be specified in advance, which can often be difficult to know (Provost and Fawcett 2013). This clustering technique was used in relation to hyperspectral data analysis of peat cores in the Granlund, Keinänen, et al. (2021) study, using eight clusters to analyse stratigraphic patterns of peat cores. These clusters were generally linked to vegetation type and level of humification, suggesting that k-means clusters could be applied to data to find PFT classes.

Principal Component Analysis (PCA) analysis spectrally reduces the number of bands in a dataset with the aim to extract the most useful information from an image (Nordin et al. 2019). This removes noise from the dataset, however, can lead to the removal of relevant information from the image (Tochon et al. 2015). As well as being used to reduce hyperspectral data, it can also be used to suggest whether predictions made using the dataset will be strong (Patil and Dwivedi 2021). It can therefore be applied to the train-test dataset to assess the strength of predictions made using the hyperspectral data to determine PFTs.

2.5 Gaps in the Literature

The research in this study fills two key research gaps. Firstly, the analysis of air borne hyperspectral data in the Flow Country as previously only handheld spectrometers have been used (e.g. Lees 2019). This is primarily due to the lack of available data, with the hyperspectral data in northeast Scotland only being collected in July 2021.

The second gap is the use of hyperspectral data in conjunction with InSAR PSM data to add more details to the data outcomes. Additional understanding should come from higher spatial resolution and the large number of bands covering a larger part of the spectrum. However, this difference in spatial resolution, and the lack of temporal hyperspectral data may affect the potential of hyperspectral analysis to add to the understanding of PSM outcomes.

Aim and objectives

3.1 Aim

To determine whether hyperspectral data can be used to understand the association between peatland surface motion (as measured by InSAR data) and land cover in the Flow Country.

3.2 Objectives

1. Assess the extent to which supervised and unsupervised machine learning algorithms can be used to classify plant functional types.
2. Determine whether machine learning can be used to show a relationship between plant functional types and peat surface motion.

Methodology

4.1 Data Sources and Software

The hyperspectral sensor used in this study is Airborne Visible/Near-Infrared Imaging Spectrometer-Next Generation (AVIRIS-NG) which was developed by NASA/JPL and has a spectral range of 0.38-2.52 µm, up to 430 bands and a spectral sampling interval of 5 nm (Jia et al. 2020). AVIRIS-NG instruments utilise the pushbroom imaging mode outlined in Figure 4.1. As there is no mechanical scanning mechanism, the weight and volume are lower than the alternative whiskbroom technique and the signal-to-noise ratio is also better (Jia et al. 2020). Whereas, the main disadvantage of this image is the challenge of balancing the field of view and instantaneous field of view, although this can be compensated for by detector technology (Jia et al. 2020).

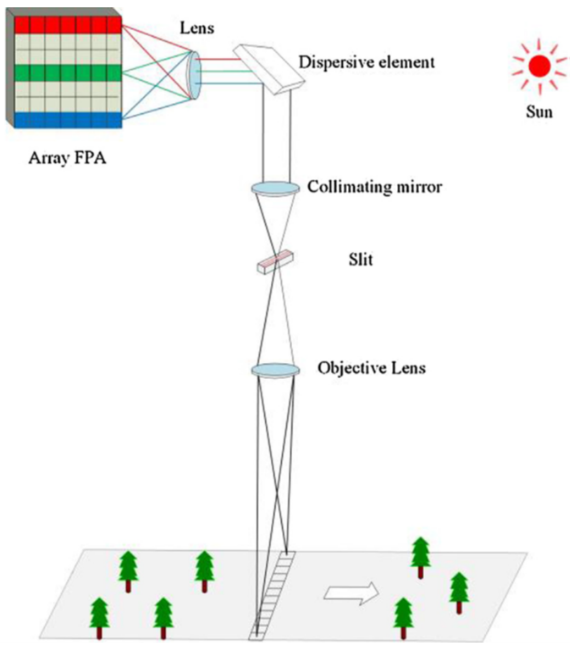


FIGURE 4.1: Pushbroom imaging mode, where FPA stands for focal plan array Jia et al. 2020

Hyperspectral data has the potential to improve understanding of landscapes due to the level of detail that can be acquired from images with many wavebands. The AVIRIS-NG data was collected 15th July 2021 during a collaborative mission conducted by ESA, NASA/JPL and UZH; the Flow Country was one of 17 flights conducted over sites across western Europe (Figure 4.2) between May and July 2021

(ARES 2021). This data ranged from 377-2500 nm with each pixel having a spatial resolution of 5x5 m for three of the sites, but slightly over 5 m for the erosion site. This resolution is finer than that used in studies using InSAR PSM data (80x90 m) in the Flow Country (Marshall et al. 2022 Alshammari et al. 2018 Bradley et al. 2021 Marshall et al. 2021) meaning that there is potential to not only add to their findings, but also add detail to current knowledge about this blanket bog.

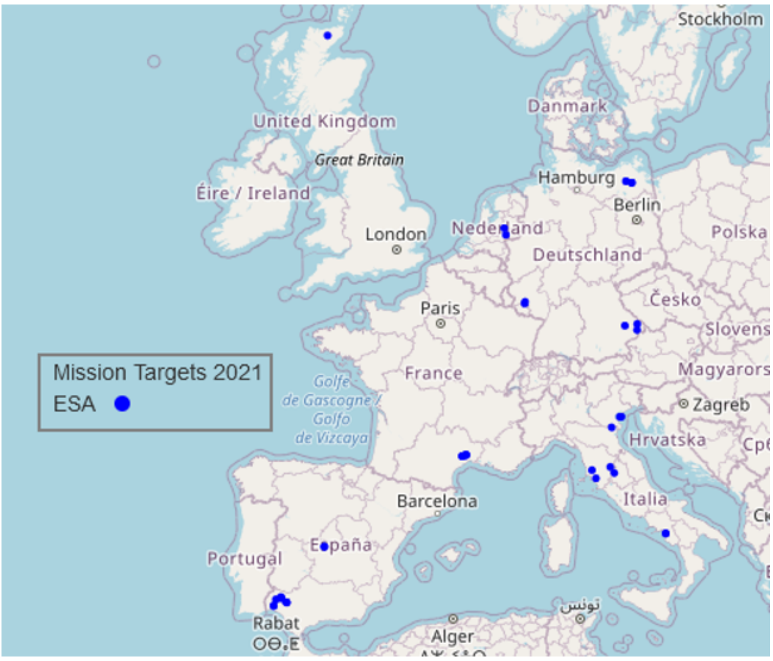


FIGURE 4.2: The locations of the AVIRIS-NG Europe 2021 campaign sites (ARES 2021)

The two datasets used in this study are summarised in Table 4.1, with initial data investigation conducted in QGIS, using the EnMAP Box and AVHYAS plugins which can visualise hyperspectral data. All analysis was undertaken in Python (https://github.com/rachelzwalker/Flow_Country_HSI_and_PSM), with some data transformations and data joins based on location carried out in GIS. The majority of GIS work was completed in QGIS as this is an open-source software, however, some processes, such as matching different data projections were more accurate in ArcMap. Results were visualised in Python, QGIS and Microsoft Excel.

TABLE 4.1: A summary of the hyperspectral and InSAR data

Data Type	Spatial Resolution	Temporal Range	Data Source	Data Collection	Data Description
Hyperspectral	5 x 5 m	Collected in one day (15/07/2021)	https://ares-observatory.ch/esa_chime_mission_2021/	JPL/NASA	AVIRIR-NG data containing 425 bands
InSAR	80 x 90 m	Timeseries data from 2015-2019	https://catalogue.ceh.ac.uk/documents/7c2778bf-b498-4ba2-b8cb-60a2081e5ba7	Terra Motion	PSM vector data

4.2 Site Selection

The overall study area was determined by the NASA/JPL flight paths predominantly covering the Forsinard Flows Nature Reserve (Figure 4.3); InSAR data had already been processed for this area, but only analysed in localised areas.

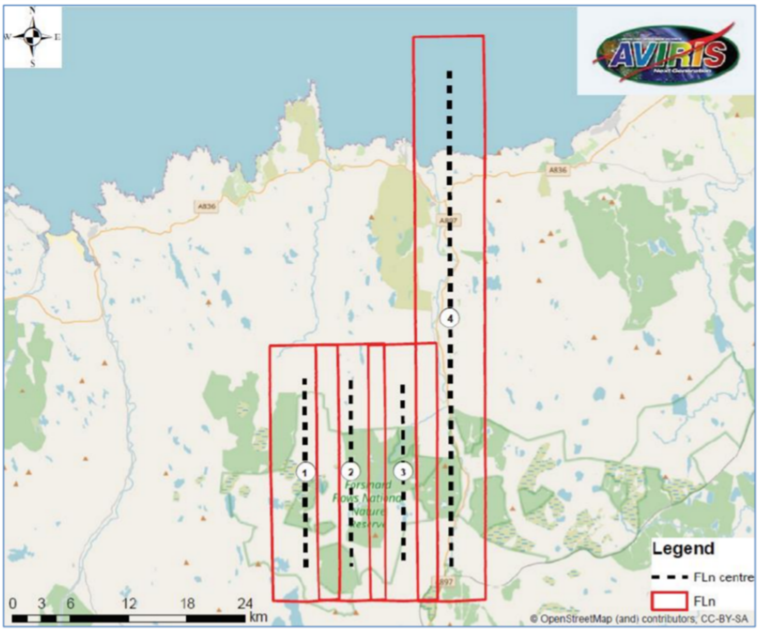


FIGURE 4.3: The flight paths used to collect the hyperspectral data, July 2021 ((NASA JPL 2022)

Specific sites were identified based on their characteristics (Figure 4.4). Due to the scope and scale of this study, four sites were studied, as these areas cover a range of environments which have been focal points of previous PSM studies. However, had the scope been larger, additional sites with active peat cutting and active drains could also have been analysed. Cross Lochs was one of the locations studied in detail by Alshammari et al. (2020) and Bradley et al. (2021), and was used as the control site and area of initial exploration due to its near-natural characteristics and as a result had fewer plant functional types. Sites undergoing restoration become more stable over time, progressing faster where restoration began after 2010 (Marshall et al. 2021), therefore two restoration sites were analysed to determine the temporal impact of restoration on peat surface motion. The first restoration site was being converted in the 2000s, with the second restoration site conversion occurring a decade later. Trees and timber were removed and the main collector drains blocked; some brash was removed, however, some was left which can been seen as white lines on the satellite images. Furrows were then ploughed to reduce microtopography and flatten the site, and the peat was then left to recover. The erosion site was affected by peat cutting in the past and still contains some vegetation, such as Sitka, pine and agricultural grasses, which have been planted by humans to replace the native species, and are only separated from the rest of the peatland by a wire fence. As a result, each site will have different vegetation assemblages which interrelate with the PSM measures of peat condition and link directly to this study's aim. Each site consists of an area approximately 1 km^2 as InSAR measurement sites were this size (Marshall et al. 2022).

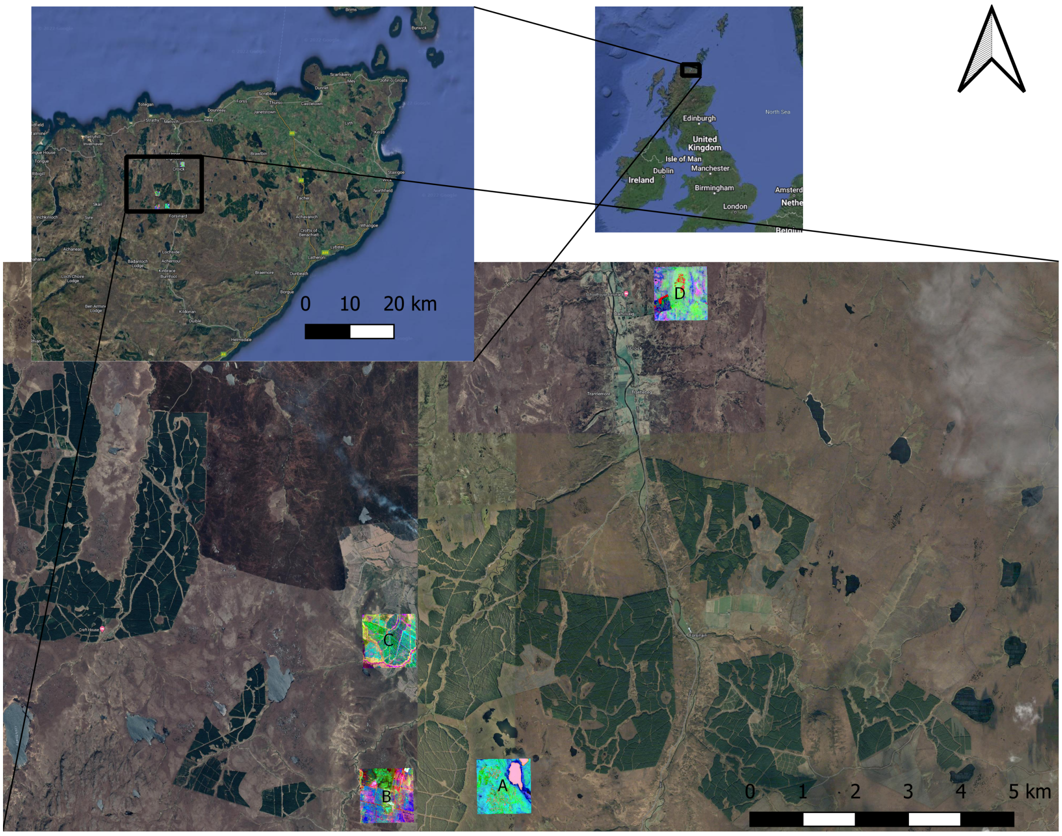


FIGURE 4.4: Four 1km^2 site locations within the Flow Country. Cross Lochs (A) is a near-natural site (control site) with central coordinates of -3.937388, 58.386329, there are two restoration sites (B and C) at different stages of transition from plantation to peat with central coordinates of -3.975073, 58.385097 and -3.973670, 58.411235, and an erosion site (D) with area of bare peat and cuttings with central coordinates of -3.879604,58.469920.

4.3 Data Preparation for Machine Learning

4.3.1 Pre-processing

The hyperspectral data was already radiometrically and geometrically corrected (Priyadarshini et al. 2019) using the World Geodetic System 84, projection, however, the erosion site required further geometric correction (when the data was compared to Google satellite imagery). which was undertaken in GIS. Although the data had undergone atmospheric corrections (NASA 2022), removal of bands associated with atmospheric water vapour are encouraged (Priyadarshini et al. 2019, Erudel et al. 2017), therefore the bands with wavelengths 1350-1450 nm, 1810-1940 nm and 2400-2500 nm were removed. Additional pre-processing, such as the application of the Savitzky-Golay filter, could have been undertaken (Erudel et al. 2017), however, when attempted the outcomes were not as usable as the data without the filter. The InSAR data was already pre-processed when received for use in this study with only small changes to the attribute table required.

4.3.2 Data Transformations

Different transformations (first derivative, second derivative and continuum removal) are linked to the absorption features of vegetation, and these have the potential to differ between PFTs (Erudel et al. 2017):

$$\text{first derivative} = \frac{\rho\lambda j - \rho\lambda i}{\Delta\lambda} \ . \tag{4.1}$$

$$\text{second derivative} = \frac{\rho\lambda j - 2\rho\lambda i + \rho\lambda k}{\Delta\lambda^2} \ . \tag{4.2}$$

$$\text{continuum removal} = \frac{\rho\lambda}{C\lambda} \ . \tag{4.3}$$

As a result, they may be able to predict the PFTs to a higher degree of accuracy than the original dataset. Other potential transformations to be used include the brightness-normalised spectral signature and continuum removal derivative reflectance (Erudel et al. 2017), but the scope of this study did not include these transformations due to the complexity of these signatures and performance of the normalised signature.

4.3.3 Spectral Ranges

It may not be necessary to use the whole dataset to accurately predict the different PFT classes, which would save processing power and time, especially if a larger sample of the original hyperspectral data were to be analysed. Therefore, spectral sub-ranges were analysed in addition to the full dataset (Table 4.2, Figure 4.5).

TABLE 4.2: The full sepctral range and five sub-ranges of the electromagnetic spectrum used to analyse the reflectance of green vegetation (Erudel et al. 2017)

Wavelength Range (nm)	Bands	Spectral Range	Spectral Reflectance of Vegetation	References
377-2500	0-358	Full	Includes all of the of the wavelengths included in the dataset, with variations in absorption and transmission throughout.	
400-700	6-65	Visible	Chlorophyll and carotene (a biological pigment) absorptions cause low reflectance and transmittance of radiation (Section one of Figure 8).	Salisbury and Ross 1992 Erudel et al. 2017
680-750	62-75	Red-edge	Plant biochemical and physical parameters are correlated with reflectance (Red edge in Figure 8)	Mutanga and Skidmore 2007 Erudel et al. 2017
700-1300	66-185	NIR	Scattering of photons in the spongy mesophyll in the leaves causes high reflectance and transmittance, with low absorption (Second main section in Figure 8)	Woolley 1971 Erudel et al. 2017
1300-2500	186-358	SWIR	Water content of vegetation causes strong absorption of radiation, with some absorption by lignin and other biochemical components. There is low reflectance. (Final section of Figure 8)	Woolley 1971 Erudel et al. 2017

4.4 Machine Learning for PFT Prediction

The machine learning analysis included in the investigation is summarised in Figure 4.6.

