Comparing Sentinel 2, SAR and Terrain Data Composites for Heathland Classification in Conne River using Random Forest Classifier

Prepared by: Alexander Johnston
For: Dr. Ben DeVries

GEOG*6060
Fall 2023

Table of Contents

Abstract

Classified maps of heathlands provide insights into the distribution of these culturally and ecologically significant ecosystems, aiding in conservation efforts, land management, and ecological studies. Data composites were tested based on their ability to classify heathland, forest, bogs, water, and, sand and exposed rocks classified in a Random Forest model. This classification model serves to demonstrate Google Earth Engine's abilities and acts as a precursor to more sophisticated classification models. Data composites consisted of combinations of sentinel 2, sentinel 1 SAR, and a 5m DEM. Heathlands exhibited high spectral separability, particularly in the green, red and near-infrared bands. The standalone spectral data achieved an overall accuracy of 94.63%, and the inclusion of terrain and SAR data resulted in a marginal increase, raising the overall accuracy to 95.12%.

Introduction

To detect and map heathlands across Miawpukek First Nation (MFN) traditional territory, a preliminary step involves identifying the most suitable variables for land cover classification in the study area. Combining multiple data sources through data fusion has proven useful for classifying land cover. Different data sources capture different properties of land cover. For instance, optical imagery like Sentinel 2 captures visible light, which is great for identifying vegetation types, while Sentinel 1 SAR data can capture features like soil moisture, surface roughness, and vegetation structure (Tempfli et al., 2009). Complimentary data can often increase the model's classification accuracy. For example, using an object-based random forest classification of wetlands in Newfoundland, Mahdianpari et al. (2018) were able 79.14% accuracy using SAR, 83.79% accuracy using optical data, and 88.37% when combining the two (Mahdianpari et al., 2018). Google Earth Engine (GEE) is fitting for heathland classification due to its seamless access to extensive data via its data catalogue, efficient processing and analysis of large datasets through cloud computing, and direct integration of classifiers such as Random Forest, simplifying the entire workflow.

Forest, Water, Bog and Exposed Rock were selected as classes as they are all found within heathland complexes. Exploring different land cover types beyond the target class can enrich our understanding of landscapes, enhancing classification accuracy and offering valuable insights into ecosystem dynamics. Furthermore, investigating heathland succession post-disturbance through time series analysis stands as a promising research direction. This approach would utilize the classification of various land classes to unveil both primary and secondary succession patterns within heathland ecosystems.Heathlands hold both cultural and ecological importance, housing cultural landmarks and serving as habitats for rare, specialized plants, as well as cultural keystone species (Oberndorfer & Lundholm, 2009; Schaefer et al., 2016).

It was announced in 2022 that Miawpukek First Nation would receive $3 million from the Ministry of Environment and Climate Change to support their area-based conservation work (*The Government of Canada and Miawpukek First Nation in Newfoundland and Labrador Take First Steps toward a New Indigenous Protected and Conserved Area*, 2022). The Government of Canada has made commitments to conserve 25 percent of land and inland waters in Canada and 25 percent of oceans by 2025 and is working toward 30 percent of each by 2030. The data from this project could help support indigenous-led conservation efforts in MFN by providing an inventory of habitats in their traditional territory.

Methods & Data

Study Area

The traditional territory of Miawpukek First Nations (MFN) spans nearly 23,000 km^2, extending across the central and maritime barren ecoregions of Newfoundland. Little is known about the extent of heathlands within MFN's traditional territory. Twenty-six heathland sites, each with three 20x20m plots were sampled this summer as part of a study on functional traits. GPS points were gathered for each plot and many additional points were collected as ground-truthed heathland points. Eighty-six points fall within the study area and will be used as sample points in the training and testing dataset. This classified land cover map is of the Sentinel 2 tile that contains the majority of the field sites visited. Within the study area, there are two distinct classes of heathlands: edaphic and successional heathlands. Successional heathlands develop as a result of stand-replacing disturbances, such as after a fire or clear-cutting, and will gradually return to forest over time through secondary succession (Meades, 1983). Edaphic heathlands, on the other hand, form in regions with specific environmental conditions such as severely wind-exposed or nutrient-poor areas (Meades, 1983). For this project, both classes will be identified under the umbrella category heathland. Future models using environmental variables will be used to delineate these two classes.

