EVALUATING MULTITEMPORAL SENTINEL-2 DATA FOR FOREST MAPPING USING RANDOM FOREST

Marc Nelson, Metria   marc.nelson@metria.se

ABSTRACT

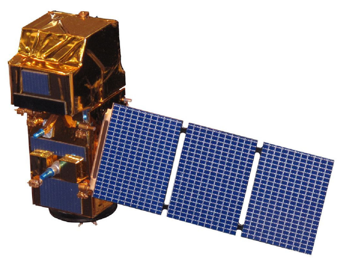
The mapping of land cover using remotely sensed data is most effective when a robust classification method is employed. Random forest is a modern machine learning algorithm that has recently gained interest in the field of remote sensing due to its non-parametric nature, which may be better suited to handle complex, high-dimensional data than conventional techniques. In this study, the random forest method is applied to remote sensing data from ESA's Sentinel-2 satellite program, which was launched in 2015 yet remains relatively untested in scientific literature with respect to classifying forest types. In a study site of boreo-nemoral forest in Ekerö mulicipality, Sweden, a classification is performed for six forest classes based on CadasterENV Sweden, a multi-purpose land cover mapping and change monitoring program. The performance of Sentinel-2 is investigated in the context of time series to capture phenological conditions, optimal band combinations, as well as the influence of ancillary inputs. Using two images from the spring and summer of 2016, an overall map accuracy of 86.0% was achieved. The red edge, shortwave infrared, and visible red bands were confirmed to be of high value. Important factors contributing to the result include the timing of image acquisition, use of a feature reduction approach to decrease the correlation between spectral channels, and the addition of ancillary data that combines topographic and edaphic information. The results suggest that random forest is an effective classification technique that is particularly well suited to high-dimensional remote sensing data.