4.4.1 Supervised Learning

Due to the lack of spectral libraries for peatland vegetation species and the dynamic nature of the signatures depending on the conditions, as Harris et al. (2005) demonstrated with changes in water stress, fieldwork was planned. However, permissions could not be attained until the latter stages of the project due to the time of year (breeding season). Therefore, expert knowledge was used to develop site descriptions and identify potential vegetation assemblages in the landscape from Google satellite imagery. This enabled the development of a spectral library of PFTs for Cross Lochs (this was the site where the PFTs were easiest to distinguish without fieldwork, using only satellite images). This resulted in six classes for the Cross

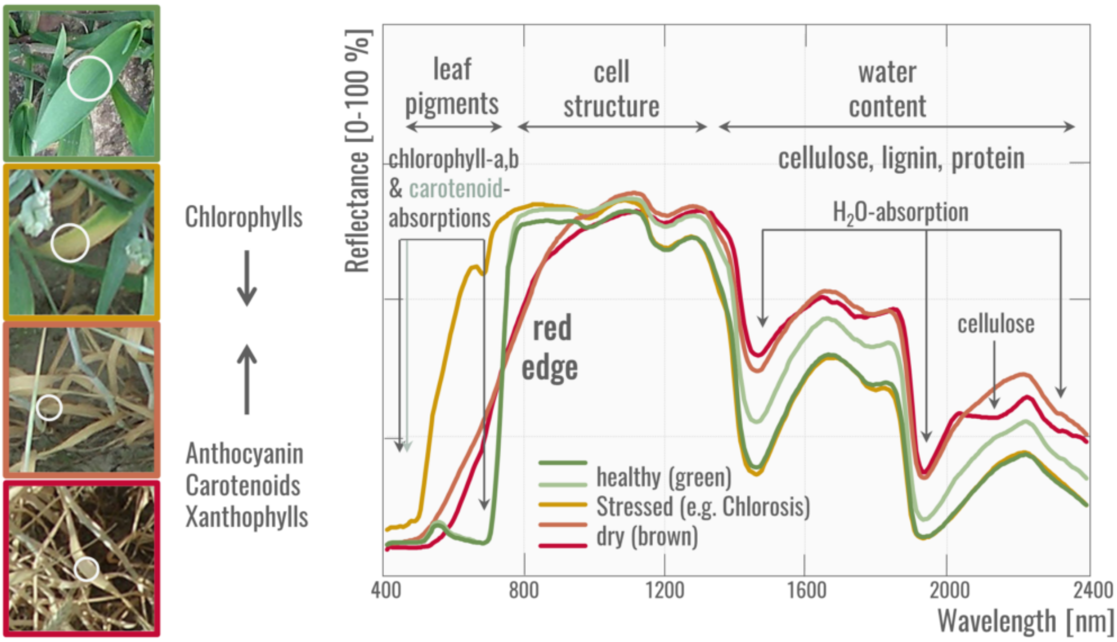


FIGURE 4.5: Reflectance spectra of green, stressed and drying vegetation (EO College 2021)

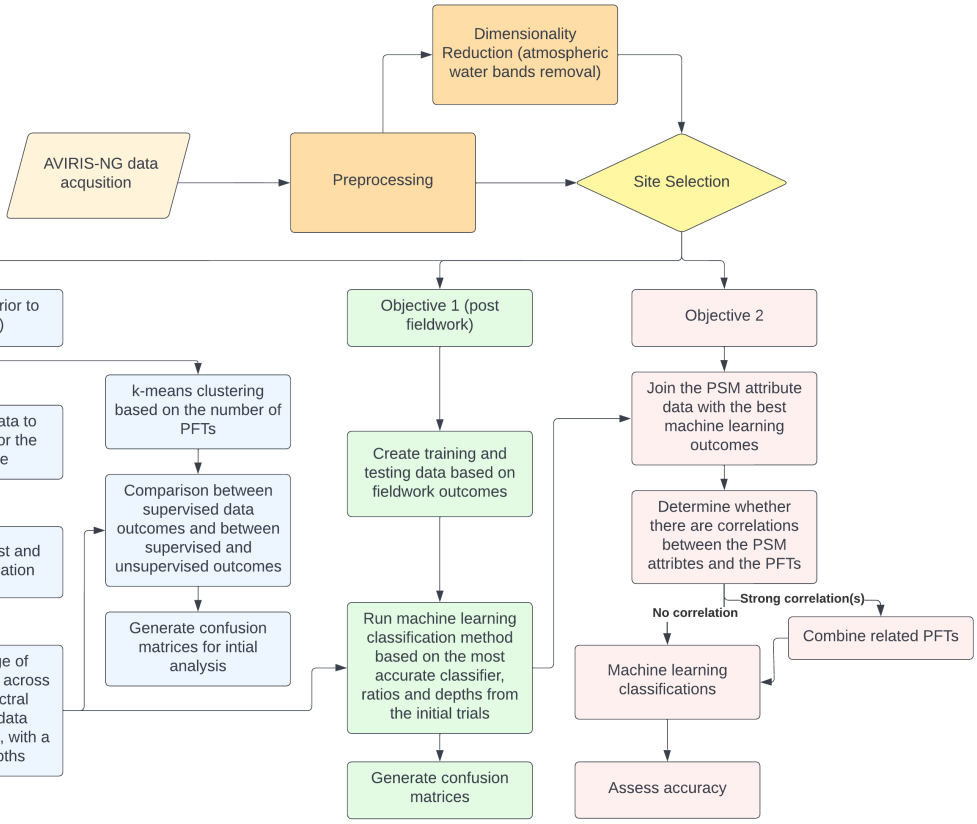


FIGURE 4.6: Summary of machine learning methods used for each objective

Lochs area (as shown in the spectral library for Cross Lochs Figure 4.7 with equation 4 (Mather and Tso 2009) determining that 98 samples of each PFT should be recorded to generate the training and testing data for the supervised learning analysis.

$$n = \frac{B\Pi(i)(1\check{}\Pi(i))}{b(i)^2} \quad . \tag{4.4}$$

Small sample sizes meant that only 588 of the 42,021 pixels needed allocating to different PFTs, saving time and processing power when training the data. As fieldwork could not be conducted until less than a month before the end of the project, initial analysis was conducted for Cross Lochs to determine which machine learning techniques were better at predicting PFTs. The plant-based PFTs were focused on the dominant species and there were two water-based groups, 'water' and 'pool bogbean'. 'Water' is open water with a greater depth than 'pool bogbean', which is also contains bogbean (Menyanthes trifoliata), changing the spectral signature. Prior to running the machine learning analysis, a PCA test was performed to judge the potential strength of predictions (Figure 4.8). Some PFTs, such as the 'calluna mix' had a clear cluster, however, others, such as 'water' did not, suggesting that predictions could be made, but their strength would vary.

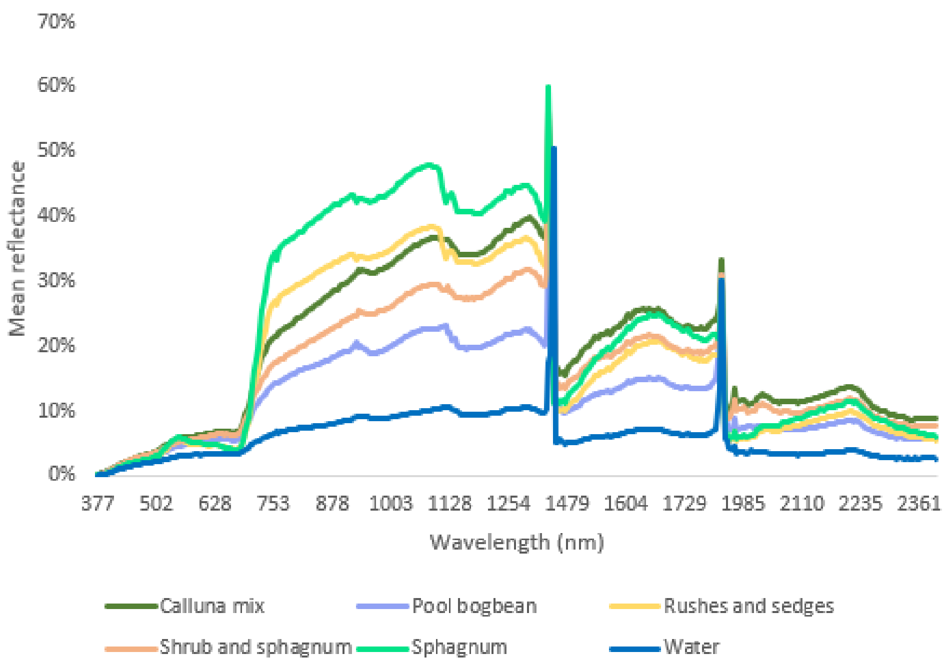


FIGURE 4.7: Spectral library for Cross Lochs developed from expert descriptors and satellite imagery to generate mean spectral signatures for each PFT

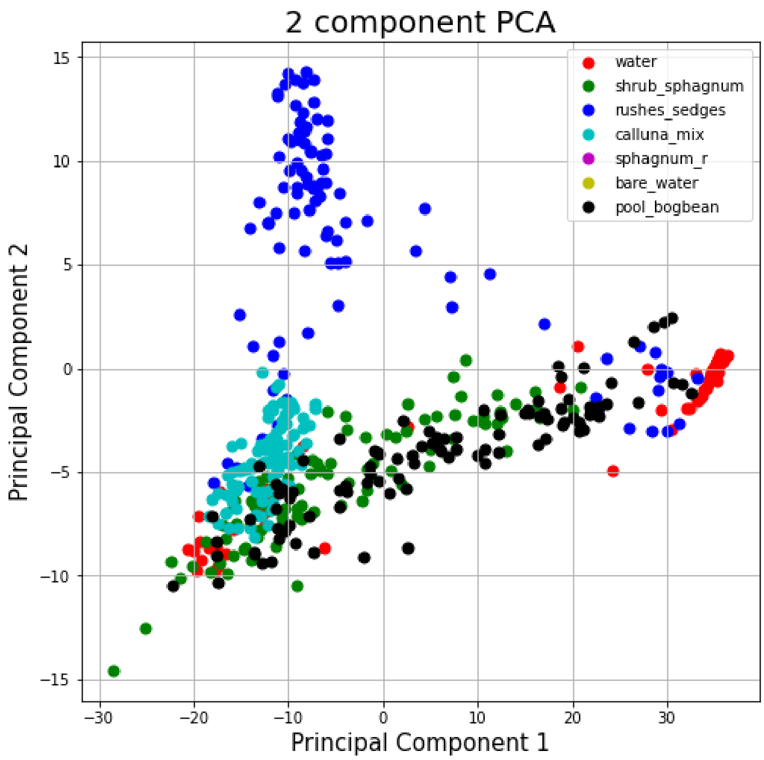


FIGURE 4.8: Scatter plot for the two dominant PCA bands regarding the train-test data for the exploratory analysis of Cross Lochs

The Scikit-learn Python package was used to perform Random Forest, SVM and k-means clustering analysis on the data prior to comparison. Other machine learning algorithms could also have been used, such as the partial least squares regression (used by Cole et al. 2013), however, due to the coarser nature of the hyperspectral data used in this study, it was not included. The Random Forest depths tested ranged from 2-5 to prevent over- and under-fitting of data, with those with the highest accuracy used in the comparative analysis. A range of training-testing ratios were also used to determine which gave the best performance for the data. This was conducted for each of the data transformations and spectral ranges. Each supervised learning outcome included classification accuracy and standard deviation, with those with the highest accuracy compared to the k-means cluster analysis. The outputs from the Random Forest classifier, SVM and k-mean cluster were PFT maps illustrating the potential distribution of the different vegetation groups (key outputs in Appendix D. Confusion matrices with overall, producer's and user's accuracies were outputted to assess the difference between the outcomes. These demonstrated that random forest analysis was more successful than both SVMs at predicting PFTs, with a tree depth of five and training size of 70-75 % as the most effective (Figure 4.9). This code was then combined in PyCharm to improve efficiency and repeatability, with map outcomes automatically generated for the best predictions.

FIGURE 4.9: Supervised learning PFT identification accuracies for exploratory analysis of Cross Lochs. The machine learning classifications were run for a range of train-test ratio, with better outcomes for the random forest classifier than SVM classification

The first restoration site interlinks most with the other sites so has the potential to be used to predict most of the PFTs across the locations. Immediately following fieldwork, the annotated satellite images (especially the first restored site) were added to (Figure 4.10) and a spectral library developed for this site (Figure 4.11). The fieldwork and satellite images were then used to create the PFT training and testing data, which were added to the attribute table in GIS. This was then combined to train all four sites and predict the PFT clusters, however, analysis was also undertaken using the first restoration site data alone to predict for this site as the training data for this site was the most accurate as most of the fieldwork was carried out there. These were then inputted with the various data transformation and sub-samples into PyCharm for analysis. The number of points required for the different PFTS was determined using the fourth equation, resulting in 1,117 training and testing points compared to more than 160,000 pixels in total. The majority of the PFTs had 90 sets of training and testing data, however, the agricultural grasses did not visibly cover a large area so only had 55 points and the rushes and sedges were more sporadic and difficult to identify from the satellite image so only had 67 samples from the fieldwork site. 51 % of the training data was from the first restoration site (where most of the field data was collected) and 24 % from the erosion site (the rest of the field data). However, additional points were required from the Cross Lochs site (13 %) for the water and pool-bogbean PFTs as these were much more frequent in this landscape, similarly with the second restoration site (12 %) and the brash and bare peat PFTs.

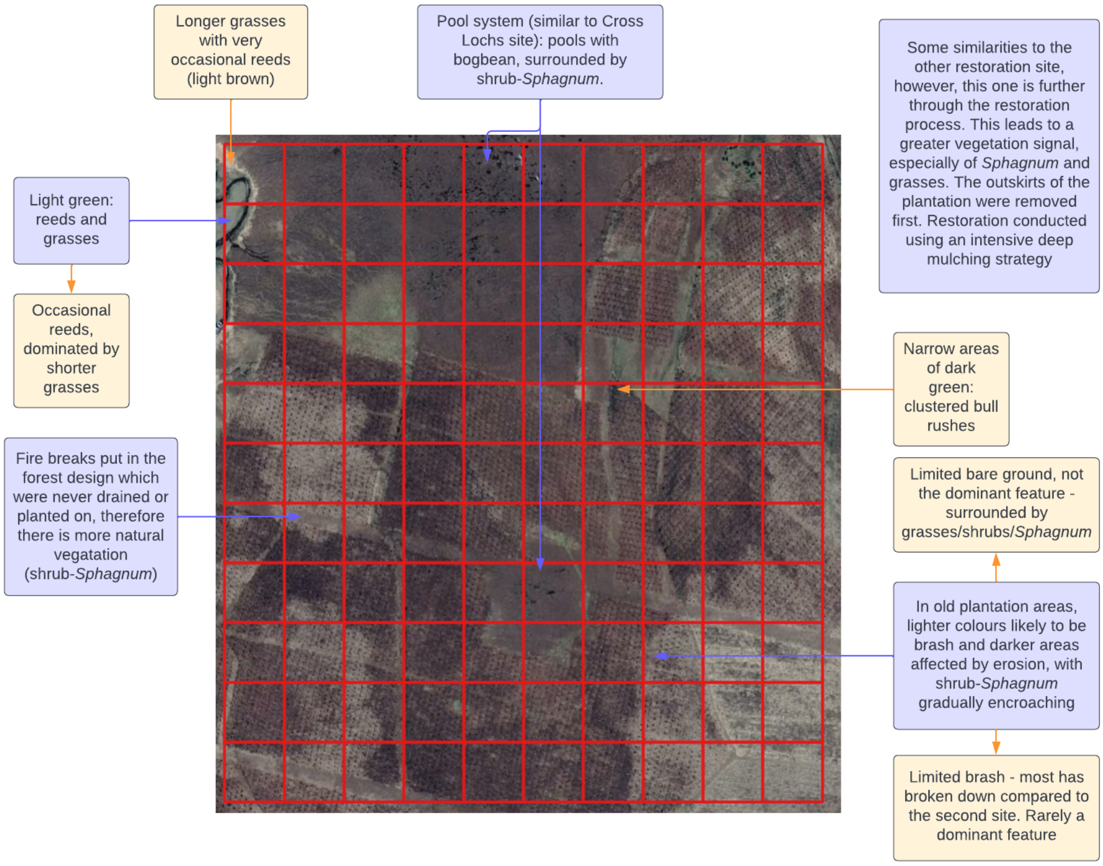


FIGURE 4.10: Annotated satellite image; blue annotations were from expert description of the area with yellow additions post fieldwork

The top outcomes for each of the first restoration site analysis were compared with the training and testing data using confusion matrices, with overall, producer's and user's accuracy to assess the difference between outcomes. The matrices were then compared to each other using the McNemar test equation (Agresti 2006, G.M Foody 2004).

$$z = \frac{FN - FP}{\sqrt{FN + FP}} \ . \tag{4.5}$$

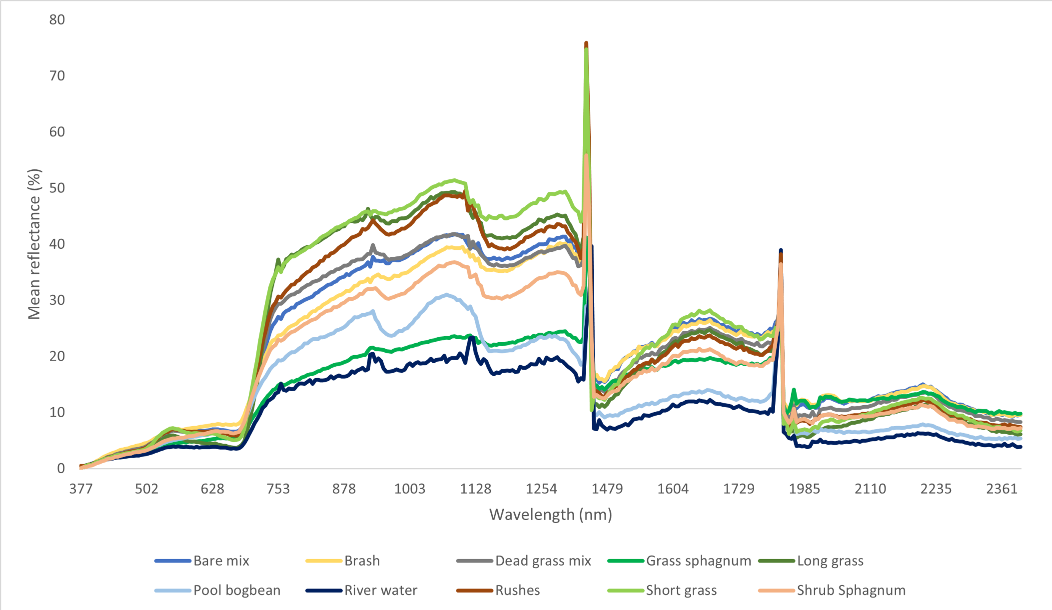


FIGURE 4.11: Spectral library for the first restoration site based on data collected on fieldwork with signatures of each PFT identified

4.4.2 Unsupervised Learning

K-means clustering was undertaken for each site base on the number of predicted PFTs (from the machine learning) and the number of expected PFTs from expert descriptions and fieldwork to see if there was a direct relationship between the spectral signature and the PFTs. Alternative clustering methods could have been applied to the data, such as hierarchical clustering, however, k-means is usually the most efficient and often the most accurate clustering technique (Provost and Fawcett 2013).

4.4.3 Field Sites

Due to visiting restrictions during the breeding season, permission from the RSPB could not be attained meaning fieldwork could not be undertaken until the end of July. This in addition to Covid-19 recovery meant that more limited fieldwork could be conducted than initially intended; the first restoration site was used as the focus as this had the most overlap with the other sites. Prior to visiting the Flow Country, a risk assessment was developed (Appendix G) and field sites were determined. Stratified random sampling was used to collect measurements with plant functional types predicted using expert knowledge of the site and Google satellite images. Areas with these PFTs were selected in a GIS and ten random points from each PFT selected using Python code (Figure 4.12). Due to limited time and accessibility issues, data was not collected from every point (Figure 4.13), however all potential

PFT categories were included in data collection. When in the field, these were located using GPS and a GNSS receiver used to record location (these locations were checked, once out of the field, and compared to the pixel locations from the hyperspectral data). The GNSS receiver typically had access to over 20 satellites and was set to record measurements when within 20 cm accuracy of the receiver's location. However, in some sub-sites, especially to the northwest of the site near the river, the internet connection was lost and the accuracy reduced to over 5 m at times. There will also be some additional errors due to the movement of tectonic plates since 1984 (Mesibov 2012). Whilst in the Flow Country, the Cross Lochs and Erosion sites were also visited with notes taken and some measurements recorded at the Erosion site of features that had not been seen at the restoration site.