Sentinel 2 and Sentinel 1 SAR Image Collections

Newfoundland experiences “chronic cloud cover” and snow and ice cover in the winter which limits the number of usable observations (Mahdianpari et al., 2018). To overcome this issue, a composite of Sentinel 2 images from the 2022 growing season was used.

The Sentinel 2 and Sentinel 1 SAR image collections used in this project were narrowed down to the 2022 growing season, spanning from May 1st to September 1st. This timeframe was selected because summer months yield ample spectral data, and the 2022 dataset offers recent information without significant cloud cover interference.

Dual polarity VV and VH Interferometric Wide Swath (IW) data was used to create the SAR composite. The composite consisted of the mean VV and VH backscatter over the study period. SAR provides soil moisture, surface roughness, and vegetation structure

Sentinel 2 MSI image collection filtered by our date range was used to create a cloud-free composite. The methods outlined in the s2cloudless tutorial by Justin Braaten were used as a basis to create the cloud free composite (Braaten, 2023). The parameters for the sentinel 2 collection and cloud mask were adjusted based on visual interpretation of the cloud mask and a layer containing the number of observations at each pixel in the cloudless composite. The maximum allowable cloud cover percent will be set to 60%, the cloud probability threshold was set to 50%, the near-infrared reflectance threshold was set to 0.15, the max cloud shadow distance was set to 1km, and the buffer was set to 50m. These values produced optimal cloud masking outcomes without creating observation gaps across the study area. First, the Sentinel 2 MSI image collection is combined with the Sentinel-2 cloud probability collection. The cloud mask is conditioned based on the cloud probability and the probability threshold value set. Cloud shadows are then identified based on dark NIR pixels that are within our specified distance from our cloud and that are not considered water based on the scene classification (SCL) band in the Sentinel 2 Image collection. The cloud and cloud shadow masks are then applied to each image in the collection. We then take the median of each cloud-masked image in our collection to produce the final cloud-free composite.

Spectral Indices

Normalized Vegetation Index (NDVI) is commonly used to assess and monitor the health and density of vegetation (Tempfli et al., 2009). It is calculated using the following formula.

$$NDVI = \frac{NIR - R}{NIR + R}$$

Band 4 and Band 8 For this application, NDVI can help differentiate between land cover types. NDVI is sensitive to photosynthetically active biomass (Tempfli et al., 2009). Land cover classes with higher plant biomass such as closed canopy forests, will have higher NDVI than heathlands which contain less plant biomass.

Tasselled Cap Transformation (TCT) is a technique used to transform satellite image bands into fewer bands associated with physical scene features such as brightness, greenness, and wetness (Shi & Xu, 2019). TCT is similar to a PCA but captures distinct physical properties within the data (Crist & Cicone, 1984). The coefficients used in the transformation matrices are determined through regression or statistical methods (Crist & Cicone, 1984). TCT was originally developed for Landsat data but the underlying principle of transforming spectral bands into composite bands representing different land surface characteristics remains applicable to other satellite sensors with similar spectral bands (Shi & Xu, 2019). The TCT for this project was calculated using Sentinel 2 MSI coefficients from Shi and Xu, 2019 (Shi & Xu, 2019). The calculation involves matrix multiplication of the original bands by transformation coefficients to produce the new composite bands.

Tasselled Cap Transformation	Blue	Green	Red	NIR	SWIR	SWIR 2
Brightness	0.3510	0.3813	0.3437	0.7196	0.2396	0.1949
Greeness	-0.3599	-0.3533	-0.4734	0.6633	0.0087	-0.2856
Wetness	0.2578	0.2305	0.0883	0.1071	-0.7611	-0.5308

Table 1: Tasselled Cap Transformation coefficients for 6 band image.

Terrain Data

Elevation was derived from a 5m DEM from the Newfoundland GeoHub. This data was acquired from 1:20,000 scale aerial photographs with an accuracy of +/-2 Meters. Slope was originally derived from the DEM but was found to have the lowest relative importance score in RF and was removed from the composites and from further analysis.