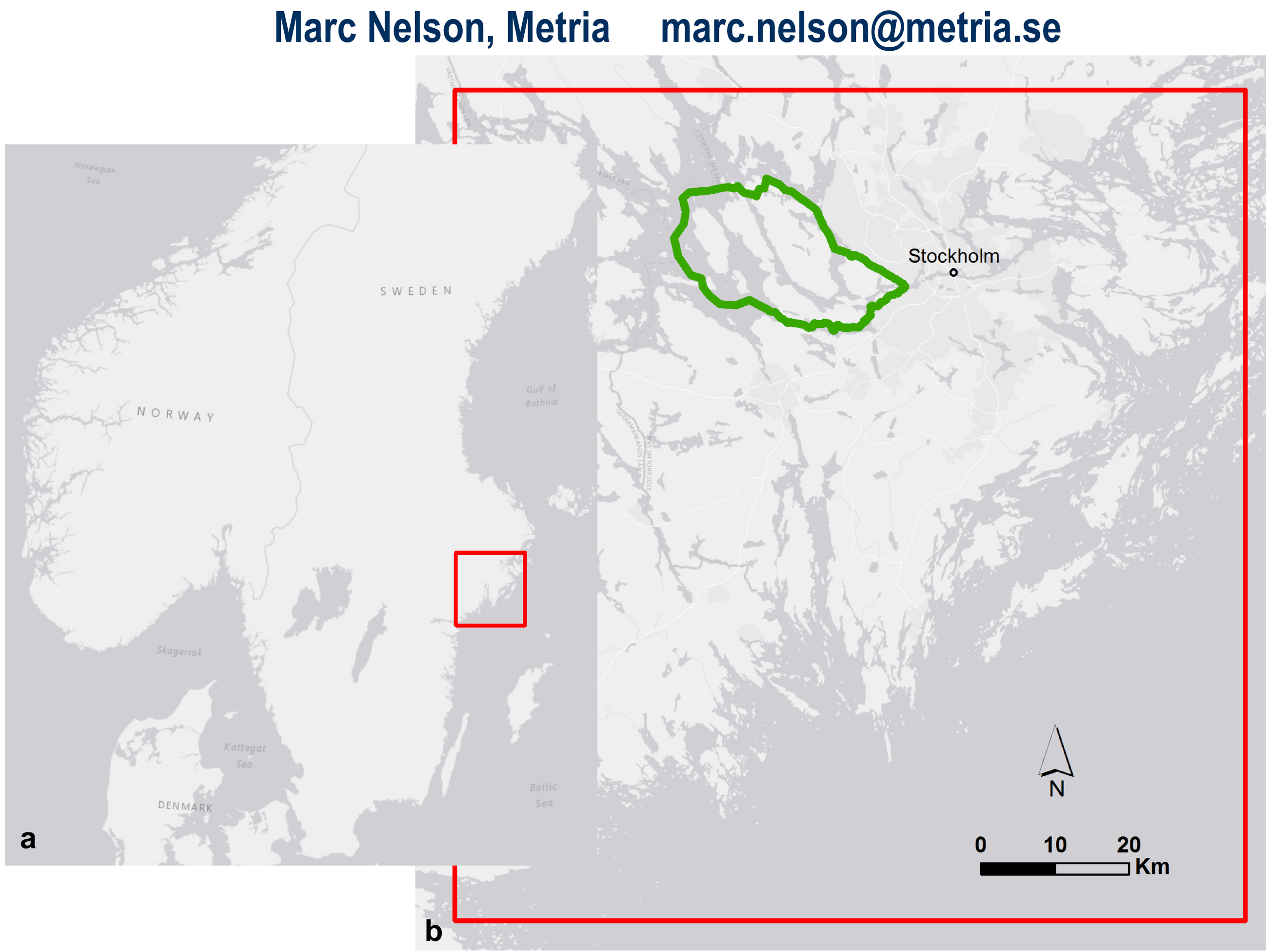
INTRODUCTION

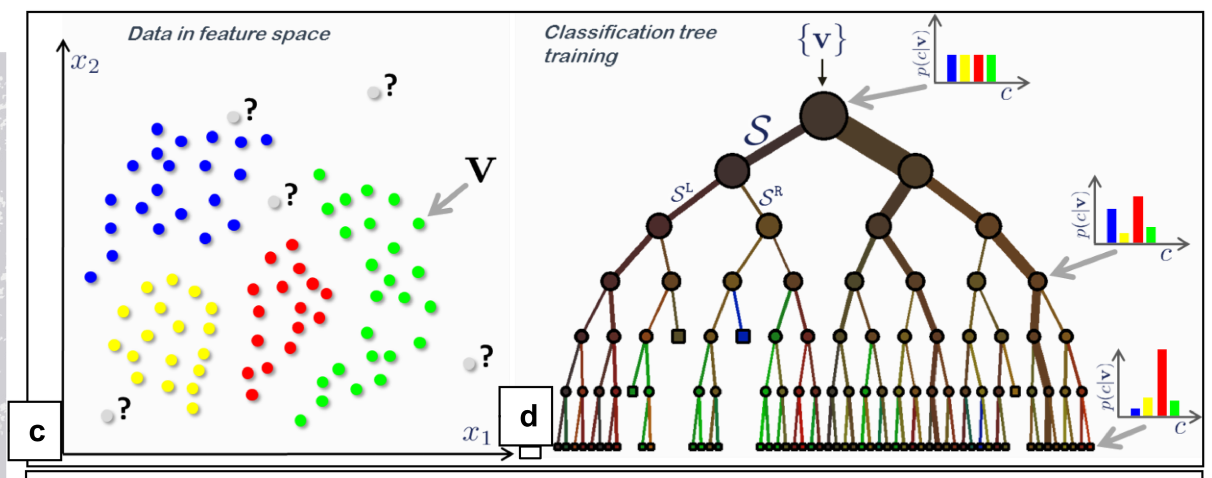
Machine learning techniques have become increasingly popular within a remote sensing context, due in part to the evolving nature of satellite datasets, which have become progressively larger and denser over time. Random forest is a type of non-parametric machine learning algorithm, which does not rely on assumptions of data distribution e.g. normality, and are generally more accurate than parametric techniques such as the widely-used maximum likelihood technique. By employing the bootstrap aggregation technique, random forest is less sensitive to overfitting. This method has been selected to perform a pixel-based classification for this study, due to its combination of ease-of-use, robustness to noise, as well as its demonstrated performance. Additionally, random forest is computationally light as well as being simple to set up and automate compared to other non-parametric classifiers such as Support Vector Machine, and can be easily implemented in open-source platforms such as R and Python's scikit-learn library.