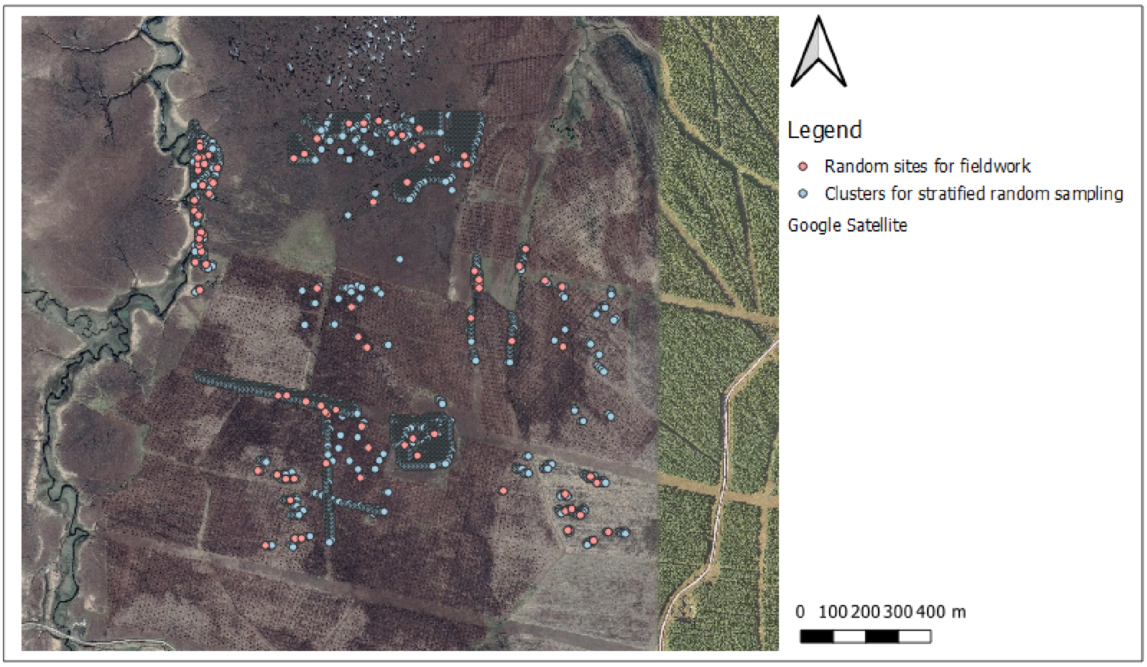


FIGURE 4.12: Clusters of areas containing different potential PFTs based on analysis of satellite imagery and expert descriptions with random sites selected for data collection in the field

Once a sub-site had been located using GPS and Google Earth, the GNSS receiver was used to record the location (Figure 4.14) and four tape measures used to create a 5 m^2 quadrat with the receiver as the central point. The receiver was placed in a globular foot to reduce the impact on the peat surface and ensure consistency as the surface area of the foot prevented the receiver from sinking into the ground. The quadrat was 5 m^2 to reflect the size of the hyperspectral pixels. For each quadrat the approximate proportion of the different plant types was recorded (Figure 4.15), with the key species made up of different types of moss, heather and grass. This could have been improved with the use of additional lines within the quadrat, splitting it

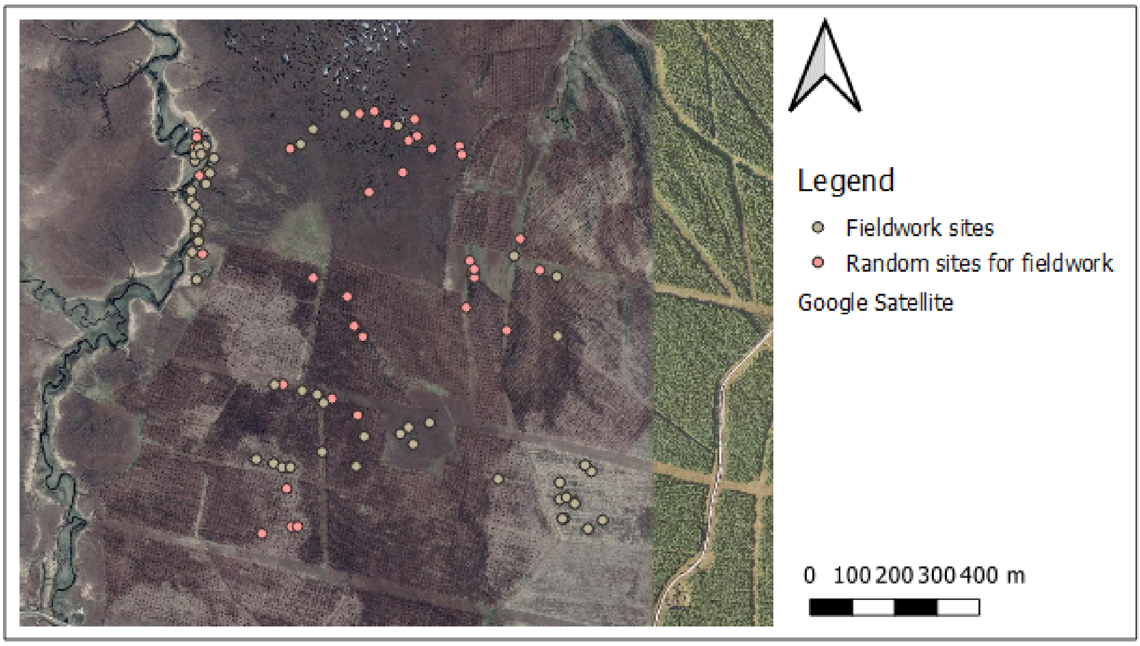


FIGURE 4.13: The field sites where data was ultimately collected from

into a 10x10 grid to more accurately determine the proportion of different species, however, due to time constraints, this was not feasible. An example recording sheet can be found in Appendix A with the proportion of species at ten of the sub-sites which are named based on the random points generated in Python, GPS number, and the time of recording to ensure pictures of vegetation could be linked to each sub-site. Additional notes were made for some sub-sites if there were unexpected features or there was layering of the mosses, shrubs and grasses.

The conditions when recording the data were mainly sunny, however, there was rain for a couple of hours. The main weather factor which affected data collection was the wind which made it more difficult to keep the GNSS receiver perpendicular to the ground, which reduces its accuracy. A key issue with this fieldwork was time related. As the sites were challenging to access, time in the field was limited. In addition to this, Covid-19 recovery meant that fieldwork could not occur at an optimal rate, resulting in less data collected than planned. Despite this, the data gathered was valuable and seeing the sites on the ground improved my understanding of the landscape, enabling better interpretation of the satellite images.

4.5 Peatland Surface Motion and PFT Comparison

The peatland surface motion data required geometric correction in relation to the PFT prediction data and spatial reduction to the size of the first restoration site.

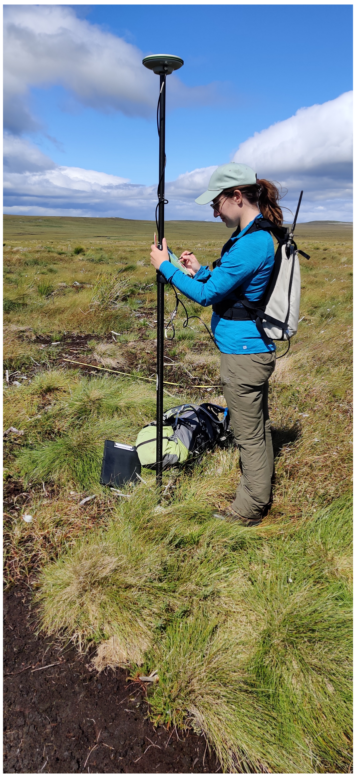


FIGURE 4.14: Locating sites using the GNSS receiver in the first restoration site, with the quadrat then build centred around the receiver

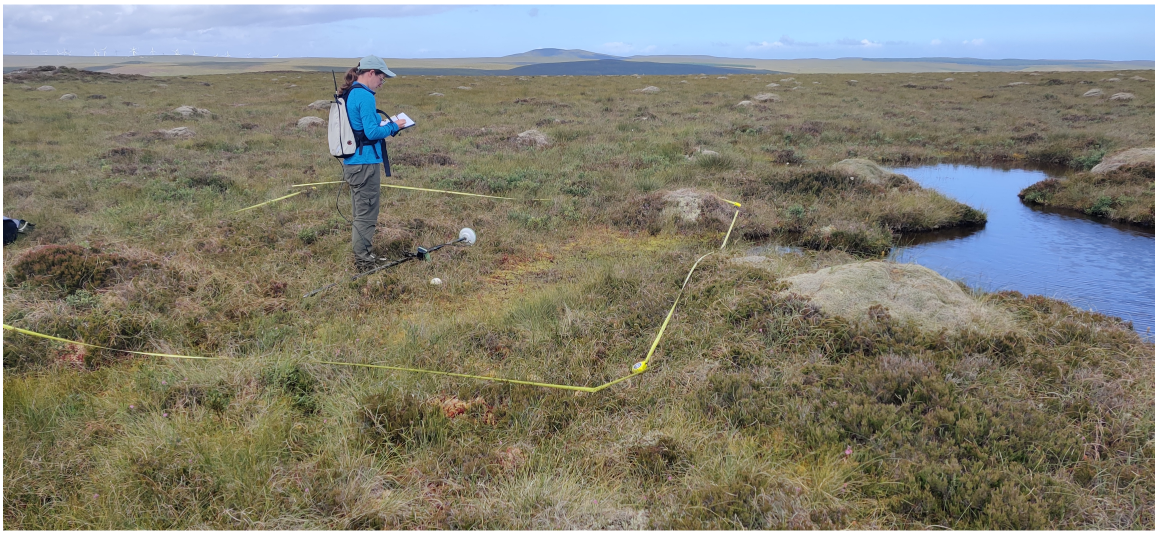


FIGURE 4.15: Species identification and proportioning in a quadrat at the erosion site

This site was chosen as most of the fieldwork was conducted there. To determine whether there is potential to add to the understanding gleaned from PSM data and PFTs generated from the hyperspectral data, analysis was undertaken to assess whether there is a relationship between the velocity, and mean amplitude, peaks and troughs of ground movements and the PFT. The vector points were then converted to a raster dataset for each of these PSM attributes. The raster pixels were then enlarged to meet each other using the r.neighbor function in QGIS and saved as a GDAL virtual format with one pixel every five metres to allow spatial joins with the PFT dataset using k-nearest neighbour, creating a new upscaled vector file. This was then used to perform various classifications: logistic regressions, random forest, decision tree and SVM. The SVM analysis was conducted using the 'RBF' kernel as this has been more successful during exploratory analysis than the linear kernel (Figure 4.5).

Descriptive statistics and histograms (such as those in Appendix C.1 and C.2) were looked at to see initial relationships between the PFTs and PSM attributes. As some histograms looked reasonably similar, correlation matrices were used to assess whether PFT classes were highly correlated with each other. If they were, they would be merged to create a new category for the machine learning classifications. As the classes varied in size, the results were iterated through with random values chosen for those with larger proportions (e.g. Appendix C.3).

Results

5.1 PFT Classifications

5.1.1 Random Forest PFT Predictions

The random forest results were variable depending on the site, whether the whole dataset was used and which data transformation was classified. All of the highest random forest outcomes used the full dataset, with 17 of the 40 highest outputs from the original data, 11 and 10 from the first and second derivative transformations respectively and just two from the continuum removal transformation. The first restoration site, using focused train-test data, random forest accuracies were greater than all other sites in five of the eight scenarios, reaching 78.91 % in the best case (Figure 5.1), when the training and testing data only from this site was applied. However, when all the training and testing data was applied to this site, the accuracy declined, resulting in lower accuracies than other sites, especially when the cross validation was three, depth 4 and testing size 0.3 (Figure 5.1).

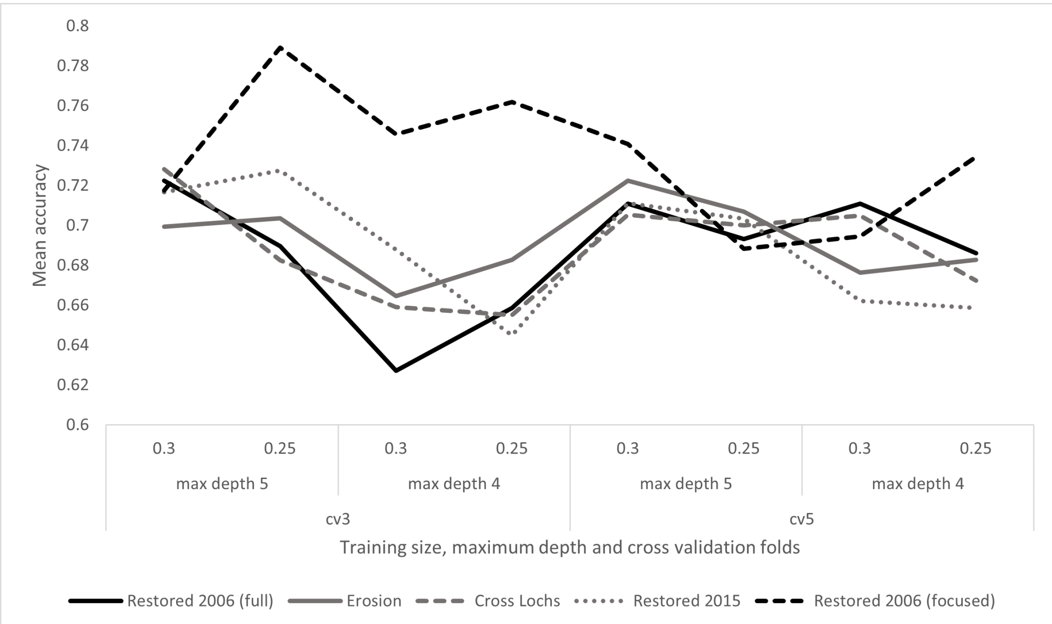


FIGURE 5.1: The highest outcomes for each site and for each train-test ration, maximum depth and internal cross validation

The highest five outputs were recorded and outputted for each site. These were combined for the first restoration site (highest ten outputs summarised in Table 5.1

(full table found in Appendix C)) and used to assess in more depth which spectral ranges and data transformations were more useful (Figure 5.2). This confirmed that the full dataset was required for determining PFTs using random forests, with limited benefits of assessing the smaller spectral ranges, other than potentially the visible part of the spectrum. In terms of the data transformations, overall for this site, the most accurate predictions used the second derivative transformation with the full dataset, (87.5 % of the highest outcomes). The standard deviation was low across all outcomes, suggesting the data is reliable due to close clustering around the mean.

TABLE 5.1: Highest ten outputs for the first restoration site using the fieldwork focused train-test data

Spectral range	Data transformation	Mean accuracy	Standard deviation	Training size	Max depth	Number of cross validation folds
Full	Original	0.77	0.05	0.25	5	3
Full	Second derivative	0.76	0.03	0.25	4	3
Visible	Second derivative	0.76	0.06	0.25	4	3
Full	First derivative	0.75	0.04	0.30	4	3
Red edge	Second derivative	0.74	0.03	0.25	4	3
Full	First derivative	0.74	0.08	0.30	5	5
Full	Second derivative	0.73	0.06	0.25	4	5
Full	Original	0.73	0.05	0.30	4	3
Visible	Original	0.73	0.04	0.25	5	3
Full	First derivative	0.72	0.08	0.25	4	5

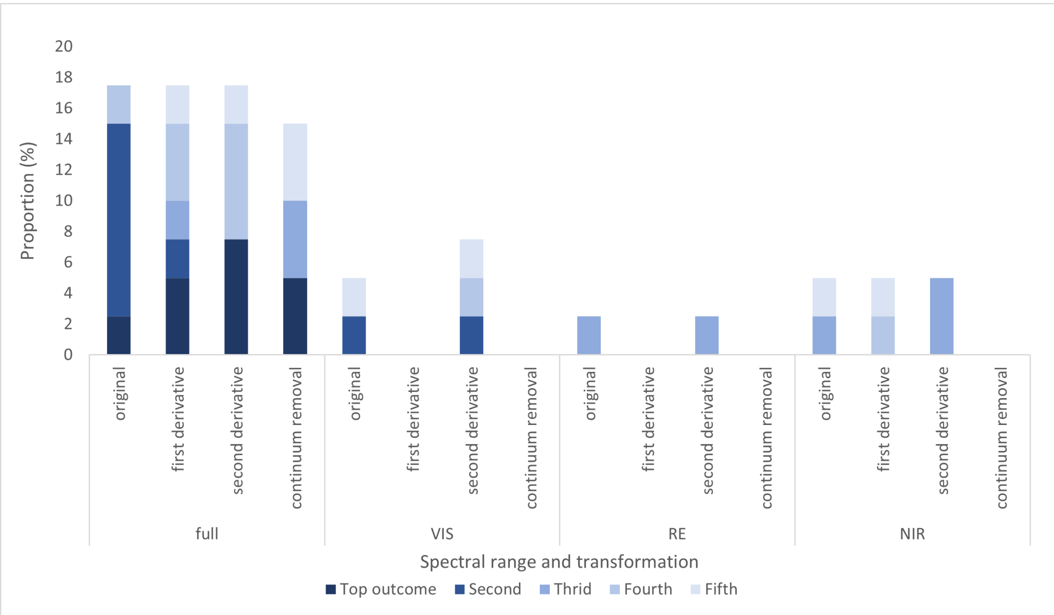


FIGURE 5.2: A summary of the highest five random forest classification outcomes for each iteration through the first restoration site using the fieldwork-focused train-test data

The map outcomes for the first restoration site vary significantly as demonstrated in Figures 5.3 and 5.4a. This suggests that the spectral signatures of shrub and *Sphagnum*, and grass and *Sphagnum* are very similar. This is further reflected by the confusion matrices (Table 5.2 and Table 5.3). With an overall accuracy of 86 %, it is much more likely that the map output associated with the more focused train-test data was closer to the actual PFTs, compared to an accuracy of 72 % for the site when the full train-test data was used in the random forest classification. This is reflected by the McNemar z scores from the comparison of confusion matrices. When comparing the outcomes with the best accuracies from the first restoration site (based on the focused train-test data), there were consistent outcomes with a z score of 2.708 which is less than the critical value of 3.841 (Table 5.6, therefore there is similarity between the predictions. However, there is a significant difference between outcomes when comparing the first restoration site outcomes when using the full train-test dataset compared to the focused one, with a z score of 5.196 (Table 5.5).

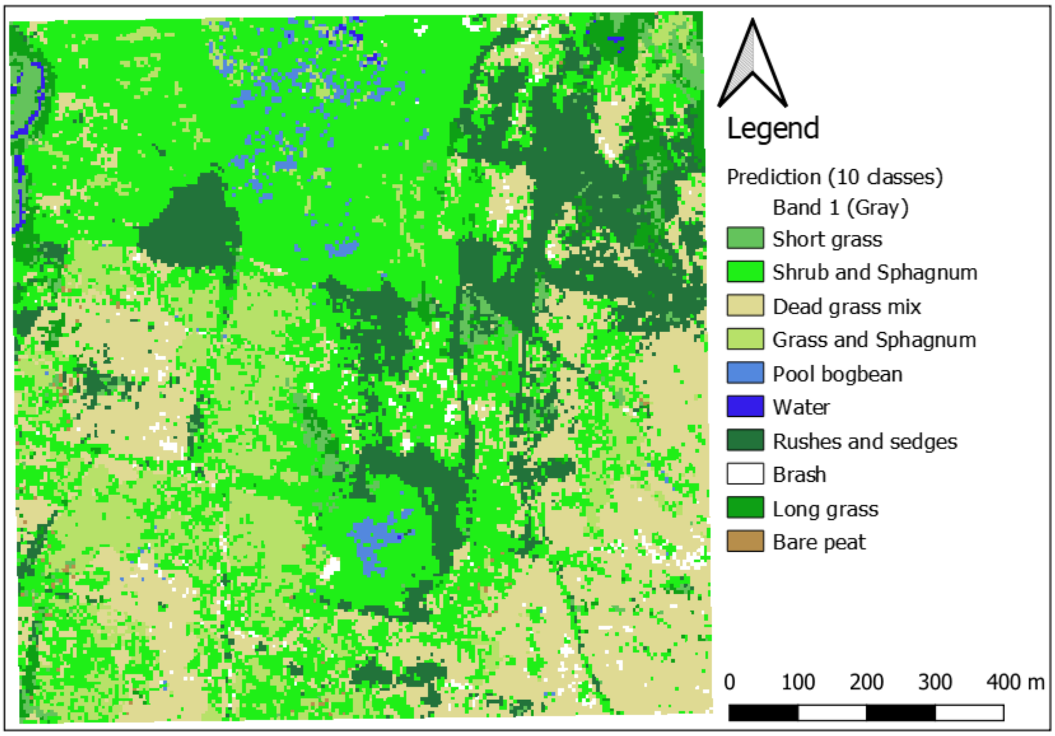
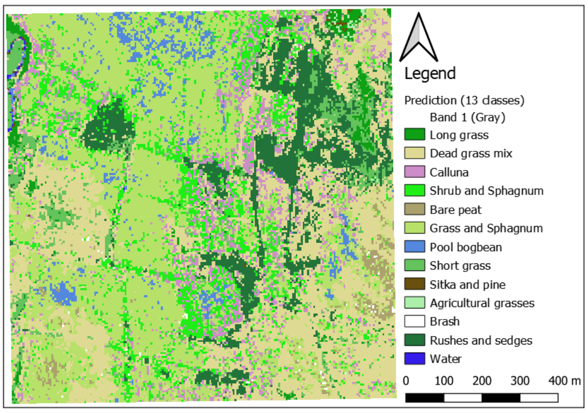
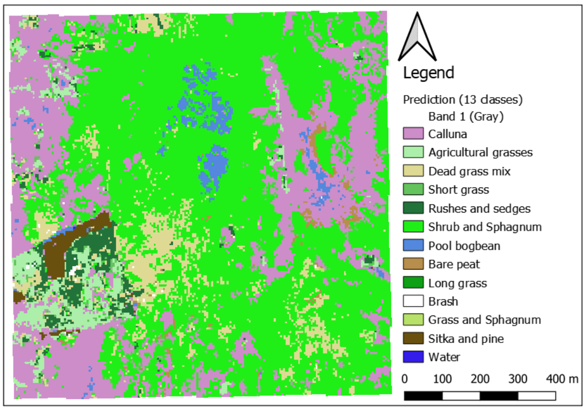


FIGURE 5.3: Random forest classification map of predicted PFTs for the first restoration site with 10 classes using fieldwork-based train-test data (accuracy 0.770.05