The topographic position index (TPI) was calculated for each point based on a 50m square neighbourhood. Heathlands can often be found on wind-exposed ridges (Meades, 1983). A measure of topographic exposure may help their classification. TPI is used to distinguish topographic features such as valley bottoms, exposed ridges, and flat plains (Newman et al., 2018). TPI is calculated by subtracting the elevation at a point by the mean elevation of a 10 x 10-pixel window surrounding the pixel. TPI is a relatively simple measure of exposure, only considering elevation within a local neighborhood which makes it scale sensitive (Newman et al., 2018). Furthermore, it does not consider other topographic factors influencing exposure, such as aspect, slope, or curvature (Newman et al., 2018).

Testing and Training Sample

Once the cloud-free composite was created, sample points were gathered for the following classes: heathlands, forest, water, bog, and road. This was a non-random sample, 561 sample points were collected based on visual interpretation of the Sentinel 2 cloud-free composite

Class	**Number of samples**
Heathlands	182
Forest	185
Bogs	168
Sand and Exposed Rocks	86
Water	176

Table 2: Table showing the number of sample points for each class in the model.

The sample points were split 70/30 for training and validation. GPS points collected during this summer's field season were used as secondary validation to "ground truth" the model.

Data Composites

The following data composites were tested as predictors for land cover classification:

Data Composite 1	Surface reflectance
Data Composite 2	Surface reflectance and terrain data
Data Composite 3	Surface reflectance and SAR data
Data Composite 4	Surface reflectance, SAR, and terrain data
Data Composite 5	SAR and Terrain
Data Composite 6	Terrain, SAR and Indices

Table 3: Table showing the data composites that were tested for the ability to identify target classes.

Random Forest

For this project, random forest (RF) models were used to classify pixels based on training points and the data composites. RF is non-parametric, meaning it does not assume any specific distribution of the data. This allows it to handle complex relationships and makes it less prone to overfitting than statistics based classifiers (Adugna et al., 2022; Mahdianpari et al., 2018; Salwa Thasveen & Suresh, 2021). RF combines the predictions of multiple decision trees (Salwa Thasveen & Suresh, 2021). Each decision tree is trained on a subset of the data and a random subset of the features. During classification, the algorithm aggregates the predictions of these individual trees to make a final classification decision (Salwa Thasveen & Suresh, 2021).