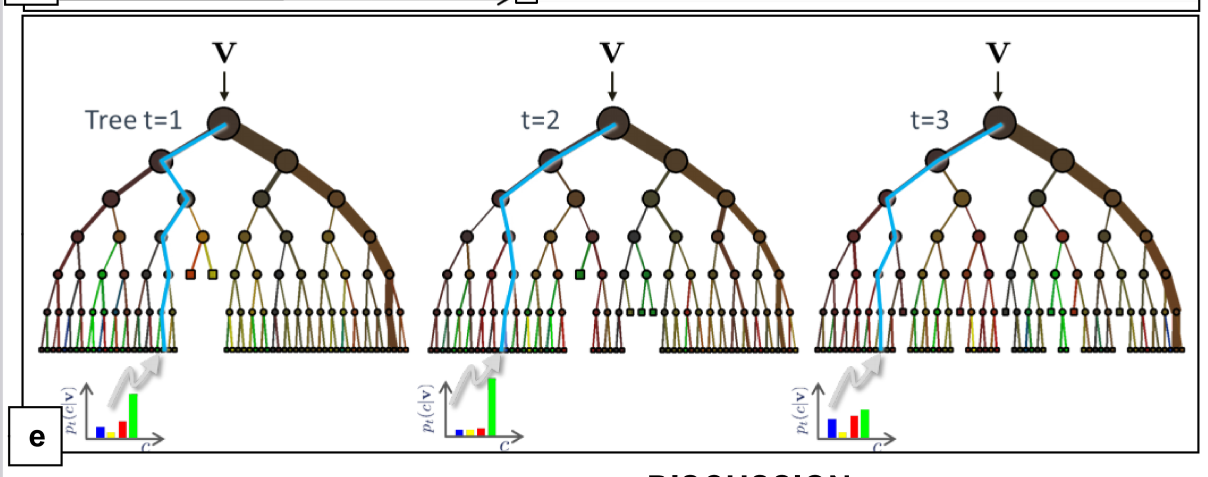
OBJECTIVE

This study seeks to contribute to the state-of-the-art by investigating the following research questions: (1) how can Sentinel-2 be best utilized for future forest classification studies, at a regional to national scale? (2) How do band combinations, time series, and ancillary data affect map accuracy, and how can these variables be adjusted to produce an optimal result? (3) How effective are machine learning techniques in the context of remote sensing? The study area for this project is the municipality of Ekerö, a group of islands in lake Mälaren in southeast Sweden with a land surface area of ~218 km2.