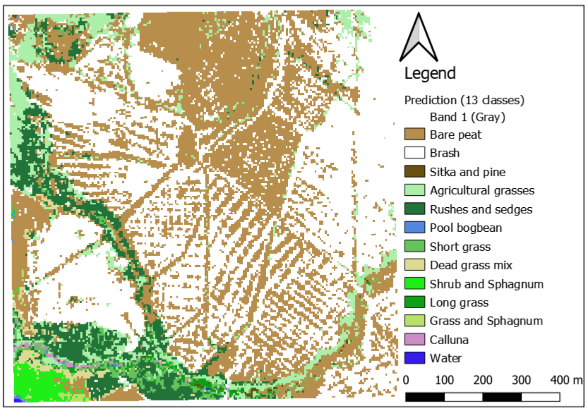
The other sites, which relied more on training and testing data from the first restoration site had similar accuracies of 0.722, 0.728 and 0.727 (Figures 5.4b, 5.4c, 5.4d). When comparing the initial training and testing outlined in the methodology for Cross Lochs using satellite imagery and expert knowledge to develop PFTs, had



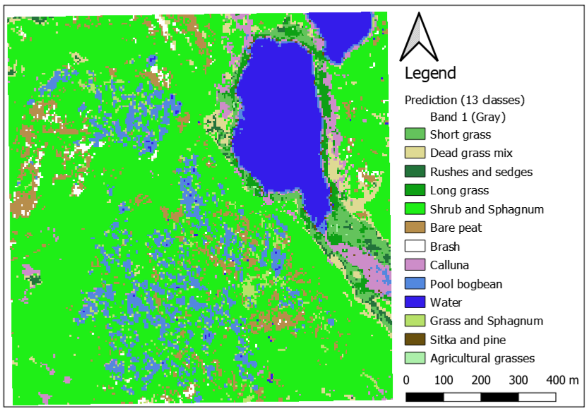
(A) Predicted PFTs for the first restoration site (accuracy: 0.720.03



(B) Predicted PFTs for the erosion site (accuracy: 0.7220.03



(C) Predicted PFTs for the second restoration site (accuracy: 0.7280.03



(D) Predicted PFTs for the Cross Lochs site (accuracy: 0.7270.03

FIGURE 5.4: Random forest classification map of predicted PFTs for the four sites using the whole train-test dataset which was applied across the sites

better random forest accuracies, with highest outcomes of 73 % and 79 %, which are slightly higher than the outputs with the full train-test dataset, but less consistent. This is likely to be linked to 13 PFTs being included rather than six. The dominant PFT is shrub and *Sphagnum* with the location of water and pools being very clear. Other PFTs are less dominant with agricultural grasses and Sitka and pine (which were not found in this location) making up a small number of points (20 and 72 pixels respectively, compared to the 26,114 shrub and *Sphagnum* pixels). The second restoration site outcomes are similar to what was expected from the expert knowledge and satellite imagery and the erosion site outputs mainly reflected what was observed in the field. However, the erosion site prediction did contain more brash and dead grass mix than expected. The overall accuracy of the combined sites compared to the train-test data was 75 %, with most PFTs being correctly identified much of the time. The shrub and *Sphagnum*, and water PFTs were the classes most likely to be misidentified as reflected by the producer's accuracies of 0.36 and 0.21 respectively, whereas, all other PFT producer accuracies were aver 65 % with Sitka and pine and short grass highest performing with producer's accuracies of 91 % and 90 % respectively. This suggests that their spectral signatures are more consistent across the sites, with little variation in the conditions of the areas containing these species.

TABLE 5.2: Confusion matrices for the first restoration site predictions compared to the full train-test dataset, with an overall accuracy of 0.72.

	Predicted PFT													
Sample PFT	Bare mix	Brash	Calluna	Dead grass mix	Grass and *Sphagnum*	Long grass	Pool bogbean	Rushes and sedges	Short grass	Shrub and *Sphagnum*	Sitka and pine	Water	Row total	Producer's accuracy
Bare mix	**1**	0	4	8	4	0	0	0	2	3	0	0	22	1.00
Brash	0	**5**	2	1	2	0	0	0	0	10	0	0	20	1.00
Calluna	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0.00
Dead grass mix	0	0	0	**75**	5	0	1	3	5	1	0	0	90	0.68
Grass and *Sphagnum*	0	0	1	2	**61**	0	7	0	0	0	0	0	71	0.56
Long grass	0	0	1	3	0	**83**	2	0	4	0	2	0	95	0.95
Pool bogbean	0	0	0	0	3	0	**10**	0	0	0	0	0	13	0.40
Rushes and sedges	0	0	3	5	1	1	0	**56**	1	0	0	0	67	0.92
Short grass	0	0	3	2	0	2	1	0	**82**	0	0	1	91	0.80
Shrub and *Sphagnum*	0	0	4	13	33	1	0	2	8	**29**	0	0	90	0.67
Sitka and pine	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0.00
Water	0	0	1	1	0	0	4	0	0	0	1	**12**	19	0.92
Column total	1	5	19	110	109	87	25	61	102	43	3	13	578	
User's accuracy	0.05	0.25	0.00	0.83	0.86	0.87	0.77	0.84	0.90	0.32	0.00	0.00		

TABLE 5.3: Confusion matrices for the first restoration site predictions compared to the site-specific train-test dataset (from fieldwork) with an overall accuracy of 0.86

	Predicted PFT											
Sample PFT	Bare mix	Brash	Dead grass mix	Grass and *Sphagnum*	Long grass	Pool bogbean	Rushes and sedges	Short grass	Shrub and *Sphagnum*	Water	Row total	Producer's accuracy
Bare mix	**5**	0	0	0	0	0	0	0	0	0	5	0.23
Brash	1	**14**	0	0	0	0	0	0	2	0	17	0.70
Dead grass mix	7	2	**81**	3	2	0	3	1	7	0	106	0.90
Grass and *Sphagnum*	1	0	0	**65**	0	1	0	0	0	0	67	0.92
Long grass	2	0	1	0	**84**	0	1	6	1	0	95	0.88
Pool bogbean	0	0	1	0	0	**9**	0	0	2	0	12	0.69
Rushes and sedges	1	0	6	0	0	0	**61**	0	0	0	68	0.91
Short grass	0	0	0	0	6	0	1	**82**	1	2	92	0.90
Shrub and *Sphagnum*	5	4	1	3	0	3	1	1	**77**	0	95	0.86
Water	0	0	0	0	3	0	0	1	0	**17**	21	0.89
Column total	22	20	90	71	95	13	67	91	90	19	578	
User's accuracy	1.00	0.82	0.76	0.97	0.88	0.75	0.90	0.89	0.81	0.81		

5.1.2 K-means Cluster Outputs

K-means analysis maps were outputted using the same number of clusters as PFTs to demonstrate whether unsupervised learning techniques could be used to estimate the location of different PFTs. The maps (Figure 5.5) demonstrate that with a

TABLE 5.4: Confusion matrices for the predictions from all four sites compared to the full train-test dataset, with an overall accuracy of 0.75

	Predicted PFT														
Sample PFT	Agricultural grasses	Bare mix	Brash	Calluna	Dead grass mix	Grass and *Sphagnum*	Long grass	Pool bogbean	Rushes and sedges	Short grass	Shrub and *Sphagnum*	Sitka and pine	Water	Row total	Producer's accuracy
Agricultural grasses	**46**	1	0	2	0	0	0	0	0	0	0	1	0	50	0.84
Bare mix	0	**59**	5	0	0	0	0	0	0	0	0	0	0	64	0.66
Brash	2	3	**90**	0	0	0	0	0	0	0	0	0	0	95	0.82
Calluna	0	5	2	**59**	0	1	1	0	3	3	4	1	1	80	0.65
Dead grass mix	0	8	1	8	**75**	11	3	0	5	2	13	0	1	127	0.83
Grass and *Sphagnum*	0	4	2	0	5	**62**	0	8	1	0	33	0	1	116	0.69
Long grass	0	0	0	0	0	0	**83**	0	1	2	1	0	0	87	0.87
Pool bogbean	0	2	0	0	1	7	2	**67**	0	1	0	3	8	91	0.74
Rushes and sedges	2	0	0	1	3	2	0	0	**56**	0	2	3	0	69	0.84
Short grass	5	2	0	0	5	0	4	0	1	**82**	8	0	0	107	0.90
Shrub and *Sphagnum*	0	6	10	21	1	7	0	15	0	0	**34**	0	19	113	0.36
Sitka and pine	0	0	0	0	0	0	2	0	0	0	0	**82**	1	85	0.91
Water	0	0	0	0	0	0	0	0	0	1	0	0	**59**	60	0.21
Column total	55	90	110	91	90	90	95	90	67	91	95	90	90	1144	
User's accuracy	0.92	0.92	0.95	0.74	0.59	0.53	0.95	0.74	0.81	0.77	0.30	0.96	0.98		

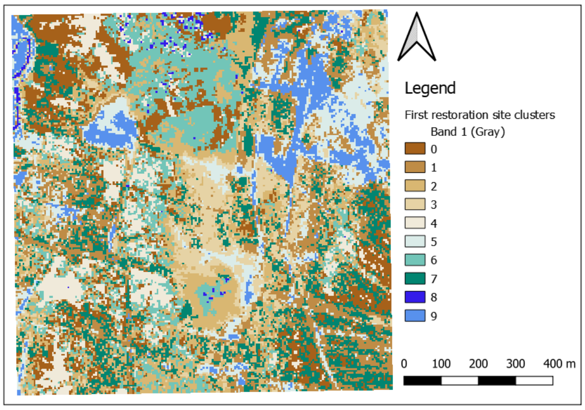
TABLE 5.5: The comparison of the classifications derived from the confusion matrices (Tables 5.2 and 5.3, generated from the predictions from the first restoration site using the focused train-test dataset and the full train-test dataset. z = 5.196

		10 classes First Restoration Site	
	Allocation	Correct	Incorrect
13 classes First Restoration Site	Correct	333 (TP) 58 %	81 (FN) 14 %
	Incorrect	162 (FP) 28 %	2 (TN) 0.35 %

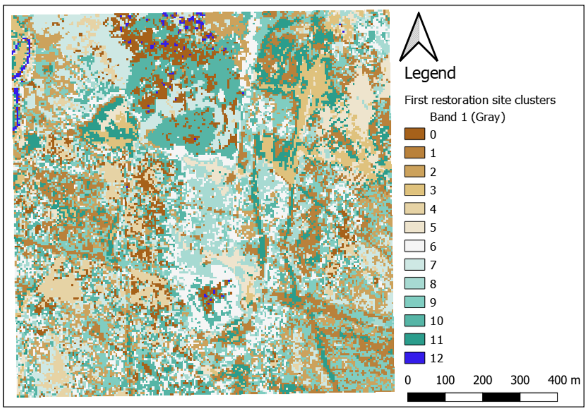
TABLE 5.6: The comparison of the classifications derived from the highest two confusion matrices for the first restoration site with a testing size of 0.25 and maximum depth of 5, generated from the predictions from the first restoration site using the focused train-test dataset. z = 2.708

		Full spectral range, original transformation	
	Allocation	Correct	Incorrect
Visible range, original transformation	Correct	451 (TP) 78 %	22 (FN) 3.8 %
	Incorrect	44 (FP) 7.6 %	61 (TN) 11 %

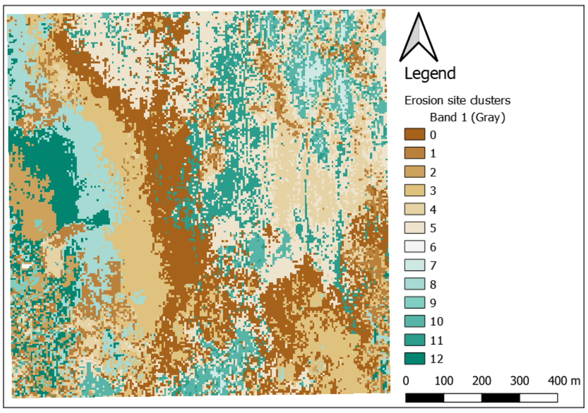
large number of PFTs, it is only possible to discern larger bodies of water with other groups being less easily classified. However, the second restoration site much better reflects the PFTs than the other k-means clustering outcomes.



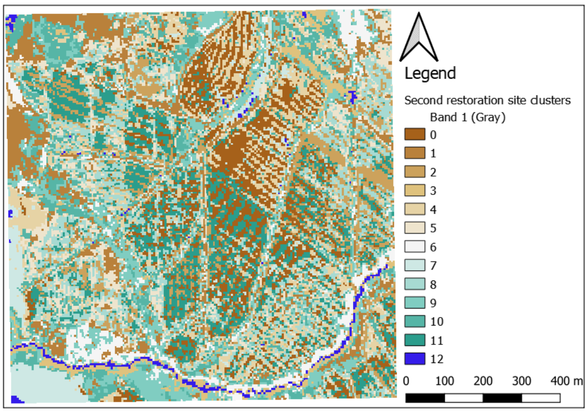
(A) A ten k-means cluster for the first restoration site



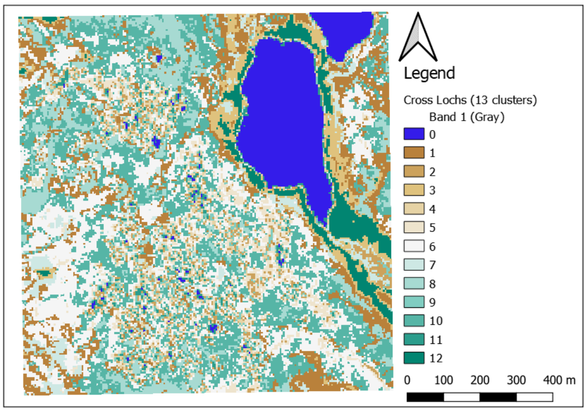
(B) A 13 k-means cluster for the first restoration site



(C) A 13 k-means cluster for the erosion site



(D) A 13 k-means cluster for the second restoration site



(E) A 13 k-means cluster for the Cross Lochs site

FIGURE 5.5: K-means clustering for the four sites with 'k' detemined by the number of PFTs

5.2 PSM Relationship With PFTs

When conducting similarity tests using correlation matrices for each PSM attribute, relationships were not strong (e.g. Appendix C.3), therefore no PFTs were combined prior to machine learning classifications. The accuracy of the PSM machine learning classification for the first restoration site was much lower than when creating the PFTs (highest accuracy 53 %). The random forest classifications outperformed the others (Figure 5.6) for both sets of PFTs (with 10 and 13 PFT classes). Machine learning was not as effective in this case requiring improvements to be made to the method prior to analysis. The low resolution of the PSM data makes it more challenging to compare the different attributes to the PFTs, as reflected by the k-means clusters for this site (Figure 5.7).

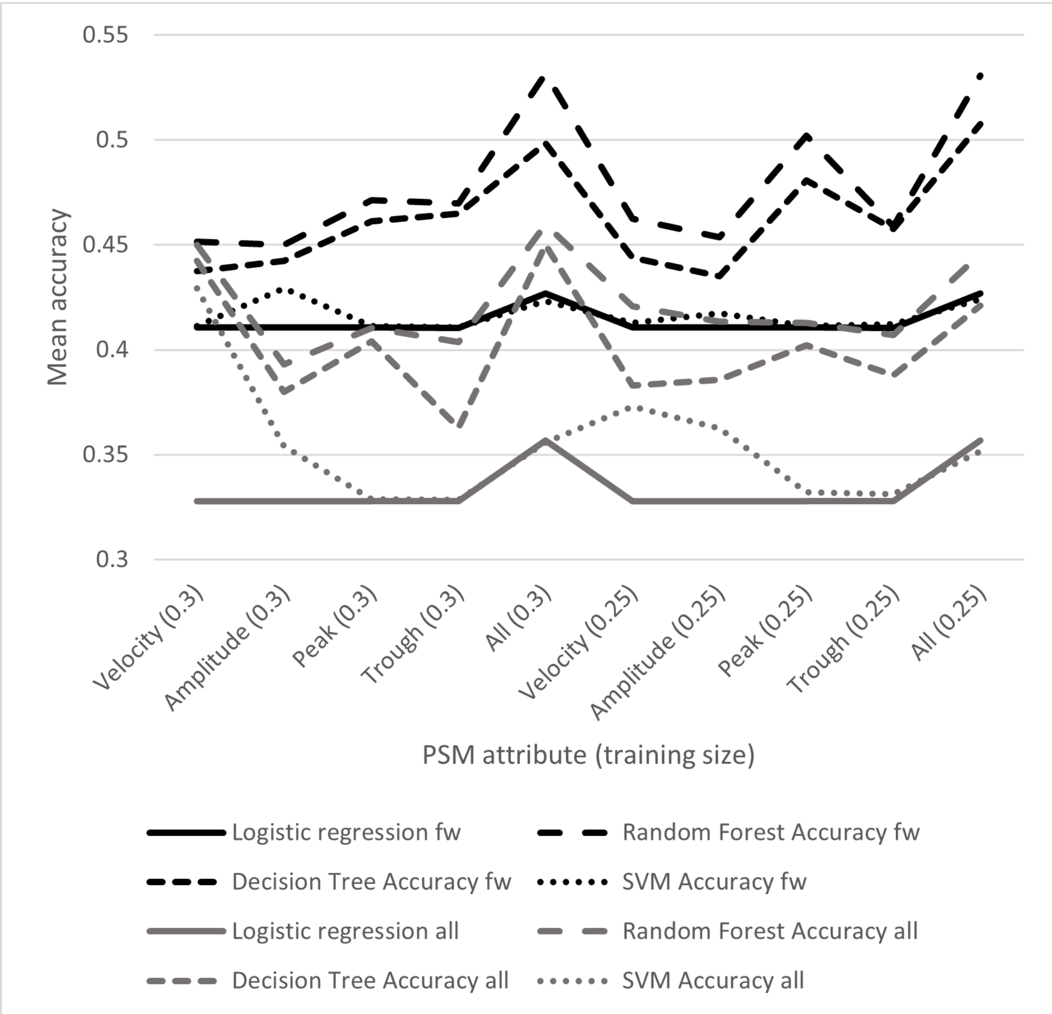
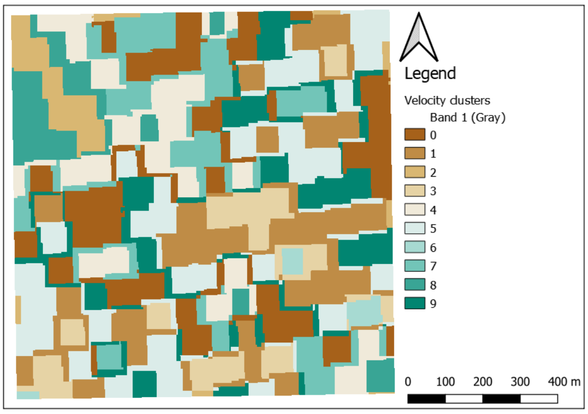
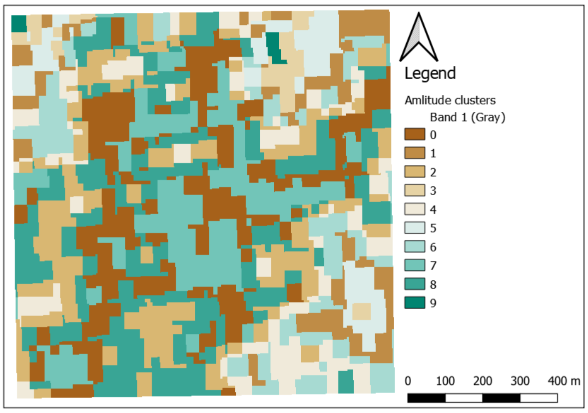


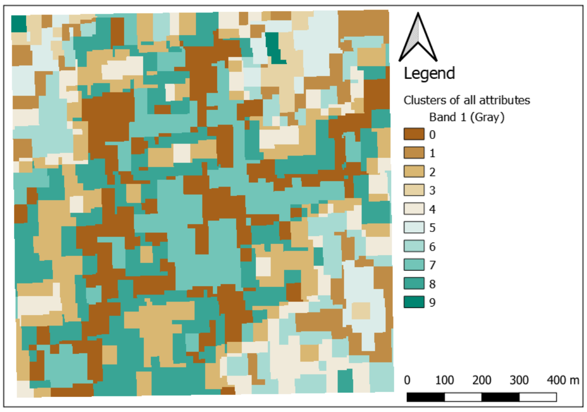
FIGURE 5.6: Mean accuracy of difference machine learning classifications when using PSM to predict PFTs. 'Fw' relates to the predictions determined from the fieldwork-focused train-test data and 'all' refers to predictions determined using the full train-test dataset



(A) Cluster for the velocity PSM attribute



(B) Cluster for the amplitude PSM attribute



(C) Cluster for the velocity, amplitude, peak and trough timings

FIGURE 5.7: K-means clustering, with ten clusters, for a range of PSM attributes in the first restoration site

Discussion

6.1 Machine Learning PFT predictions

6.1.1 Random Forest PFT Predictions

The results demonstrate that a random forest classier can be used to determine likely PFTs in an area, however, accuracy is higher when there are fewer classes and when more fieldwork is undertaken. This is reflected by the first restoration site where the majority of field data was collected and suggested there were ten PFTs at this site which had a higher accuracy than when trained with data from additional sites, with less fieldwork (Figures 5.3 and 5.4). When in the field, some of the 5 m^2 sites contained a large range of species, making it more difficult to assign a specific category, with the most dominant vegetation determining the group. However, some vegetation may dominate the spectral signature due to the layering of vegetation with mosses often covered by grasses, shrubs and rushes/sedges. Some of the spectral signatures are similar for different PFTs (Figure 4.11) which could affect accuracy, especially where pixels included areas of transition between different vegetation groups. When in the field, it was observed that there were strong transitions between some PFTs, such as transitioning from 100 % long grass to 100 % short grass with a sharp boundary. However, other transitions occurred more gradually, such as from areas with predominantly shrubs and *Sphagnum* to mainly grass and *Sphagnum*. Additionally, in areas with pool bogbean, it is very likely that part of the pixel would contain shrubs and *Sphagnum* in addition to the pool. The spectral similarity between the shrub and *Sphagnum*, and grass and *Sphagnum* PFTs could be why the fieldwork-focused and full training-testing maps for the first restoration site differ so much and that is reflected in the McNemar similarity score of 5.196 (Table 5.2), which is greater than the critical value of 3.841 meaning there is a statistically significant difference between the two confusion matrices for the site at the 0.05 level. Whereas, when the highest two outputs from the fieldwork focused training and testing data were compared, there was not a significant difference (similarity value of 2.708; Table 5.3). This demonstrates that the additional data from the erosion site probably had a different spectral signature, thereby changing the machine learning prediction. The shrub and *Sphagnum* PFT is a category which contains several species (see data collection sheet Appendix B) and if some were less dominant in the first restoration site, but more frequent in other sites, it is more likely that they would be misidentified. The potential of this issue was reduced by including samples from other sites and from two different parts of the first restoration site as the firebreak and pool systems were both dominated by shrub-*Sphagnum* mixes. This demonstrates that there are challenges when applying data collected from one site to others, especially when the species within the PFTs could vary depending on the land use.