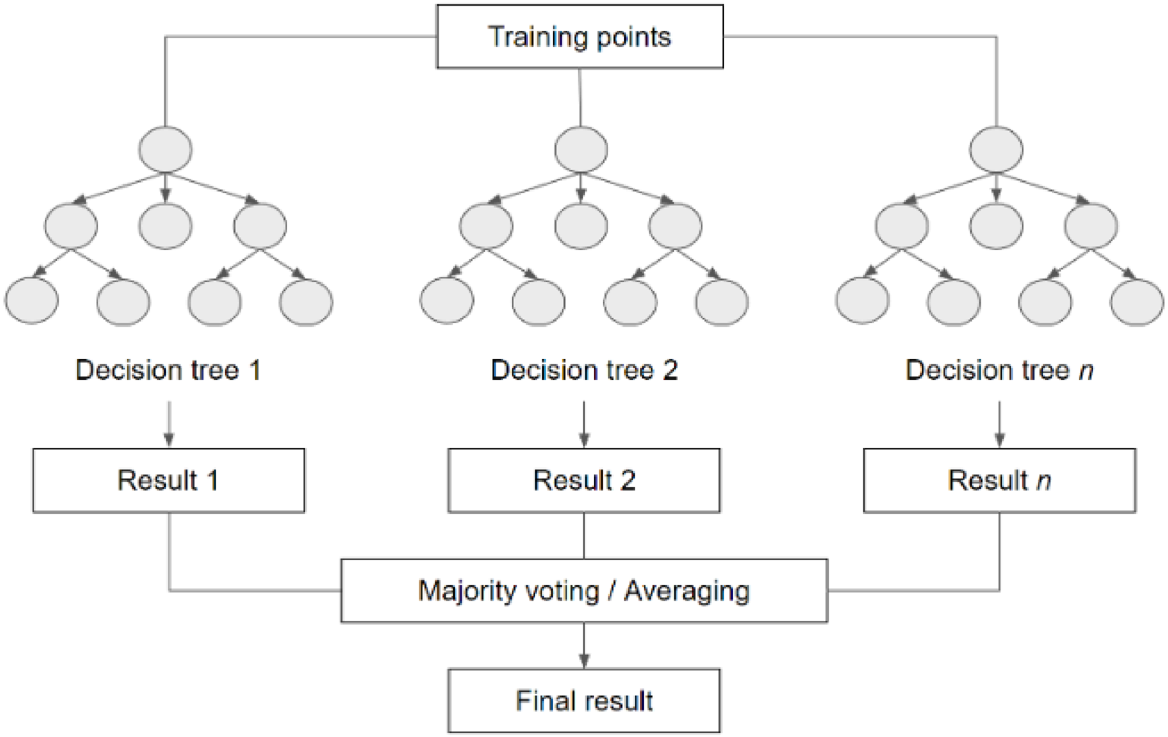

Figure 1: Concept of Random Forest decision tree aggregation. From “Cloud-based remote sensing with Google Earth Engine: fundamentals and applications” By Cardille et al., 2023, *Springer*, F2 p.39
RF can be tuned by adjusting the number of trees (Ntree). The Ntree parameter was assessed at values (25, 50, 75, 100, 125, 150). A value of 100 trees was found to be appropriate for this project. At this value, the combined overall accuracies across data composites were maximized.

Accuracy Assessment

Confusion matrices were generated for each data composite from which the producer, user, and overall accuracy were derived. The overall accuracy is the proportion of the data classified correctly by the model (Cardille et al., 2023). It is calculated as the sum of correctly identified pixels divided by the total number of pixels in the sample (Cardille et al., 2023).

$$Overall\ Accuracy\ =\ \frac{Number\ of\ correctly\ classified\ pixel}{Sample\ Size}$$

The overall accuracy does not take into account classes imbalances

User Accuracy (UA) is the probability that a pixel is classified by the algorithm as its correct class (Cardille et al., 2023).

$$User\ Accuracy\ (UA)\ = \frac{Number\ of\ correctly\ classified\ pixels\ for\ a\ class}{Total\ number\ of\ pixels\ classified\ as\ that\ class}$$

Producer Accuracy (PA) is the probability that a pixel in a particular class is correctly classified as that class (Cardille et al., 2023).

$$Producer\ Accuracy\ (PA) = \frac{Total\ number\ of\ pixels\ that\ belong\ to\ that\ class}{Number\ of\ correctly\ classified\ pixels\ for\ a\ class}$$

Lastly, is the kappa coefficient, which evaluates how well the classification performed compared to a random assignment of categories (Cardille et al., 2023). The value of the kappa coefficient can range from −1 to 1 (Cardille et al., 2023). Negative values indicate that the classification is worse than a random assignment of categories (Cardille et al., 2023). Positive values indicate that the classification is better than random (Cardille et al., 2023).

$$Kappa\ Coefficient\ = \frac{overall\ \ accuracy - chance\ agreement}{1 - chance\ agreement}$$

Chance agreement is the sum of the product of row and column totals for each class (Cardille et al., 2023).

Class Separability

A CSV of the training points was extracted from Google Earth Engine. The reflectance values of each band at each point were extracted. Using Seaborn, a python data visualization library, box and whisker plots were created of the key Sentinel 2 and Sentinel 1 SAR bands (See Table 4) were plotted to assess class separability.

Results and Discussion

Importance Assessment

Importances scores were extracted from a 20m random forest model trained on all of the input data layers created for this project. This 20m model using all bands achieved an overall 93.66% and a kappa of 0.9194. Many of the high-importance bands have a 10-meter spatial resolution (See Table 4). To leverage the higher spatial resolution of these critical 10-meter bands, a data composite using only 10m bands was tested. By excluding 20m bands and classifying at 10m resolution the overall accuracy increased to 95.12% (See Table 5).

Bands	Relative Importance Value	Spatial Resolution
B3	58.46	10m
B2	54.72	10m
NDVI	53.18	10m
B4	46.70	10m
Greenness	46.29	20m
VH	42.14	10m
TPI	41.10	10m
B5	35.48	20m
B8	34.29	10m
B6	33.91	20m
VV	33.54	10m
Brightness	33.21	20m
B7	30.05	20m
B12	29.47	20m
B8A	29.39	20m
B11	28.96	20m
Elevation	28.37	10m
Wetness	22.84	20m

Table 4: Table displaying the bands in order of relative importance values extracted from a random forest model trained with all the input layers.