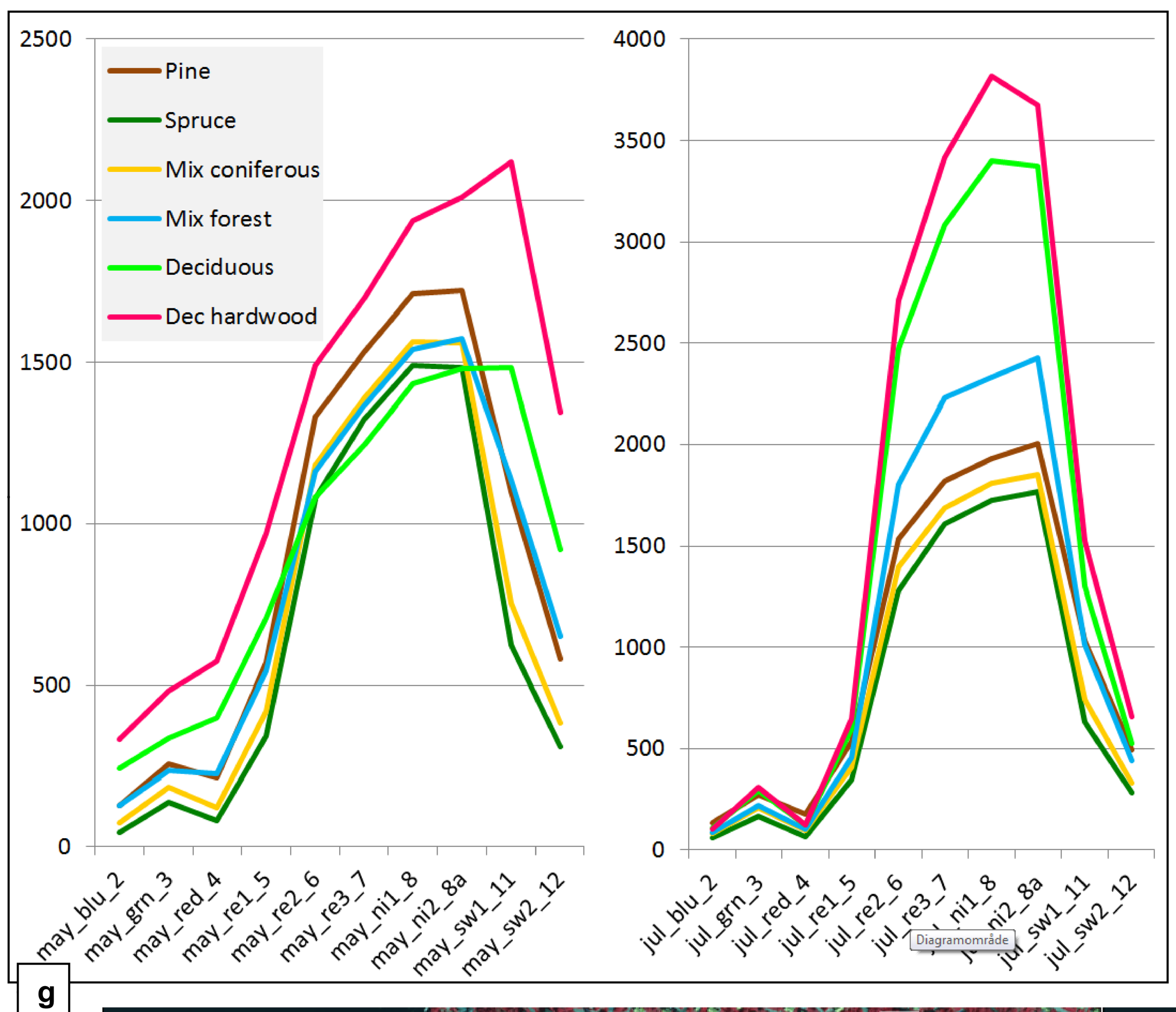


Descriptions of figures and tables. a: Location of Sentinel-2 tile 33VXF (outlined in red). b: location of Ekerö municipality (outlined in green). c: Visualization of random forest classification. Reference data (v) is labeled according to in-situ species composition. d: This data set is fitted to a random forest model in the training step where a 'forest' of binary classification trees of a random subset of input variables is 'grown'. e: The forest is then used in the prediction of new data (Criminisi et al., 2011). f: Correlation coefficient matrix for selected Sentinel-2 bands (study area). g: Mean BOA reflectance of CadasterENV classes for selected bands derived from reference data, May 2 (left) and July 21 (right). h: Classification map produced using random forest, Ekerö municipality (inset of i outlined in blue). i: Example portion of CadasterENV classification (left), 2015 CIR orthophoto for comparison (right).

	blue	green	red	RE 1	RE 2	RE 3	NIR 1	NIR 2	SWIR 1	SWIR 2
blue	1.000									
green	0,968	1.000								
red	0,957	0,963	1.000							
RE 1	0,845	0,923	0,918	1.000						
RE 2	0,224	0,386	0,244	0,499	1.000					
RE 3	0,084	0,242	0,098	0,353	0,976	1.000				
NIR 1	0,068	0,235	0,089	0,336	0,935	0,952	1.000			
NIR 2	0,070	0,232	0,102	0,366	0,972	0,990	0,952	1.000		
SWIR 1	0,648	0,719	0,738	0,828	0,523	0,420	0,410	0,459	1.000	
SWIR 2	0,808	0,835	0,873	0,877	0,348	0,220	0,210	0,241	0,939	1.000