The conditions across the sites could also vary, affecting the application from one area to another. Microtopography, wetness and stresses caused by restoration rate, consumers and human activity could affect the spectral reflectance of different species. This is why there are no large scale spectral libraries for peatland species; the spectral reflectance can vary significantly for each species, especially *Sphagnum* due to their capacity to store water (with some species able to store 20 times their own weight in water due to the fibrous structure of the moss Marsden and Ebmeier 2012). Grazing would have affected species growing at the erosion site where there were no deer fences, increasing the damage to vegetation, especially mosses which are less able to recover (Marsden and Ebmeier 2012). This in addition to changes in water table reduce the stability of the moss, making it more vulnerable to invasion from other species (Hambley 2016).

This is also reflected in the Cross Lochs site where the initial exploration took place and it was the site with the best classification accuracy after the first restoration site. In the initial exploration, only six PFTs were identified, these had much more distinctive spectral signatures than those from the restoration site. This increased the likelihood of more accurate allocation, with some of the PFTs barely registering in this location (agricultural grasses and Sitka and pine, with likely misidentification for these species due to the near-natural nature of the site). Additionally, as this area was in a near natural state, pool systems dominated the landscape, with little impact from human activity or large herbivores, fewer stresses are put on the vegetation. This is in strong contrast to the second restoration site where brash and bare peat still dominate the landscape following the felling of trees over the past seven years. It takes time for succession to occur and the landscape will be less stable with less developed peat limiting the resilience of the landscape to small changes in climate.

Although field data were collected in July 2022 (therefore at approximately the same time of year as the hyperspectral data was collected), weather conditions were different in 2021, meaning that different vegetation could dominate. In the Forsinard area, from January to July 2021 there was 29 mm less precipitation than in the same period in 2022 and 105 mm less from April to July (Figure 6.1). Whereas the temperature was likely to be more than 1°C higher in 2022 as reflected by the nearest Met Office monitoring station in Wick where maximum and minimum temperatures were more than 1.1°C for the maximum temperature ranges from January to July and April to July 2022 than 2021, with minimum temperatures averaging over 1.3°C more (Figure 6.1). A warmer, wetter 2022 could lead to difference vegetation mixes compared to a cooler, drier 2021 as more precipitation increases the likelihood of pools systems growing with more extensive *Sphagnum* mosses, however, higher temperatures could affect evaporation rates, limiting their growth. Higher temperatures would also increase the rate of oxidation and decomposition, decreasing the chance of peat build up and reducing the amount of dead grass in 2022 compared to 2021. This could cause slight changes in PFTs at all sites, however, additional annual changes are more likely to occur at sites that have been subject to restoration or erosion (Hancock et al. 2018). The more recent the restoration, the more change there should be (Hancock et al. 2018, Marshall et al. 2021) meaning that that PFTs at the second restoration site may not reflect ground data. Similarly, areas that have been eroded and are vulnerable to further degradation are likely to change more rapidly than more natural sites.

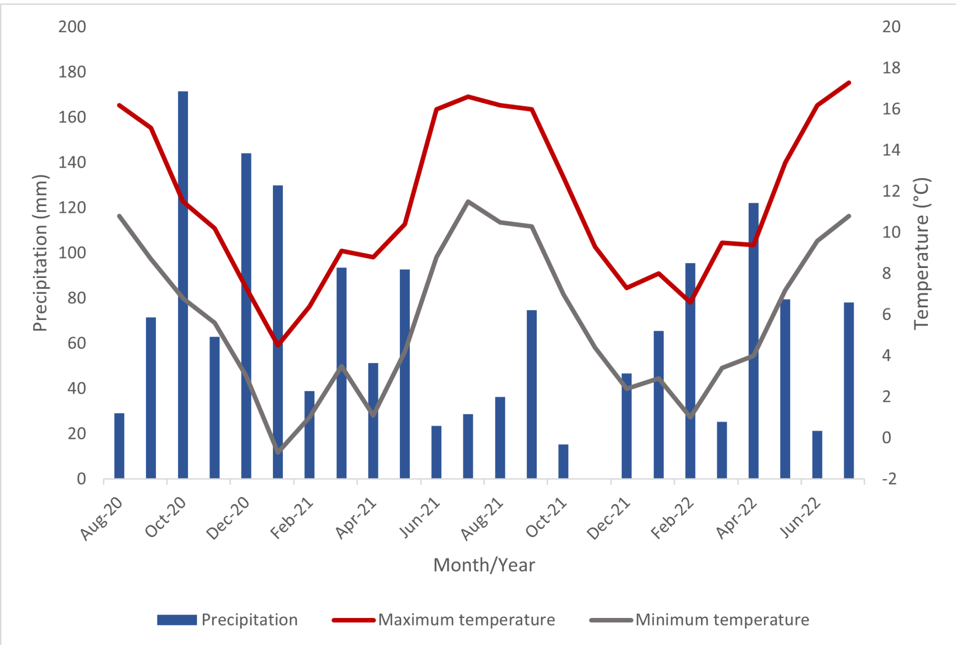


FIGURE 6.1: Precipitation (at Forsinain) and temperature (at Wick airport) each month from August 2020 to July 2022 (Met Office 2022, SEPA 2022)

High accuracy does not necessarily reflect reality, in both map outcomes for the first restoration site, there were large areas of rushes and sedges, however, it was observed in the field that although there were some clusters of rushes and sedges covering over 5 m^2, the majority of were a small part (less than 10 %) of other 5x5 m subsites. It could be that once rushes and sedges make up a certain proportion of the pixel (whether or not the dominant species), it dominates the spectral signature, causing it to be over represented. This is further reflected by the PCA analysis (Figure 6.2) which demonstrates that the different PFT clusters are not distinct and interrelate with each other. However, it could be that in these locations, which contained no specific field sites but were observed during movement between points, there were younger rushes and sedges which did not visually dominate the landscape, but did register spectrally. The analysis using the full train-test data, returned some values with agricultural grasses, Sitka and pine, and calluna. Although, there was calluna present at this site it did not dominate the landscape, however, it was unsurprising that this PFT was included in the prediction as concentrations in some areas which were not measured specifically could have included relatively high concentrations of this species. There were few agricultural grasses outputs and these were all to the east of the site where there was less data collection and observation, therefore it is possible that there were some of these grasses sporadically located in that area. More fieldwork is required to determine the accuracy of this prediction. Conversely, there were no trees at the site suggesting spectral similarity with some vegetation mixes causing an incorrect prediction. Overall, however, most of the outcomes were as expected based on fieldwork, satellite image analysis and expert descriptions across the sites.

Comparisons with Other Studies

Erudel et al. (2017) used machine learning to identify specific species (rather than PFTs) in a French pine peat-bog and found machine learning accuracies of over 80 %, which was higher than the mean accuracies for all sites with the highest outcome of 78.91 % for the classification using the fieldwork-focused data. This difference is likely to be due to the high spatial resolution of the French study data (1 m resolution using a handheld spectroradiometer). Their data was also collected under cloudless conditions, however some parts of the complete Flow Country data (prior to spatial reduction) contained cloud. Whereas the Hati et al. (2021) study focusing on mangrove forest classification also used 5 m spatial resolution AVIRIS-NG data with a classification accuracy of 87.61 % classifying 24 groups using SVM, mainly at species level. This is likely to have been more accurate than this study due to the clustering of species and more pre-processing of data prior to use. Therefore, future use of the Flow Country AVIRIS-NG data should convert top of the atmosphere radiance to surface reflectance and use a radiative transfer model to improve atmospheric correction in addition to the band removal applied in the methods. Erudel

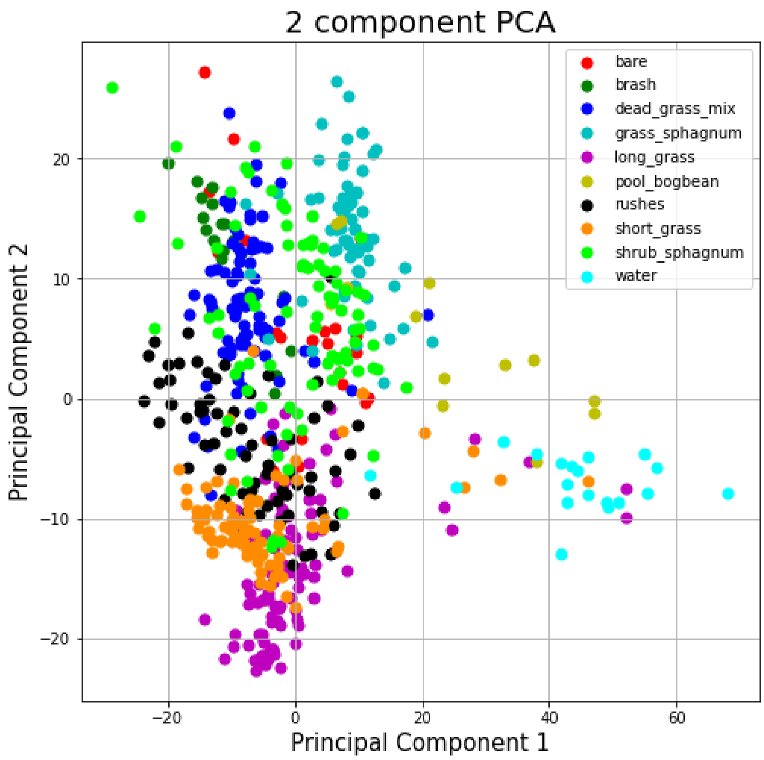


FIGURE 6.2: Scatter plot for the two dominant PCA bands regarding the fieldwork-focused train-test data.

et al. (2017) also advise using a Savitzky-Golay filter to reduce noise.

As discussed by Räsänen et al. (2019) it makes more sense to create PFTs based on a community species mix more likely to be found in the field (as undertaken in this study), than just on a specific collection of biologically related species (shrubs, graminoids and mosses) as most of the time they do not occur in isolation. In their study on a Finnish peatland, the overall accuracy of random forest classification was only 72 %, compared to the overall accuracy of 86 % in this analysis when comparing the predictions to the training and testing data of the fieldwork-focused site. This could be due to the number of classes; they only created five general classes, which may need to be broken down into smaller groups. Similarly, the overall accuracy from the confusion matrices was higher in this study than the best outcome of 77.21 % in the Erudel et al. (2017) study. Whereas, the overall accuracy was much higher in the Polish study focused on the invasive species, steeplebush (*Spiraea tomentosa*), which affects central European peatlands, using random forest to predict its location in a woodland (Kopeć et al. 2020). Their overall accuracies ranged from 92.73 to 98.57 %, which were probably due to the high spatial (1 m) and spectral resolution (400-2,500 nm) of the data used to identify the species and that data outputs were split into two categories; *Spiraea tomentosa* and 'background'. Locating one species in the Flow Country dataset would only be viable if it were a dominant species and due to the coarse resolution, it is likely much would be missed due to the heterogeneous nature of the landscape. Whereas the Polish study site was not complex in terms of the species variety and *Spiraea tomentosa* grows in large single-species patches (Kopeć et al. 2020), so was more identifiable.

Producer's accuracies in the higher scoring confusion matrix generated by Erudel et al. (2017), ranged from 36.61 to 96.36 %, however, these were generally lower for the restoration site ranging from 23 to 92 %. The species/group with the lowest producer's accuracy both sites were not common to the other location (butterworts in Bernadouze, France (Erudel et al. 2017), compared to bare peat in the first restoration site). However, the species (spike sedge) producer's accuracies was the highest (Erudel et al. 2017) as was the rushes and sedges PFT, suggesting they can be more easily identified using spectral signatures. It is generally more useful to focus on producer's accuracy rather than user's accuracy or overall accuracy as it helps determine whether the data is suitable for making predictions (Foody 2008). Although difficult to compare outcomes with the Finnish study, similarly, the producer's accuracies were variable, ranging from 32 % (Tussock flark) to 80 % (Wet flark), suggesting that either more groups needed to be created or data with a higher spectral resolution used for analysis as they only included five PFTs and the full spectral range was not used. This suggests that the poorer spatial resolution, but higher spectral resolution used in this study was more valuable than the sub 3 m spatial resolution with less spectral range used by Räsänen et al. (2019). However, the Polish study had both high spatial and spectral resolution and although the producer's accuracy for *Spiraea tomentosa* was generally high (ranging from 68.18 to 84.33 %) there were probably errors when predicting the location of the species.

Compared to other studies the mean accuracy is not as high, mainly due to the coarse resolution of the hyperspectral in this study. Other studies that used handheld spectrometers and UAV data, typically had a high spatial resolution up to 2.5 cm accuracy (Honkavaara et al. 2016). This increases the probability of heterogeneous sites with more mixed vegetation and more boundaries between different vegetation types, especially for those affected by changes in topography and water levels. However, this reduces the clarity of the PFTs and increases spatial confusion. The use of fieldwork improved clarity for the first restoration site, but the use of satellite images and site descriptions exacerbates these issues, as does the number of PFTs and different conditions at different locations.

Erudel et al. (2017) found that the first derivative transformation produced the most accurate predictions, however, this not the case in this study, with 17/40 (42.5 %) of the highest outcomes from the original data and 11/40 (27.5 %) from the first derivative which was similar to the second derivative (25 %). This could have been due to sensitivity originating from the atmospheric correction of the airborne data, which was not required by studies using handheld spectrometers (Erudel et al. 2017).

6.1.2 K-means Clustering

The k-means cluster analysis demonstrates that the higher the number of clusters the less linked it is with the vegetation as other influences from the environment such as water content, soil properties and geology impact the spectral signature in addition to vegetation properties. The k-means cluster analysis in the exploratory phase of the research found 50 % accuracy with the train-test data (Appendix D, however, the maps outputs were much less accurate with the higher number of classes based on fieldwork.

6.2 Using PSM to Predict PFTs

As the whole random forest prediction from the first objective was used as the training and testing data, the chances of higher accuracy should have been increased. However, maximum accuracy (random forest classification using the velocity, and mean amplitude, peak timing and trough timing with a training size of 70 %) of 53 % with a high standard deviation of 0.52. This accuracy is low, but high enough to encourage further refinement of the methodology of this novel analysis, which was pretty crude due to the timescale of this project. The other machine learning classifications were less successful, as with the PFT analysis, suggesting that random forests are the best classification method for this data. Key issues to overcome are differences in spatial resolution, how to scale the data, projection issues and the number of classes.

The resolution of the PSM data was much coarser than the PFT data, causing scale-associated vulnerability (Marshall et al. 2022), meaning that scaling was required. In this study, the PSM data upscaled to 5 m, with each new pixel being allocated the same value as the others in the larger 80x90 m area. This could be improved by taking considering other nearby values as it is likely that the boundaries between the initial pixels are much starker than in reality where there would be gradual change. An alternative approach potential approach is to degrade the PFT information to the larger pixel size. This would require either the dominant PFT within that area to be chosen, or new broader PFTs created. To develop this, it would be better to use larger sites, otherwise there would be limited train-test data. The resolution difference affects the ability to analyse the boundaries between PFTs. Many of the PFTs appeared to have distinct boundaries between them in the field, especially those that do not contain *Sphagnum*, although most of these boundaries could not be assessed within the 80 x 90 m pixel. This is further exacerbated by the less distinctive boundaries between shrub and *Sphagnum* dominated landscapes, which is a key area to focus on (Marshall et al. 2022), therefore a finer spatial resolution is required for both the PSM and hyperspectral data.

The two datasets used different coordinate reference systems (WGS 84 and OSGB36) causing projection issues. Although this was improved by using ArcGIS to project the data and perform geometric corrections on the PSM data, it was challenging to line up the points correctly, especially due to the different spatial resolutions. This could have reduced the accuracy and would need to be improved in future analysis.

To improve accuracy, it would be worth focusing on sites with less variation, such as Cross Lochs, to demonstrate whether it is feasible to relate the two datasets, trialling different methods to test for accuracy, then retrying for sites with more complex vegetation systems, before extending across the Flow Country. It would also be worth analysing the PSM data in more depth to extract the key attributes that are more likely to link to vegetation types and then defining PFTs based on broader categories (shrub-based, *Sphagnum*-based and grass-based, however this would be challenging due to the layering of the species). The PSM analysis undertaken by Marshall et al. (2022) demonstrates that PSM varies at multiple scales with hummocks being more dynamic than lawns which are more dynamic than hollows, however, the coarseness of the data prevents detailed analysis of this. If the PFT data could be related to the PSM data more accurately, this would enable additional analysis, especially in areas where there is poorer understanding (in the less dynamic parts of the peatland) (Marshall et al. 2022). The microtopographies also affect the timing of maximum seasonal swelling (Alshammari et al. 2020, Bradley et al. 2021). Therefore, hyperspectral data could be used to assess areas more prone to swelling in the summer, although there would be limitations with this as most swelling occurs from August to November (Alshammari et al. 2018). One aspect of this is the impact of drought; it would be worth removing these values from the dataset (especially in pool areas) as the PSM was underpredicted by up to 42 mm/yr (Marshall et al. 2022), which would affect the machine learning accuracies and outcomes. More fieldwork would also be beneficial, recording temporal changes in PFTs (with the same timings as the PSM data) and other factors such as water level, microtopography and erosion rates as the combination of these factors affect PSM (Marshall et al. 2022, Marshall et al. 2021). Water level directly corresponds to carbon accumulation/loss and is strongly affected by erosion rates and microtopographic changes, which also affect vegetation assemblage (Bradley et al. 2021).

The use of higher spatial resolution hyperspectral data (less than 1 m) would create species-specific spectral libraries, enabling more detailed analysis of the landscape. This would enable better understanding of which species dominate the landscape and how they interact with each other, including the sharpness of boundaries between different species and at this scale, the relationship of vegetation with microtopography could be analysed. However, if UAVs are used, they need to be suitable for the task; they should not contain reflectance panels, otherwise part of the vegetation signature will be missed (Honkavaara et al. 2016).