Class Separability in Spectral and SAR Data

The land cover classes chosen exhibited a high degree of spectral separability, particularly in red, green and near-infrared bands. Using spectral data from B2: Blue, B3: Green, B4: Red, and B8: NIR alone achieved a 94.63% overall accuracy (See Figure #). The class separation among VH and VV backscattering was relatively low. Forests exhibit high VV and VH backscatter when compared to the other classes, perhaps due to the presence of vertical structures or denser more complex vegetation structures.

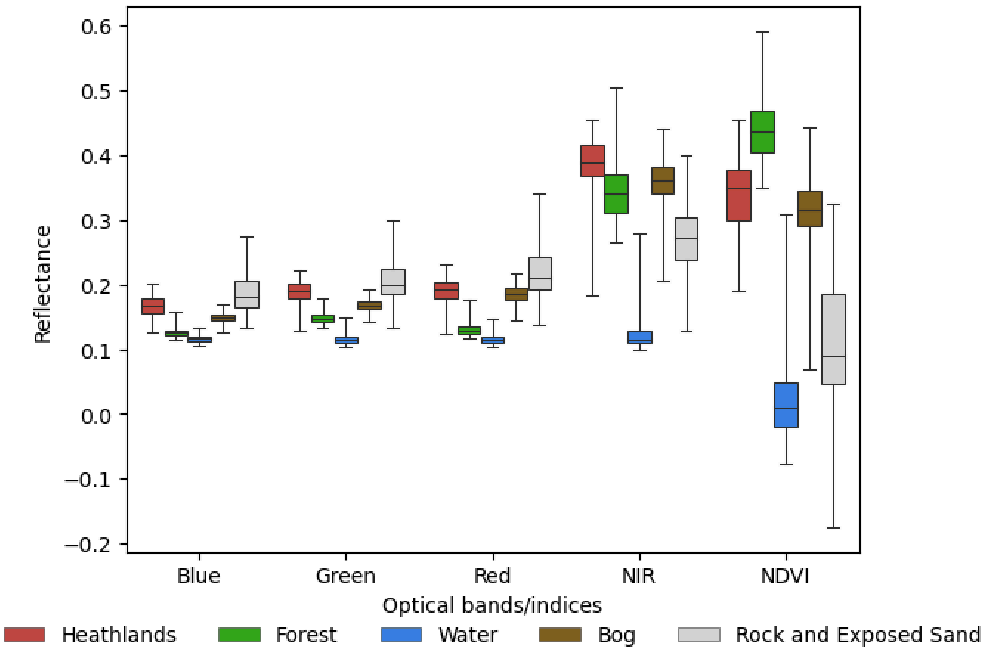

Figure 2: Box-and-whisker plot of the 2022 summer growing season composite illustrating the distribution of surface reflectance, and NDVI for land cover classes obtained from pixel values extracted from the training dataset.