The accuracy of the PSM varies depending on the topography and as the InSAR data has a spatial resolution of 80 x 90 m, the impact of microtopography cannot be assessed, although it does have an impact, especially where there microtopographies are more variable (Marshall et al. 2022). These microtopographies also (objective two). Landscape management also affects PSM (Marshall et al. 2021), with peatland affected to different extents by erosion. InSAR analysis has demonstrated that subsidence rates are high in areas with plantations, clear felled forest and areas of bare peat (Alshammari et al. 2018).

6.3 Applications

The PFT data can not only be used in conjunction with aiming to improve understanding of PSM outputs, but also peat health and to compare with other peatlands in terms of vegetation assemblage for sites under a range of conditions. If time-series PFT data is generated, then there is the potential to assess how the sites are responding to climatic changes and restoration projects (Lavorel et al. 2007). It is likely that if temperatures increase and precipitation reduces, 'shrubification' of the landscape will occur as species with longer, finer root systems extending their range (Malhotra et al. 2020). In turn each PFT affects carbon fluxes in the peatland, with emissions from shrubs, grasses and mosses increasing with higher temperatures, however, lower levels of precipitation, and therefore water table, cause reductions in methane emissions, but higher carbon dioxide emissions (Whitaker et al. 2021). Peat under graminoids (grasses) emit more methane than bryophytes (mosses) or ericoid (shrub) PFTs (Whitaker et al. 2021), therefore monitoring localised extent of grasses is key to improving understanding carbon fluxes in the Flow Country.

Currently, the PFT-PSM analysis has little application, however, with improvements to the method, could be used in conjunction with PSM analysis undertaken in the

Flow Country by Alshammari et al. (2020), Alshammari et al. (2018), Bradley et al. (2021), Marshall et al. (2022), and Marshall et al. (2021). The vegetation assemblage is a key factor which affects PSM and although there is understanding about the impact in pool systems dominated by *Sphagnum* species, the limitations of InSAR data mean that shrub dominated environments are less well understood (Marshall et al. 2022). Therefore, improvements in accuracy could lead to better understanding of the less dynamic parts of the landscape and the rate of change between different PFTs.

6.4 Future Research

6.4.1 PFT-focused

When developing the machine learning predictions for the PFTs, rather than focusing on random forest (or the highest accuracy outcomes), outputs can be combined into an ensemble (Goos et al. 1998) and with each prediction from each classifier considered as a vote for a specific class (Khurram Shahzad and Lavesson 2012). This could be tested using the majority, conservative, comparative and veto voting rules to determine which generates the most accurate outputs, with the majority strategy generally considered to be the most effective with the class with the most votes producing the outcome (Lam and Suen 1997). An alternative option would be Bayesian formulation which determines the probabilities of an outcome using notions of control (Huys and Dayan 2009). In addition to the supervised machine learning classifiers used, a partial least-squares regression could be applied to the data, if higher spatial resolution were collected. These outcomes could then be compared to other studies such as Cole et al. (2013).

To improve accuracy, more pre-processing should be undertaken of the hyperspectral data and additional fieldwork carried out across all sites, with more quadrats at each to gain a greater clarity of PFTs and how they vary between sites. This would be improved through the use of higher spatial resolution data which would enable the mapping of specific species. This should increase the accuracy of the train-test data, increasing accuracy of the predictions, both within and between sites.

When collecting field data in the future, accuracy regarding the location of field sites could be improved by considering tectonic movement when using a GNSS receiver. This is becomes increasingly important the higher the spatial resolution.

6.4.2 PFT and PSM

There is the potential to develop new methods as discussed above regarding the use of PFT data to add more understanding to the PSM outcomes, with a range of strategies to increase accuracy of outcomes and relate to more aspects of the environment including microtopography, water abundance and erosion rates. These can be measured using hyperspectral data collected with a handheld spectroradiometer or UAV with a high spatial resolution. In addition to PFTs, vegetation indices such as NDVI, PRI and cellulose absorption index to determine which spectral indices are most important spatially and potentially temporally (Cole et al. 2014).

With drought conditions expected to increase in the future, the impact of this could be analysed by collecting timeseries hyperspectral data to assess the relationship between drought, PFTs (or specific species extent) and PSM attributes. Time series hyperspectral data would also be valuable in when comparing with PSM outcomes generally to assess seasonal impacts of changes in vegetation on motion. However, it would also be worth taking the 2018 drought into account, as the underestimation of the PSM would reduce the accuracy of the data. Therefore, the removal of drought data from the PSM timeseries could improve the accuracies of the PSM-PFT analysis.

6.4.3 Wider Reaching Research

Other research could be undertaken to compare hyperspectral outcomes with other analysis, such as DOC outcomes in the Flow Country or compare with hyperspectral analysis in other peatlands. DOC The hyperspectral data could be used to retrieve DOC and compare with current results (Muller and Tankéré-Muller 2012, Vinjili 2012), before being extended over a larger area to assess the extent of leaching. Peat properties such as moisture content and humification could be compared with studies by Cole et al. (2013), Lees et al. (2020), J. McMorrow et al. (2014) and J. McMorrow et al. (2005). The hyperspectral data could also be compared to Sentinel 2 10 m resolution data to assess whether the higher 5 m resolution hyperspectral data gives a significant advantage or whether the multitemporal Sentinel data is better (however, there are issues with high levels of cloud cover in the Flow Country, reducing the frequency of useful data). The hyperspectral data could be degraded to 10 m and then classified and Sentinel 2 data upscaled.

Conclusions

Random forests can be used to classify plant functional types using AVIRIS-NG data collected over the Flow Country. Accuracy of random forest classifications are comparable to other studies, especially in the analysis conducted on the first restoration site with the fieldwork-focused train-test data. Accuracies were comparable to other studies classifying data into PFTs or vegetation species in terms of both overall accuracy and producer's accuracy. Differences in overall accuracy could be attributed to differences in spatial resolution, the number of classes and the complexity of the vegetation assemblages, with differences in producer's accuracy linked to the spectral signatures of different species/PFTs and the clarity of the classes. A key limiting factor of this study was the spatial resolution of 5 m, with better accuracies found when data with a resolution of 1 m or better were used. This coarse resolution especially affects the prediction of PFTs which interlink with others and can take up small amounts of space, such as pools. Whereas, spatial resolution was much less of an issue for PFTs which dominated a larger area such as the long and short grasses.

Accuracies could also be affected by changes in PFTs between July 2021 and July 2022 due to different precipitation levels and temperature maxima/minima between the two years, meaning that the fieldwork and satellite images would not reflect the conditions found in 2021 when the hyperspectral data was collected. These are unlikely to have a significant impact on predictions, especially at Cross Lochs which is unlikely to undergo much change year to year. However, it is expected that restoration sites change year-to-year, especially the second restoration site as tree felling has only occurred within the past ten years, increasing the potential impact on predictions.

Despite accuracy being high, there could be issues regarding over- and under-estimation of PFTs and confusion between PFTs with similar spectral signatures such as those containing *Sphagnum* mosses, meaning that high accuracy values does not necessarily mean that reality is reflected. This is supported by the outcome of the McNemar similarity test which demonstrated that there was a statistically significant difference between the focused and general train-test predictions for the first restoration site. More fieldwork needs to be undertaken across the sites to improve this issue, with new confusion matrices and McNemar similarity tests to improve PFT identification. The PFTs across the sites were mainly as anticipated, however, either from observations in the field, satellite image interpretation or from expert description of the sites.

The data used to train all four sites resulted in slightly lower accuracies than the focused training data. This demonstrates that PFTs may be slightly different across the sites, for instance the Shrub and *Sphagnum* PFT will exist at all sites, however, the species composition of this class could vary at different locations. This is likely due to different conditions linked to human activity (whether the site is near-natural, undergoing restoration or eroded), the presence of large herbivores and the local physical conditions (topography, geology and water levels).

Unlike random forest classification, k-means clusters cannot be used to determine PFTs, especially with a high number of classes, due to the complexity of the environment. Therefore, although supervised learning classifications can be used to predict PFTs, unsupervised learning cannot.

To ameliorate the classification outcomes, improvements in the pre-processing of the hyperspectral data are required, as well as more time spent collecting field data across all four sites. It would also be worth reviewing the PFTs and performing machine learning analysis using a range of classes, potentially merging groups that are more likely to be confused, such as the grass and *Sphagnum*, and shrub and *Sphagnum* PFTs.

Machine learning has the potential to be used to classify PFTs using PSM attributes, however, improvements need to be made to the method to increase accuracy and reduce standard deviation. Increased pre-processing and processing is required prior to prediction with more predictions generated to determine the best strategy. It is essential to improve projection matching between the two datasets so that they match and trial different ways of scaling the data - both degrading the PFT data and upscaling the PSM data, whilst using a combination of values to determine the PSM rather than just taking the value of the 80x90 m pixel. Different numbers of PFTs should be trialled with similar PFTs grouped together, in terms of spectral signature and the nature of the bog (how dynamic it is). More fieldwork is also required to improve PFT accuracy prior to analysis with the PSM data and measurements should also be collected regarding the water level in the peat. It would also be worth collecting higher resolution spectral data to develop more accurate spectral libraries regarding the vegetation assemblage, ensuring that the dominant species across the sites can be determined. This links specifically to the water content of the peat, erosion rates and microtopography, which interrelate with the PSM to add to understanding from previous studies.

Once improvements have been made to the methods, the application of this research will increase, enabling analysis to be undertaken in conjunction with current research on PSM in the Flow Country, especially in areas less understood by outcomes of this research. Further research should, therefore, focus on the vegetation assemblage of the less dynamic parts of the ecosystem (shrub-based environments).

Regarding the overall aim, hyperspectral data has the potential to be used to understand the association between peatland surface motion (as measured by InSAR data) and land cover in the Flow Country. However, further work is required to achieve this, with the use of higher spatial resolution data, better scaling and projections and improvements made to pre-processing.

Bibliography

Agresti, A. 2006. *An Introduction to Categorical Data Analysis Second Edition.* Technical report.

Aitkenhead, M, and M Coull. 2020. "Mapping soil profile depth, bulk density and carbon stock in Scotland using remote sensing and spatial covariates." *European Journal of Soil Science* 71, no. 4 (July): 553–567. ISSN: 13652389. https://doi.org/10.1111/ejss.12916.

Alshammari, L, D.S Boyd, A Sowter, C Marshall, R Andersen, P Gilbert, S Marsh, and D.J Large. 2020. "Use of Surface Motion Characteristics Determined by InSAR to Assess Peatland Condition." *Journal of Geophysical Research: Biogeosciences* 125, no. 1 (January). ISSN: 21698961. https://doi.org/10.1029/2018JG004953.

Alshammari, L, D.J Large, D.S Boyd, A Sowter, R Anderson, R Andersen, and S Marsh. 2018. "Long-term peatland condition assessment via surface motion monitoring using the ISBAS DInSAR technique over the Flow Country, Scotland." *Remote Sensing* 10, no. 7 (July). ISSN: 20724292. https://doi.org/10.3390/rs10071103.

Anand, R, S Veni, P Geetha, and S Rama-Subramoniam. 2021. "Extended morphological profiles analysis of airborne hyperspectral image classification using machine learning algorithms." *International Journal of Intelligent Networks* 2:1–6. ISSN: 26666030. https://doi.org/10.1016/j.ijin.2020.12.006.

ARES. 2021. *Airborne Research of the Earth System.* https://ares-observatory.ch/esa_chime_mission_2021/.

Artz, R, M Coyle, G Donaldson-Selby, and R Morrison. 2021. "Net Carbon Dioxide Emission from An Eroding Atlantic Blanket Bog," https://doi.org/10.21203/rs.3.rs-991712/v1.

Bonn, A, M.S Reed, C.D Evans, H Joosten, C Bain, J Farmer, I Emmer, J Couwenberg, A Moxey, R Artz, F Tanneberger, M von Unger, M.A Smyth, and D Birnie. 2014. "Investing in nature: Developing ecosystem service markets for peatland restoration." *Ecosystem Services* 9 (September): 54–65. ISSN: 22120416. https://doi.org/10.1016/j.ecoser.2014.06.011.

Bradley, A.V, R Andersen, C Marshall, A Sowter, and D.J Large. 2021. "Identification of typical eco-hydrological behaviours using InSAR allows landscape-scale mapping of peatland condition." *Earth Surface Dynamics* 10 (2): 261–277. https://doi.org/10.5194/esurf-2021-58. https://doi.org/10.5194/esurf-2021-58.

Cameron, L. 2019. *Peatland fire may have released six days of carbon | Scotland | The Times.* https://www.thetimes.co.uk/article/peatland-fire-may-have-released-six-days-of-carbon-3k9pbkvl7.

Carless, D, D.J Luscombe, N Gatis, K Anderson, and R.E Brazier. 2019. "Mapping landscape-scale peatland degradation using airborne lidar and multispectral data." *Landscape Ecology* 34, no. 6 (June): 1329–1345. ISSN: 15729761. https://doi.org/10.1007/s10980-019-00844-5.

Chi, M, A Plaza, J.A Benediktsson, Z Sun, J Shen, and Y Zhu. 2016. "Big Data for Remote Sensing: Challenges and Opportunities." *Proceedings of the IEEE* 104, no. 11 (November): 2207–2219. ISSN: 15582256. https://doi.org/10.1109/JPROC.2016.2598228.

Cole, B, M Evans, and J McMorrow. 2014. "Spectral monitoring of moorland plant phenology to identify a temporal window for hyperspectral remote sensing of peatland." *ISPRS Journal of Photogrammetry and Remote Sensing* 90:49–58. ISSN: 09242716. https://doi.org/10.1016/j.isprsjprs.2014.01.010.

Cole, B, J McMorrow, and M Evans. 2013. "Empirical modelling of vegetation abundance from airborne hyperspectral data for upland peatland restoration monitoring." *Remote Sensing* 6 (1): 716–739. ISSN: 20724292. https://doi.org/10.3390/rs6010716.

EO College. 2021. *Spectral reflectance of selected surface materials – EO College.* https://eo-college.org/courses/beyond-the-visible/lessons/principles-of-imaging-spectroscopy/topic/spectral-reflectance-of-selected-surface-materials/.

Erudel, T, S Fabre, T Houet, F Mazier, and X Briottet. 2017. "Criteria Comparison for Classifying Peatland Vegetation Types Using In Situ Hyperspectral Measurements." *Remote Sensing* 9, no. 7 (July): 748. https://doi.org/10.3390/rs9070748.

ESA. *Geographical Coverage - Sentinel-1 - Sentinel Online - Sentinel Online.* https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1/satellite-description/geographical-coverage.

Ferretto, A, R Brooker, M Aitkenhead, R Matthews, and P Smith. 2019. "Potential carbon loss from Scottish peatlands under climate change." *Regional Environmental Change* 19, no. 7 (October): 2101–2111. ISSN: 1436378X. https://doi.org/10.1007/s10113-019-01550-3.

Ferretto, A, R Brooker, R Matthews, and P Smith. 2021. "Climate change and drinking water from Scottish peatlands: Where increasing DOC is an issue?" *Journal of Environmental Management* 300 (December). ISSN: 10958630. https://doi.org/10.1016/j.jenvman.2021.113688.

Foody, G. 2008. "Harshness in image classification accuracy assessment." *International Journal of Remote Sensing* 29, no. 11 (June): 3137–3158. ISSN: 13665901. https://doi.org/10.1080/01431160701442120.

Foody, G.M. 2004. "Thematic Map Comparison." *Photogrammetric Engineering & Remote Sensing* 70, no. 5 (May): 627–633. ISSN: 00991112. https://doi.org/10.14358/PERS.70.5.627.

Gallego-Sala, A.V, and C.I Prentice. 2013. "Blanket peat biome endangered by climate change." *Nature Climate Change* 3, no. 2 (February): 152–155. ISSN: 1758678X. https://doi.org/10.1038/nclimate1672.

Garnett, M.H, P Ineson, and A.C Stevenson. 2000. *Effects of burning and grazing on carbon sequestration in a Pennine blanket bog, UK.* Technical report.

GISG, GISGeography. 2021. *Hyperspectral Imaging in Space.* Technical report.

Goos, G, J Hartmanis, J Van, L Board, D Hutchison, T Kanade, J Kittler, J.M Kleinberg, F Mattern, E Zurich, J.C Mitchell, M Naor, O Nierstrasz, B Steffen, M Sudan, D Terzopoulos, D Tygar, M.Y Vardi, and G Weikum. 1998. *LNCS 3686 - Pattern Recognition and Data Mining.* Technical report.

Granlund, L, M Keinänen, and T Tahvanainen. 2021. "Identification of peat type and humification by laboratory VNIR/SWIR hyperspectral imaging of peat profiles with focus on fen-bog transition in aapa mires." *Plant and Soil* 460, nos. 1-2 (March): 667–686. ISSN: 15735036. https://doi.org/10.1007/s11104-020-04775-y.

Granlund, L, V Vesakoski, A Sallinen, T.H.M Kolari, F Wolff, and T Tahvanainen. 2021. "Recent Lateral Expansion of Sphagnum Bogs Over Central Fen Areas of Boreal Aapa Mire Complexes," https://doi.org/10.1007/s10021-021-0072. https://doi.org/10.1007/s10021-021-0072.

Hambley, G. 2016. *The Effect of Forest-to-Bog Restoration on Net Ecosystems Exchange in Flow Country Peatlands.* Technical report. http://research-repository.st-andrews.ac.uk/.

Hambley, G, R Andersen, P Levy, M Saunders, N.R Cowie, Y.A. Teh, and T.C. Hill. 2018. "Net ecosystem exchange from two formerly afforested peatlands undergoing restoration in the flow country of northern Scotland." *Mires and Peat* 23. ISSN: 1819754X. https://doi.org/10.19189/MaP.2018.DW.346.

Hancock, M.H, B England, and N.R Cowie. 2018. "Knockfin Heights: A high-altitude flow country peatland showing extensive erosion of uncertain origin." *Mires and Peat* 23. ISSN: 1819754X. https://doi.org/10.19189/MaP.2018.OMB.334.

Harris, A, R.G Bryant, and A.J Baird. 2005. "Detecting near-surface moisture stress in Sphagnum spp." *Remote Sensing of Environment* 97, no. 3 (August): 371–381. ISSN: 00344257. https://doi.org/10.1016/j.rse.2005.05.001.

Hati, P.J, S Samanta, R.N Chaube, A Misra, S Giri, N Pramanick, K Gupta, S Datta Majumdar, A Chanda, A Mukhopadhyay, and S Hazra. 2021. "Mangrove classification using airborne hyperspectral AVIRIS-NG and comparing with other spaceborne hyperspectral and multispectral data." *Egyptian Journal of Remote Sensing and Space Science* 24, no. 2 (August): 273–281. ISSN: 20902476. https://doi.org/10.1016/j.ejrs.2020.10.002.

He, K.S, D Rocchini, M Neteler, and H Nagendra. 2011. *Benefits of hyperspectral remote sensing for tracking plant invasions,* 3, May. https://doi.org/10.1111/j.1472-4642.2011.00761.x.

Honkavaara, E, M.A Eskelinen, I Polonen, H Saari, H Ojanen, R Mannila, C Holmlund, T Hakala, P Litkey, T Rosnell, N Viljanen, and M Pulkkanen. 2016. "Remote Sensing of 3-D Geometry and Surface Moisture of a Peat Production Area Using Hyperspectral Frame Cameras in Visible to Short-Wave Infrared Spectral Ranges Onboard a Small Unmanned Airborne Vehicle (UAV)." *IEEE Transactions on Geoscience and Remote Sensing* 54, no. 9 (September): 5440–5454. ISSN: 01962892. https://doi.org/10.1109/TGRS.2016.2565471.

Humpenöder, F, K Karstens, H Lotze-Campen, J Leifeld, L Menichetti, A Barthelmes, and A Popp. 2020. "Peatland protection and restoration are key for climate change mitigation." *Environmental Research Letters* 15, no. 10 (October). ISSN: 17489326. https://doi.org/10.1088/1748-9326/abae2a.

Huys, Q.J.M, and P Dayan. 2009. "A Bayesian formulation of behavioral control." *Cognition* 113, no. 3 (December): 314–328. ISSN: 00100277. https://doi.org/10.1016/j.cognition.2009.01.008.

Jia, J, Y Wang, J Chen, R Guo, R Shu, and J Wang. 2020. *Status and application of advanced airborne hyperspectral imaging technology: A review,* January. https://doi.org/10.1016/j.infrared.2019.103115.

Kale, K.V, M.M Solankar, D.B Nalawade, R.K Dhumal, and H.R Gite. 2017. *A Research Review on Hyperspectral Data Processing and Analysis Algorithms,* 4, December. https://doi.org/10.1007/s40010-017-0433-y.

Kennedy, G. W, and J. S Price. 2005. "A conceptual model of volume-change controls on the hydrology of cutover peats." *Journal of Hydrology* 302, nos. 1-4 (February): 13–27. ISSN: 00221694. https://doi.org/10.1016/j.jhydrol.2004.06.024.

Kerr, J.T, and M Ostrovsky. 2003. "From space to species: Ecological applications for remote sensing." *Trends in Ecology and Evolution* 18, no. 6 (June): 299–305. ISSN: 01695347. https://doi.org/10.1016/S0169-5347(03)00071-5.

Khurram Shahzad, R, and N Lavesson. 2012. *Comparative Analysis of Voting Schemes for Ensemble-based Malware Detection *.* Technical report.

Kopeć, D, A Sabat-Tomala, D Michalska-Hejduk, A Jarocińska, and J Niedzielko. 2020. "Application of airborne hyperspectral data for mapping of invasive alien Spiraea tomentosa L.: a serious threat to peat bog plant communities." *Wetlands Ecology and Management* 28, no. 2 (April): 357–373. ISSN: 15729834. https://doi.org/10.1007/s11273-020-09719-y.

Kozma-Bognár, V, and J Berke. 2010. *New Evaluation Techniques of Hyperspectral Data.* Technical report. https://www.researchgate.net/publication/264423679.

Kumar, A, J Yadav, and R Mohan. 2020. "Global warming leading to alarming recession of the Arctic sea-ice cover: Insights from remote sensing observations and model reanalysis." *Heliyon* 6, no. 7 (July). ISSN: 24058440. https://doi.org/10.1016/j.heliyon.2020.e04355.

Lam, L, and C.Y Suen. 1997. *Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance.* Technical report 5.

Lavorel, S, S Díaz, J.H.C Cornelissen, E Garnier, S.P Harrison, S McIntyre, J.G Pausas, N Pérez-Harguindeguy, C Roumet, and C Urcelay. 2007. "Plant Functional Types: Are We Getting Any Closer to the Holy Grail?" Chap. 13 in *Terrestrial Ecosystems in a Changing World,* 149–164. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-32730-1{_}13.

Lees, K. 2019. *Measuring peatland carbon uptake by remote sensing.* Technical report.

Lees, K.J, R.R.E Artz, M Khomik, J.M Clark, J Ritson, M.H Hancock, N.R Cowie, and T Quaife. 2020. "Using spectral indices to estimate water content and GPP in sphagnum moss and other peatland vegetation." *IEEE Transactions on Geoscience and Remote Sensing* 58, no. 7 (July): 4547–4557. ISSN: 15580644. https://doi.org/10.1109/TGRS.2019.2961479.

Levy, P.E, and A Gray. 2015. "Greenhouse gas balance of a semi-natural peatbog in northern Scotland." *Environmental Research Letters* 10, no. 9 (September). ISSN: 17489326. https://doi.org/10.1088/1748-9326/10/9/094019.

Linden, S van der, A Rabe, M Held, B Jakimow, PJ Leitão, A Okujeni, M Schwieder, S Suess, and P Hostert. 2015. "The EnMAP-box-A toolbox and application programming interface for EnMAP data processing." *Remote Sensing* 7 (9): 11249–11266. ISSN: 20724292. https://doi.org/10.3390/rs70911249.

Lindsay, R.A, D.J Charman, F Everingham, R.M O'reilly, M.A Palmer, T.A Rowell, D.A Stroud, D.A Ratcliffe, and P.H Oswald. 1988. *The Flow Country - The peatlands of Caithness and Sutherland.* Technical report. http://www.jncc.gov.uk/page-4281.

Lunt, P.H, R.M Fyfe, and A.D Tappin. 2019. "Role of recent climate change on carbon sequestration in peatland systems." *Science of the Total Environment* 667 (June): 348–358. ISSN: 18791026. https://doi.org/10.1016/j.scitotenv.2019.02.239.

Luo, G, G Chen, L Tian, K Qin, and S.E Qian. 2016. "Minimum Noise Fraction versus Principal Component Analysis as a Preprocessing Step for Hyperspectral Imagery Denoising." *Canadian Journal of Remote Sensing* 42, no. 2 (March): 106–116. ISSN: 17127971. https://doi.org/10.1080/07038992.2016.1160772.

Malhotra, A, D.J Brice, J Childs, J.D Graham, E.A Hobbie, H.V Stel, S.C Feron, P.J Hanson, and C.M Iversen. 2020. "Peatland warming strongly increases fine-root growth," https://doi.org/10.25581/spruce.077/1607860. https://doi..

Marsden, K, and S Ebmeier. 2012. *The Sc ottish Parliament and Scottis h Parliament Infor mation C entre l ogos. Peatlands and Climate Change.* Technical report.

Marshall, C, A.V Bradley, R Andersen, and D.J Large. 2021. "nature.scot-NatureScot Research Report 1269 - Using peatland surface motion bog breathing to monitor Peatland Act." *NatureScot Research Report* 1269.

Marshall, C, H.P Sterk, P.J Gilbert, R Andersen, A.V Bradley, A Sowter, S Marsh, and D.J Large. 2022. "Multiscale Variability and the Comparison of Ground and Satellite Radar Based Measures of Peatland Surface Motion for Peatland Monitoring." *Remote Sensing* 14, no. 2 (January). ISSN: 20724292. https://doi.org/10.3390/rs14020336.

Mather, P, and B Tso. 2009. *42 Classification Methods for Remotely Sensed Data.* Technical report. http://ebookcentral.proquest.com/lib/nottingham/detail.action?docID=570488..

McMorrow, J, M Cutler, A Al-Roichi, and M Evans. 2005. *HYPERSPECTRAL REMOTE SENSING OF PEAT HUMIFICATION.* Technical report.

McMorrow, J, M.E.J Cutler, A. Al-Roichdi, M.G Evans, and M.E Cutler. 2014. *The effect of moisture content and humification on the hyperspectral reflectance of peat.* Technical report. https://www.researchgate.net/publication/267220811.

McMorrow, J. M, M.E.J Cutler, M. G Evans, and A Al-Roichdi. 2004. "Hyperspectral indices for characterizing upland peat composition." *International Journal of Remote Sensing* 25, no. 2 (January): 313–325. ISSN: 01431161. https://doi.org/10.1080/0143116031000117065.

Mesibov, R. 2012. "Known unknowns, Google Earth, plate tectonics and Mt Bellenden Ker: Some thoughts on locality data." *ZooKeys* 247:61–67. ISSN: 13132989. https://doi.org/10.3897/zookeys.247.4195.

Met Office. 2022. *Historic station data.* https://www.metoffice.gov.uk/research/climate/maps-and-data/historic-station-data.

Muller, F.L.L, and S.P.C Tankéré-Muller. 2012. "Seasonal variations in surface water chemistry at disturbed and pristine peatland sites in the Flow Country of northern Scotland." *Science of the Total Environment* 435-436 (October): 351–362. ISSN: 00489697. https://doi.org/10.1016/j.scitotenv.2012.06.048.

Mutanga, O, and A.K Skidmore. 2007. "Red edge shift and biochemical content in grass canopies." *ISPRS Journal of Photogrammetry and Remote Sensing* 62, no. 1 (May): 34–42. ISSN: 09242716. https://doi.org/10.1016/j.isprsjprs.2007.02.001.

NASA. 2022. *AVIRIS-Next Generation.* https://avirisng.jpl.nasa.gov/data_processing.html.

NASA JPL. 2022. *JPL | AVIRIS-NG Data Portal.* https://avirisng.jpl.nasa.gov/dataportal/.

Nordin, S.A, Z Abd Latif, and H Omar. 2019. "Individual tree crown segmentation in tropical peat swamp forest using airborne hyperspectral data." *Geocarto International* 34 (11): 1218–1236. ISSN: 10106049. https://doi.org/10.1080/10106049.2018.1475511.

Pal, M. 2005. "Random forest classifier for remote sensing classification." *International Journal of Remote Sensing* 26 (1): 217–222. ISSN: 1366-5901. https://doi.org/10.1080/01431160412331269698. https://www.tandfonline.com/action/journalInformation?journalCode=tres20.

Parry, L.E, J Holden, and P.J Chapman. 2014. *Restoration of blanket peatlands,* January. https://doi.org/10.1016/j.jenvman.2013.11.033.

Patil, H, and A Dwivedi. 2021. "Prediction of properties of the cement incorporated with nanoparticles by principal component analysis (PCA) and response surface regression (RSR)." *Materials Today: Proceedings* 43 (January): 1358–1367. ISSN: 2214-7853. https://doi.org/10.1016/J.MATPR.2020.09.170.

Pradhan, T, V Walia, R Kapoor, and S Saran. 2014. "Optimizing land use classification using decision tree approaches." In *2014 International Conference on Data Mining and Intelligent Computing, ICDMIC 2014.* Institute of Electrical / Electronics Engineers Inc., November. ISBN: 9781479946754. https://doi.org/10.1109/ICDMIC.2014.6954256.

Priyadarshini, N.K, V Sivashankari, S Shekhar, and K Balasubramani. 2019. "Comparison and Evaluation of Dimensionality Reduction Techniques for Hyperspectral Data Analysis," 6. MDPI AG, June. https://doi.org/10.3390/iecg2019-06209.

Provost, F, and T Fawcett. 2013. *Data Science for Business : What You Need to Know about Data Mining and Data-Analytic Thinking.* 1st ed. Sebastopol: O'Reilly Media, Incorporated.

Quinn, J.A, M.M Nyhan, C Navarro, D Coluccia, L Bromley, and M Luengo-Oroz. 2018. "Humanitarian applications of machine learning with remote-sensing data: Review and case study in refugee settlement mapping." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2128). ISSN: 1364503X. https://doi.org/10.1098/rsta.2017.0363.

Räsänen, A, S Juutinen, E.S Tuittila, M Aurela, and T Virtanen. 2019. "Comparing ultra-high spatial resolution remote-sensing methods in mapping peatland vegetation." *Journal of Vegetation Science* 30, no. 5 (September): 1016–1026. ISSN: 16541103. https://doi.org/10.1111/jvs.12769.

Ratcliffe, J. 2015. "Carbon Accumulation Rates Over the Holocene in Flow Country Peatlands and the Direct Comparison of Open and Afforested Peatland Carbon Stocks Using Tephrochronology," https://doi.org/10.13140/RG.2.1.1960.3288. https://www.researchgate.net/publication/286919687.

Ratcliffe, J, R Andersen, R Anderson, A Newton, D Campbell, D Mauquoy, and R Payne. 2018. "Contemporary carbon fluxes do not reflect the long-term carbon balance for an Atlantic blanket bog." *Holocene* 28, no. 1 (January): 140–149. ISSN: 14770911. https://doi.org/10.1177/0959683617715689.

Salisbury, F.B, and C.W Ross. 1992. *Plant Physiology.* CA, USA: Wadworth: Belmont.

SEPA. 2022. *Scottish Rainfall Data for Forsinain - provided by Scottish Environment Protection Agency (SEPA).* https://www2.sepa.org.uk/rainfall//data/index/115340.

Shao, Z, L Ding, D Li, O Altan, M.E Huq, and C Li. 2020. "Exploring the relationship between urbanization and ecological environment using remote sensing images and statistical data: A case study in the Yangtze River Delta, China." *Sustainability (Switzerland)* 12, no. 14 (July). ISSN: 20711050. https://doi.org/10.3390/su12145620.

Singh, A, N Thakur, and A Sharma. 2016. "A review of supervised machine learning algorithms; A review of supervised machine learning algorithms." ISBN: 978-93-80544-20-5.

Stuart, M.B, A.J.S Mcgonigle, M Davies, M.J Hobbs, N.A Boone, L.R Stanger, C Zhu, T.D Pering, J.R Willmott, J Luis, N Gómez, and M.A Martínez-Domingo. 2021. "Low-Cost Hyperspectral Imaging with a Smartphone," https://doi.org/10.3390/jimaging. https://doi.org/10.3390/jimaging.

Tochon, G, J.B Féret, S Valero, R.E Martin, D.E Knapp, P Salembier, J Chanussot, and G.P Asner. 2015. "On the use of binary partition trees for the tree crown segmentation of tropical rainforest hyperspectral images." *Remote Sensing of Environment* 159 (March): 318–331. ISSN: 00344257. https://doi.org/10.1016/J.RSE.2014.12.020.

Ustin, S.L, and J.A Gamon. 2010. "Remote sensing of plant functional types." *New Phytologist* 186, no. 4 (June): 795–816. ISSN: 0028646X. https://doi.org/10.1111/J.1469-8137.2010.03284.X.

Vinjili, S. 2012. *Land Use Change and Organic Carbon Exports from a Peat Catchment of the Hallendale River in the Flow Country of Sutherland and Caithness, Scotland.* Technical report. http://research-repository.st-andrews.ac.uk/.

Whitaker, J, H.R Richardson, N.J Ostle, A Armstrong, and S Waldron. 2021. "Plant functional type indirectly affects peatland carbon fluxes and their sensitivity to environmental change." *European Journal of Soil Science* 72, no. 2 (March): 1042–1053. ISSN: 13652389. https://doi.org/10.1111/ejss.13048.

Whitehead, S, H Weald, and D Baines. 2021. "Post-burning responses by vegetation on blanket bog peatland sites on a Scottish grouse moor." *Ecological Indicators* 123 (April). ISSN: 1470160X. https://doi.org/10.1016/j.ecolind.2021.107336.

Woolley, J.T. 1971. *Reflectance and Transmittance of Light by Leaves.* Technical report.

Zhong, Y, X Wang, Y Xu, S Wang, T Jia, X Hu, J Zhao, L Wei, and L Zhang. 2018. "Mini-UAV-Borne Hyperspectral Remote Sensing: From Observation and Processing to Applications." *IEEE Geoscience and Remote Sensing Magazine* 6, no. 4 (December): 46–62. ISSN: 21686831. https://doi.org/10.1109/MGRS.2018.2867592.

Zhou, Z, Z Li, S Waldron, and A Tanaka. 2019. "InSAR time series analysis of L-band data for understanding tropical peatland degradation and restoration." *Remote Sensing* 11, no. 21 (November). ISSN: 20724292. https://doi.org/10.3390/rs11212592.

Zhou, Z, S Waldron, and Z Li. 2013. *Quantifying changes in land-surface height in bioenergy palm oil plantations (Sumatra) using InSAR time series | Request PDF.* https://www.researchgate.net/publication/255718866_Quantifying_changes_in_land-surface_height_in_bioenergy_palm_oil_plantations_Sumatra_using_InSAR_time_series.

Appendix A: Fieldwork Risk Assessment

TABLE A.1: Faculty of Engineering risk assessment, completed for fieldwork in the Flow Country

ACTIVITY FOR ASSESSMENT	ASSOCIATED HAZARD	PEOPLE INVOLVED	ASSOCIATED RISK	EXISTING RISK-CONTROL MEASURES
Name the activity, machine, equipment, substance in use, etc.	(something with the potential to cause harm) List the significant hazards associated with each subject e.g., electricity, moving parts, flammable liquid, hazardous materials, fumes, noise, dust, sharps, etc.	List the groups(s) of people who might be at risk	(the potential harm arising from the hazard) Give a brief statement for each hazard e.g. electrocution, crushed limbs, cut hand, eye damage, respiratory irritation, chronic illness etc.	List for each hazard the control procedures, equipment and devices used to control the risk
Extreme weather working	Extreme heat or cold / Rain or snow	Student & helper	Hypothermia / Sun stroke or sun burn/ Dehydration	Waterproof and warm clothing / Sun screen and hat /Footwear as required / Regular drinks
Extreme weather working	Thunderstorm	Student & helper		Do NOT WORK
Extreme weather working	Wildfire	Student & helper		DO NOT WORK
Transporting equipment to site	Manual handling or overloading/ Trip hazard	Student & helper	Strain, ankle/hand injuries from trips	Safety talk to discuss safe loading prior to trip
Working on uneven surfaces (peatland)	Trip hazard	Student & helper	Strain, ankle/hand injuries from trips	Appropriate clothing and equipment
Vicinity of other students and general public	Block paths	Student, helper, General public	Altercation through annoyance with public	Ensure equipment is not blocking pathways
Working in remote location	Isolation, poor emergency access	Student & helper	Getting isolated or lost	Have GPS to determine location, plenty of food and clothing layers
Lone working	Lack of communication should problems occur	Student & helper		Forbidden
Equipment use	Misuse of equipment	Student & helper	Crushed fingers, impact injuries	Clear instructions on equipment use and what is misuse
Drive to site (over 10 hours)	Collision	Student & helper	Variety of injuries	Take regular breaks, overnight stop on the way there and back in Edinburgh (about half way)

Appendix B: Data collection sheet

Site .... Date .... Conditions ....

Subsite: GPS: Start time (link to photos): **Species**	%	%	%	%	%	%	%	%	%	%
Sphagnum cuspidatum										
Sphagnum capillifolium										
Sphagnum papillosum										
Drosera rotundifolia										
Drosera anglica										
Erica cinerea										
Calluna vulgaris										
Erica tetralix										
Eriophorum angustifolium										
Myrica gale										
Narthecium ossifragum										
Menyanthes trifoliata										
Trientalis europaea										
Pinguicula vulgaris										
Dactylorhiza maculata										
Usnea spp.										
Water										
Additional notes										

FIGURE B.1: Data collection sheet for fieldwork, July 2022. The subsites were numbered and ordered prior to data collection, the GPS number was the record number on the GNSS receiver and the start time was included so that relevant photos could be linked. Species percentages were recorded with space for additional species as required and other site notes.