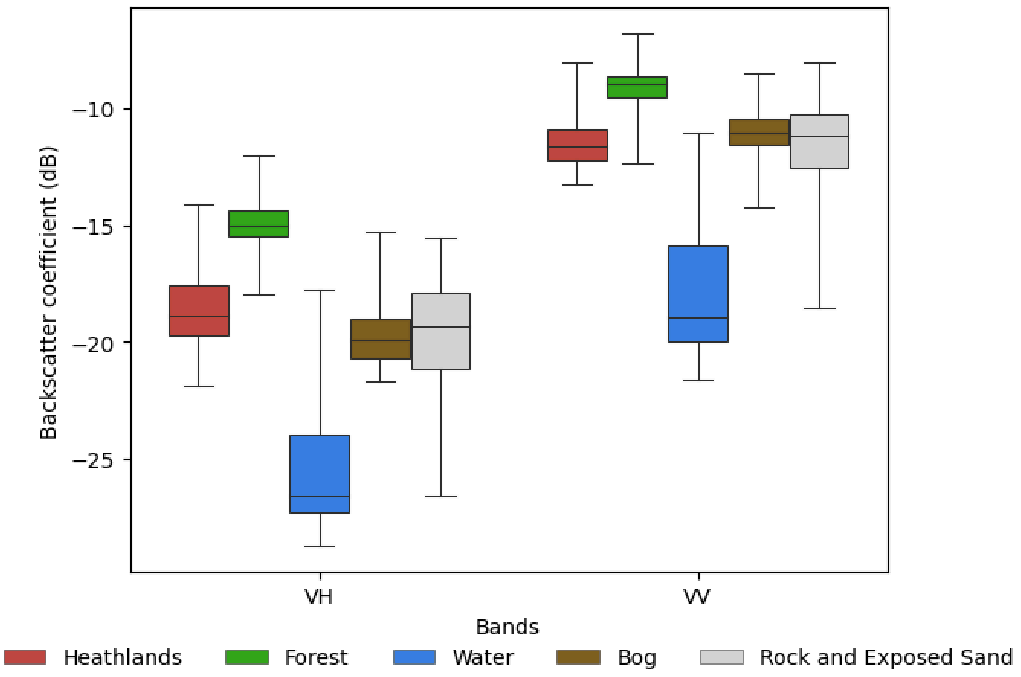

Figure 3: Box-and-whisker plot of the 2022 summer growing season composite illustrating the distribution of vertical-horizontal and vertical-vertical backscatter for land cover classes obtained from pixel values extracted from the training dataset.

Accuracy of Data Composites

Spectral data outperformed SAR data in classification, aligning withe the greater class separability as evident in Figure 2 and 3. The addition of SAR to the Spectral data composite did not increase the classification accuracy. However, the addition of SAR to the Spectral and Terrain composite did increase the classification, although marginally. Integrating SAR with spectral data alone didn't boost classification accuracy. However, when added to the spectral and terrain composite, SAR showed a slight improvement in classification.The combined Spectral, Terrain and SAR data composite achieved the highest overall accuracy at 95.12%. The output of the combined Spectral, Terrain, and SAR data underwent secondary validation, yielding an accuracy of 93.02% based on ground-truthed heathland points gathered during the summer field season.

Data Composites	**Overall Accuracy (%)**	**Kappa Coefficient**
Spectral	94.63	0.9318
Spectral and SAR	94.63	0.9318
Spectral and Terrain	94.15	0.9256
Spectral, Terrain and SAR	95.12	0.9380
Terrain and SAR	77.18	0.7089
Terrain, SAR and Indices	90.73	0.8822

Table 5: The overall accuracy and the kappa coefficient of random forest classification of data composites

Class	Spectral		Spectral and SAR		Spectral and Terrain		Spectral, Terrain and SAR		Terrain and SAR		Terrain, SAR and Indices	
	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)
Heathland	91.49	97.73	93.62	95.65	95.74	93.75	95.74	93.75	74.47	67.31	91.49	93.48
Forest	100	92.59	100	94.34	98	92.45	100	94.03	90.2	82.14	98	90.74
Water	100	97.67	100	97.67	100	97.67	100	97.67	90.48	90.48	95.24	88.89
Bog	93.18	91.11	90.91	90.91	86.36	92.68	88.64	95.12	68.18	75	86.36	92.68
Exposed Rock	81.82	94.74	81.82	94.74	86.36	95	86.36	95	45.45	62.5	72.73	84.21

Table 6: The producer and user accuracies of data composites

Potential Avenues for Research

While achieving a high classification accuracy, it's important to note that this result might be inflated due to the limited, non-probability sample used. The exclusion of mixed pixels in the training and testing samples could skew the representation of variability within each class. Implementing stratified random sampling could enhance the model's accuracy by capturing a more comprehensive range of variability.

Among the data types, spectral data alone produced the highest heathland classification user accuracy (See Table 6). However, there's potential for enhancing the effectiveness of terrain and SAR data. Radar in the L band range penetrate vegetation and capture soil moisture which may be advantageous for distinguishing heathlands and bog classes. Combining a multiscale TPI with other indices such as curvature or horizon angle or incorporating additional geospatial data such as wind energy/speed date, and proximity to the coast might yield more comprehensive insights into exposure assessments.