Appendix C: Histograms and correlation matrix

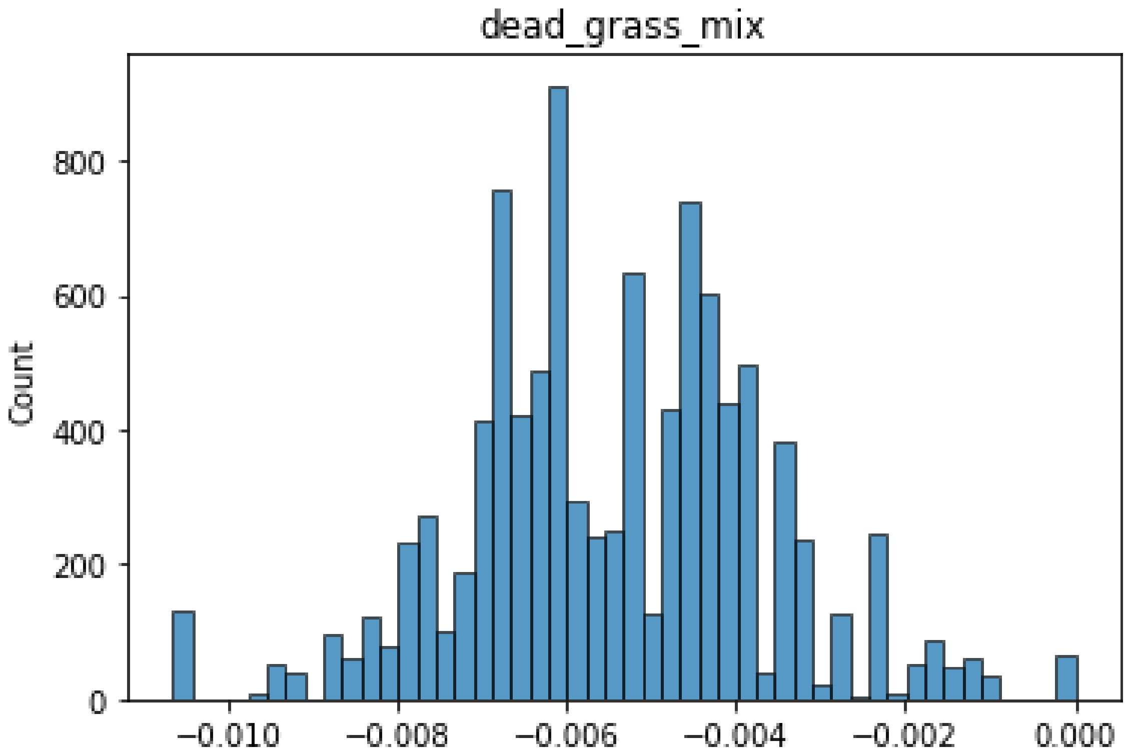


FIGURE C.1: Velocity histogram for the 'dead grass mix' PFT

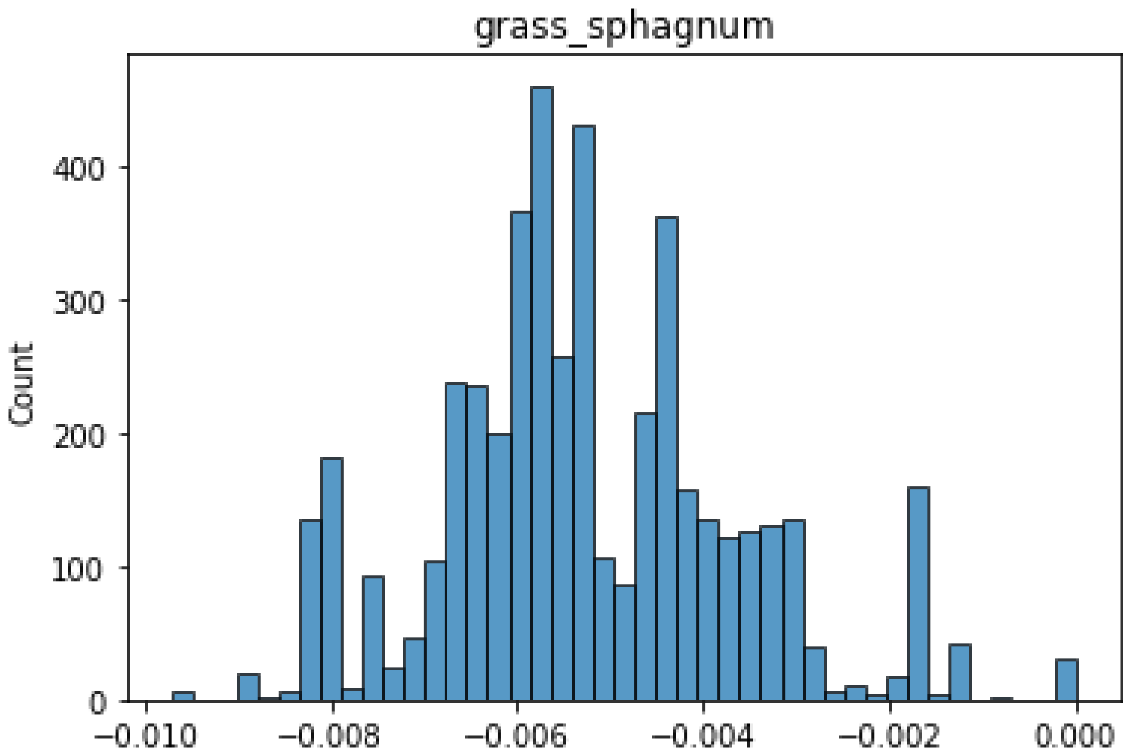


FIGURE C.2: Velocity histogram for the 'grass *Sphagnum*' PFT

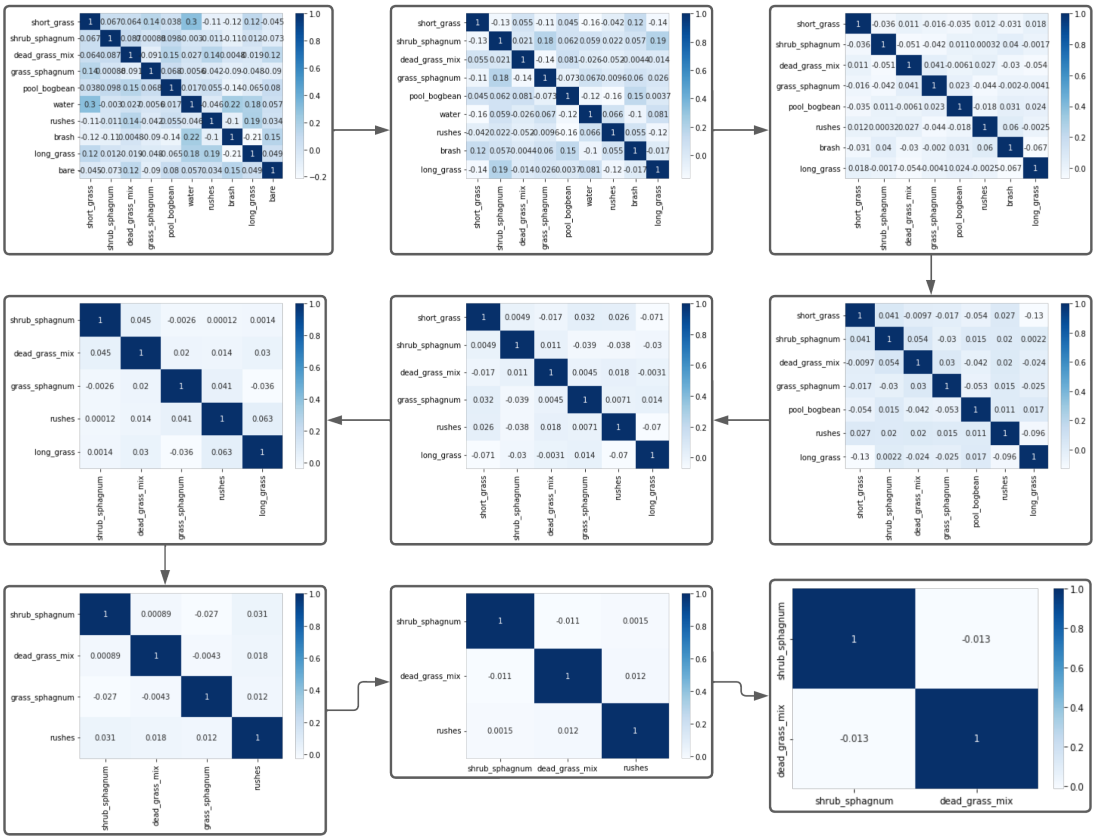


FIGURE C.3: Example correlation matrix, iterating through the PFTs (random points chosen from larger classes when a smaller class is the used as the basis of the matrix - smallest class removed with each iteration and determines the number of points for correlation)

Appendix D: Complete table of highest outcomes

TABLE D.1: Full table of all the first restoration site top outcomes (highest five for each iteration) using the focused train-test dataset

Ratio/max depth/cv	Spectral Range	Data Transformation	Mean Accuracy	Standard Deviation	Order
0.25-5-3	Full	Original	0.768707483	0.053564679	1
0.25-4-3	Full	Second derivative	0.761904762	0.034687208	1
0.25-4-3	Visible	Second derivative	0.755102041	0.060080006	2
0.3-4-3	Full	First derivative	0.745762712	0.036614354	1
0.25-4-3	Red edge	Second derivative	0.741496599	0.025453452	3
0.3-5-5	Full	First derivative	0.740952381	0.075335146	1
0.25-4-5	Full	Second derivative	0.734482759	0.060008367	1
0.3-4-3	Full	Original	0.734463277	0.052393325	2
0.25-5-3	Visible	Original	0.727891156	0.041934789	2
0.25-4-5	Full	First derivative	0.721609195	0.081115228	2
0.25-5-3	Full	First derivative	0.721088435	0.053564679	4
0.25-5-3	Red edge	Original	0.721088435	0.025453452	3
0.3-5-3	Full	Second derivative	0.717514124	0.076218856	1
0.25-4-5	Full	Continuum Red edgemoval	0.714712644	0.052594968	3
0.25-5-3	Full	Second derivative	0.714285714	0.066652782	5
0.25-4-3	Full	Original	0.714285714	0.016663196	4
0.3-5-3	Full	Original	0.711864407	0.013838925	2
0.3-5-5	Full	Original	0.711587302	0.052586418	2
0.25-4-5	NIR	First derivative	0.707126437	0.037120565	4
0.3-5-5	NIR	Second derivative	0.706507937	0.1047696	3
0.3-5-3	Full	Continuum Red edgemoval	0.706214689	0.015979814	3
0.3-5-5	Full	Second derivative	0.700952381	0.039313843	4
0.3-5-3	Full	First derivative	0.700564972	0.034827198	4
0.25-4-5	Visible	Original	0.699770115	0.081631533	5
0.3-4-3	NIR	Original	0.694915254	0.013838925	3
0.3-4-5	Full	Continuum Red edgemoval	0.694603175	0.047245626	1
0.3-5-5	Full	Continuum Red edgemoval	0.694603175	0.035643316	5
0.3-4-3	Full	Continuum Red edgemoval	0.689265537	0.042278615	5
0.3-4-3	Visible	Second derivative	0.689265537	0.021139307	4
0.3-4-5	Full	Original	0.689206349	0.059485914	2
0.25-5-5	Full	Continuum Red edgemoval	0.688275862	0.076349531	1
0.25-5-5	Full	Original	0.687586207	0.08871419	2
0.25-4-3	NIR	Original	0.68707483	0.019241001	5
0.25-5-5	Full	First derivative	0.686896552	0.051185445	3
0.25-5-5	Full	Second derivative	0.686666667	0.056989243	4
0.25-5-5	NIR	First derivative	0.680229885	0.076134655	5
0.3-5-3	Visible	Second derivative	0.677966102	0.0553557	5
0.3-4-5	NIR	Second derivative	0.677936508	0.048671957	3
0.3-4-5	Full	Second derivative	0.677301587	0.077767411	4
0.3-4-5	Full	First derivative	0.661269841	0.027736477	5

Appendix E: Cross Lochs maps from intial analysis

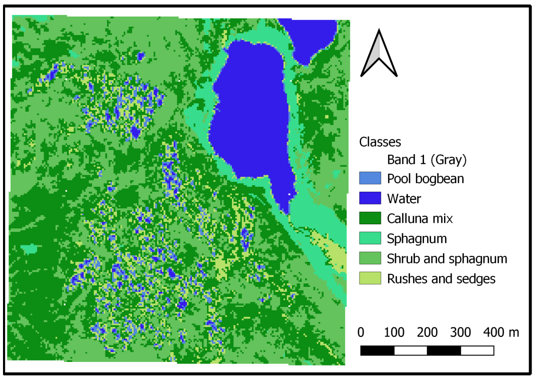


FIGURE E.1: K-means map with 6 clusters; cluster names based on location and similarities with the random forest maps.

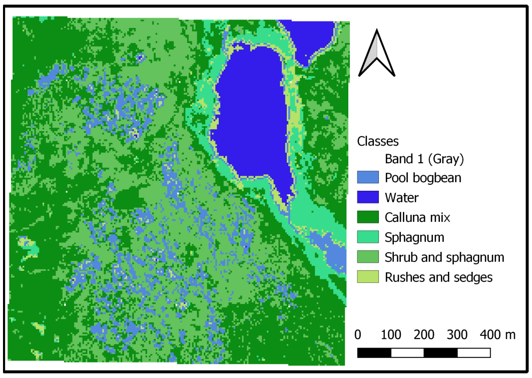


FIGURE E.2: Random forest map using the original spectra, focusing on the SWIR bands with 75% training data and 25% testing.

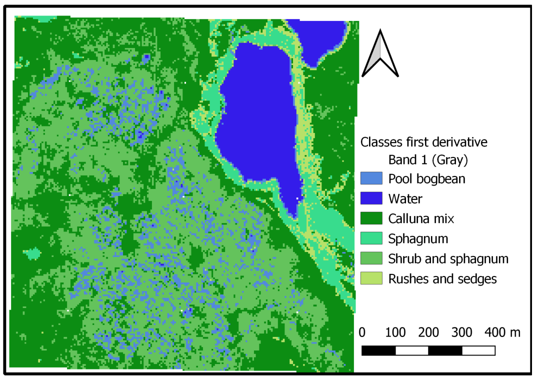


FIGURE E.3: Random forest map using the first derivative, using all 358 bands with 70% training data and 30% testing.

Appendix F: Data Management Plan

1) Provide the title and briefly describe the aim and objectives of your MRes project.

Project title: A hyperspectral approach to understand the association between PSM (as measured by InSAR data) and vegetation assemblage for a Scottish peatland

Research Aim: To determine whether hyperspectral data can be used to understand the association between peatland surface motion (as measured by InSAR data) and land cover in the Flow Country.

Objectives:

1. Assess the extent to which supervised and unsupervised machine learning algorithms can be used to classify plant functional types.
2. Determine whether machine learning can be used to show a relationship between plant functional types and peat surface motion.

2) What data will be produced? (Data types, format, standards, scale and method)

Fieldwork was be undertaken for this project to validate the predictions and add to the train-test data. The majority of the input data was digital and from open sources. It was analysed in QGIS and Python.

Secondary data includes:

- AVIRIS-NG data collected 15th July 2021, 425 bands, pixel size of 5 x 5 m, requiring 28 GB storage in total (4 sites with a total of 8 images), accessible from https://ares-observatory.ch/esa_chime_mission_2021/. This data required preprocessing.
- InSAR timeseries data collected 2015-2019, pixel size 90 x 70 m, requiring 32 MB storage, accessible from https://catalogue.ceh.ac.uk/documents/7c2778bf-b498-4ba2-b8cb-60a2081e5ba7. This data was pre-processed prior to use.

None of the data used has any special requirements.

The hyperspectral data had already undergone some atmospheric, geometric and radiometric corrections, however, the number of dimensions were reduced to remove atmsopheric water vapour. The data was cropped spatially to four 1 km^2 sites (to reflect the size of sites used in InSAR PSM studies) and the locations openly shared. It was be analysed in QGIS and Python with some use of plugins such as EnMAP-Box. The Cross Lochs (near-natural) site was used as a control site with which to compare the outcomes from the other three sites. Python code produced will made openly available on GitHub.

The hyperspectral data is being used to add detail to InSAR PSM outcomes. The data generated in the project through fieldwork will be made available with an Open Data license in the University of Nottingham Research Data Repository.

Data outputs included:

- Random forest classification maps of PFT predictions (Geotif)
- k-means cluster maps with 'k' equating to the number of clusters (Geotif)
- Graphs of the top outputs and their accuracies
- Confusion matrices regarding the predictions compared to the train-test data
- Classification comparisons and associated McNemar scores

3) What metadata standards will you use? (Metadata content and format)

The metadata was initially documented using the Data Tree template, as were the data dictionaries. Each dataset had its own metadata and data dictionary, saved in a separate 'docs' folder. They were named clearly, using the following format: 'datasetname-metadata.txt' and 'datasetname-datadict.txt'. Any other documentation was be stored in the 'docs' folder including a summary of any methods and issues.

The generated data complied with the following ISO metadata standards:

- Geographical data will comply with ISO19115 (focused on geographic information and services)
- Ecological data will comply with the Ecological Metadata Language (EML)

Some metadata can be automatically generated in QGIS, however, there are limitations and the metadata needed to be checked and added to.

4) How will your data be structured and stored? (Project storage)

The data was stored on my University of Nottingham OneDrive and back ups were regularly made on the Newcastle University OneDrive. Analysis code was be stored on GitHub in addition to within saved files in Jupyter Notebook/PyCharm. As both OneDrive accounts hold over 1TB of data, there was enough storage space for my project. Outputs were also stored locally, however, due to the size of files, not all of the input and processed data can be stored on my university laptop. Cloud storage is beneficial as files could be synchronised to use on different devices, there was protection against hardware failures and no server maintenance was required, however, there were security concerns, which is why regular backups were made.

The data was be structured in a logical manner using clearly named folders, subfolders and files. There were separate folders for 'input data', data undergoing processing/change ('processing'), and 'output data'. All data was be clearly labelled. There were also be folders for data documentation ('docs') for each data output. Each folder contained a README file which summarised the folder's contents. Other folders were created as required.

5) How will the data be shared during and after the project? (Access, data sharing and reuse)

There was not the need to share data with supervisors during the project, however, if this had been the case, I would have shared data folders from my OneDrive with them. GitHub was be used for code storage, version control and to share code and associated analysis to facilitate its reuse once the project was finished. All code, data and associated documents (including the final data management plan) will be made available 12 months after the end of the project and stored within the University of Nottingham's Research Data Repository. A DOI and URL will be created for my deposits in the data repository so that it can be accessed and reused in the future.

6) Outline the approach to data selection and long-term preservation?

All of the research data is open and, therefore, can be shared. The data was be prepared in line with the requirements of the University of Nottingham Data Repository and will be stored there for a minimum of 10 years. This will include all data generated, code scripts and additional files (including metadata, data dictionaries and README files). Copies of the data will also be kept on personal University OneDrive accounts. Versions of code will also be available on GitHub (https://github.com/rachelzwalker/Flow_Country_HSI_and_PSM).

7) Who has responsibility for implementing the DMP and are resources required?

The researcher was primarily responsible for implementing the DMP as well as making reflections and updating it. Once the data is in the repository, the University of Nottingham is responsible for maintaining data access following the project. Data validation is the responsibility of the principal researcher. Guidance will be requested as required from supervisors and IT Services at the University of Nottingham.

Appendix G: Gantt Chart

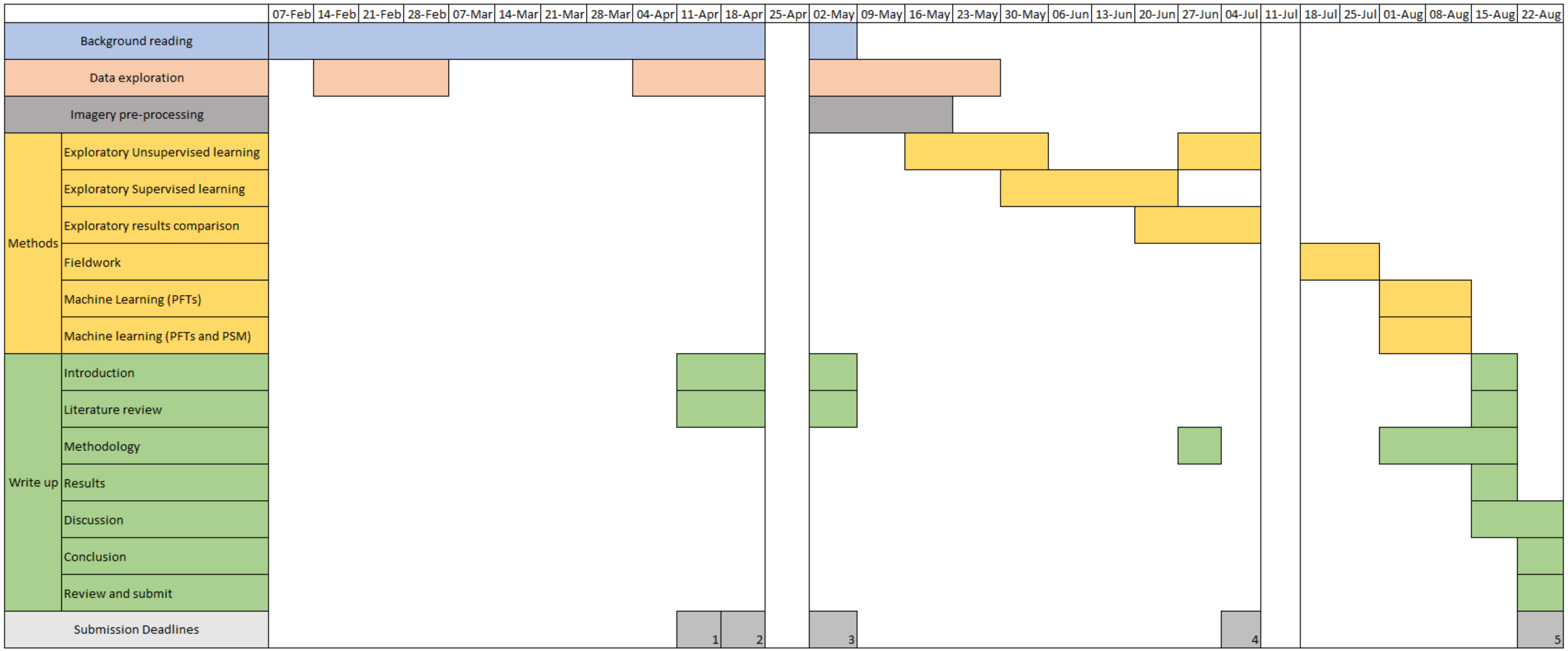


FIGURE G.1: Finalised Gantt chart regarding project management. The submission deadlines involve: the literature review poster, project proposal, data management plan, interim report and dissertation submission. Additionally, the first of each month sees the update of the data management plan, files backup and GitHub check.