Spectral data alone can achieve a high accuracy (Table 5) while reducing the dimensionality and computational complexity of the model. For classification over an extensive study area, using solely spectral data may be favourable due to reduced file size and computational needs. The classified maps produced by the data composites display noticeable salt and pepper noise, which could be reduced by using object based classification techniques. Object based classification has been shown to reduce this noise, while adding texture, and incorporating relational information, potentially enhancing classification accuracy. While object-based classification has shown success in classifying fragmented habitats like wetlands, its application in heathland classification remains an unexplored area worthy of investigation (Mahdianpari et al., 2018; Orlikova & Horak, 2019).

Conclusion

The evaluation of data composites reveals the pivotal role of spectral data in achieving high classification accuracy, particularly in identifying heathland areas. The classes chosen for this project have a high degree of spectral separability particularly in the green, red and NIR band (See Figure 2). The inclusion of SAR and terrain data, while beneficial, showed marginal improvements in classification results.However, it's crucial to note the potential biases stemming from the non sample sample. Implementing stratified random sampling could mitigate this limitation and enhance the model's accuracy by capturing a broader spectrum of variability within each class.

References

Braaten, J. (2023). *Sentinel-2 Cloud Masking with s2cloudless*. Google Earth Engine Community. https://developers.google.com/earth-engine/tutorials/community/sentinel-2-s2cloudless

Cardille, J. A., Crowley, M. A., Saah, D., & Clinton, N. E. (Eds.). (2023). *Cloud-based remote sensing with Google Earth Engine: Fundamentals and applications*. Springer.

Crist, E. P., & Cicone, R. C. (1984). A Physically-Based Transformation of Thematic Mapper Data—The TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing*, *GE-22*(3), 256–263. https://doi.org/10.1109/TGRS.1984.350619

Mahdianpari, M., Salehi, B., Mohammadimanesh, F., Homayouni, S., & Gill, E. (2018). The First Wetland Inventory Map of Newfoundland at a Spatial Resolution of 10 m Using Sentinel-1 and Sentinel-2 Data on the Google Earth Engine Cloud Computing Platform. *Remote Sensing*, *11*(1), 43. https://doi.org/10.3390/rs11010043

Meades, W. J. (1983). Heathlands. In *Biogeography and Ecology of the Island of Newfoundland* (Vol. 48, pp. 267–318). Dr. W. Junk Publishers, The Hague.

Newman, D. R., Lindsay, J. B., & Cockburn, J. M. H. (2018). Evaluating metrics of local topographic position for multiscale geomorphometric analysis. *Geomorphology*, *312*, 40–50. https://doi.org/10.1016/j.geomorph.2018.04.003

Oberndorfer, E. C., & Lundholm, J. T. (2009). Species richness, abundance, rarity and environmental gradients in coastal barren vegetation. *Biodiversity and Conservation*, *18*(6), 1523–1553. https://doi.org/10.1007/s10531-008-9539-5

Schaefer, J. A., Mahoney, S. P., Weir, J. N., Luther, J. G., & Soulliere, C. E. (2016). Decades of habitat use reveal food limitation of Newfoundland caribou. *Journal of Mammalogy*, *97*(2), 386–393. https://doi.org/10.1093/jmammal/gyv184

Shi, T., & Xu, H. (2019). Derivation of Tasseled Cap Transformation Coefficients for Sentinel-2 MSI At-Sensor Reflectance Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(10), 4038–4048. https://doi.org/10.1109/JSTARS.2019.2938388

Tempfli, K., Huurneman, G. C., Bakker, W., Janssen, L. L. F., Feringa, W. F., Gieske, A., Grabmaier, K. A., Hecker, C., Horn, J., Kerle, N., Meer, F. D., Parodi, G., Pohl, C., Reeves, C. V., Ruitenbeek, F. J. A., Schetselaar, E., Weir, M., Westinga, E., & Woldai, T. (2009). *Principles of remote sensing: An introductory textbook.* (pp. 56–85).

The Government of Canada and Miawpukek First Nation in Newfoundland and Labrador take first steps toward a new Indigenous Protected and Conserved Area. (2022). Environment and Climate Change Canada. https://www.canada.ca/en/environment-climate-change/news/2022/09/the-government-of-canada-and-miawpukek-first-nation-in-newfoundland-and-labrador-take-f

irst-steps-toward-a-new-indigenous-protected-and-conserved-area.html