very low (<0.30)
low (0.30-0.50)
moderate (0.50-0.80)
high (0.80-0.90)
very high (>0.90)

f



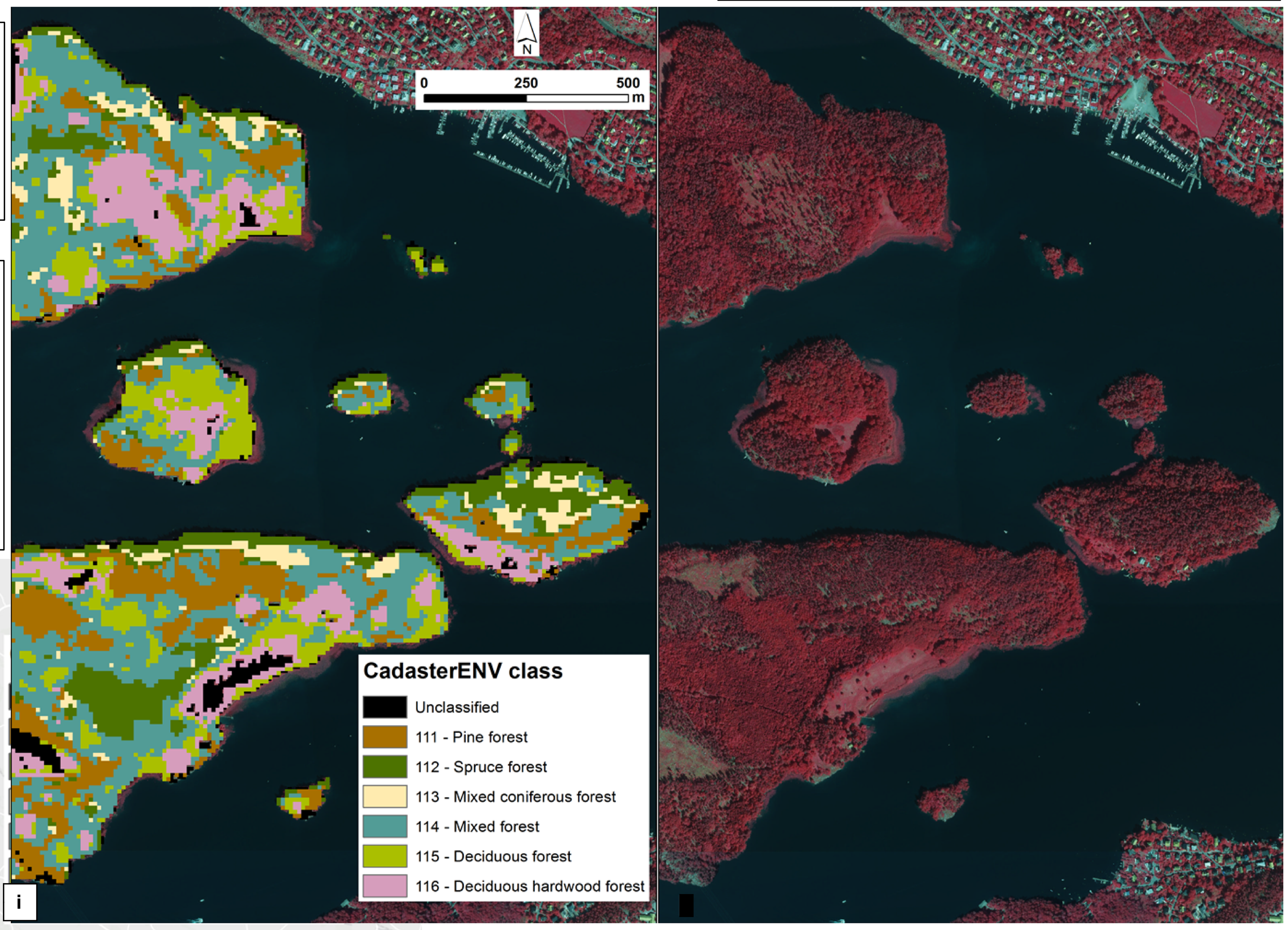
METHODS

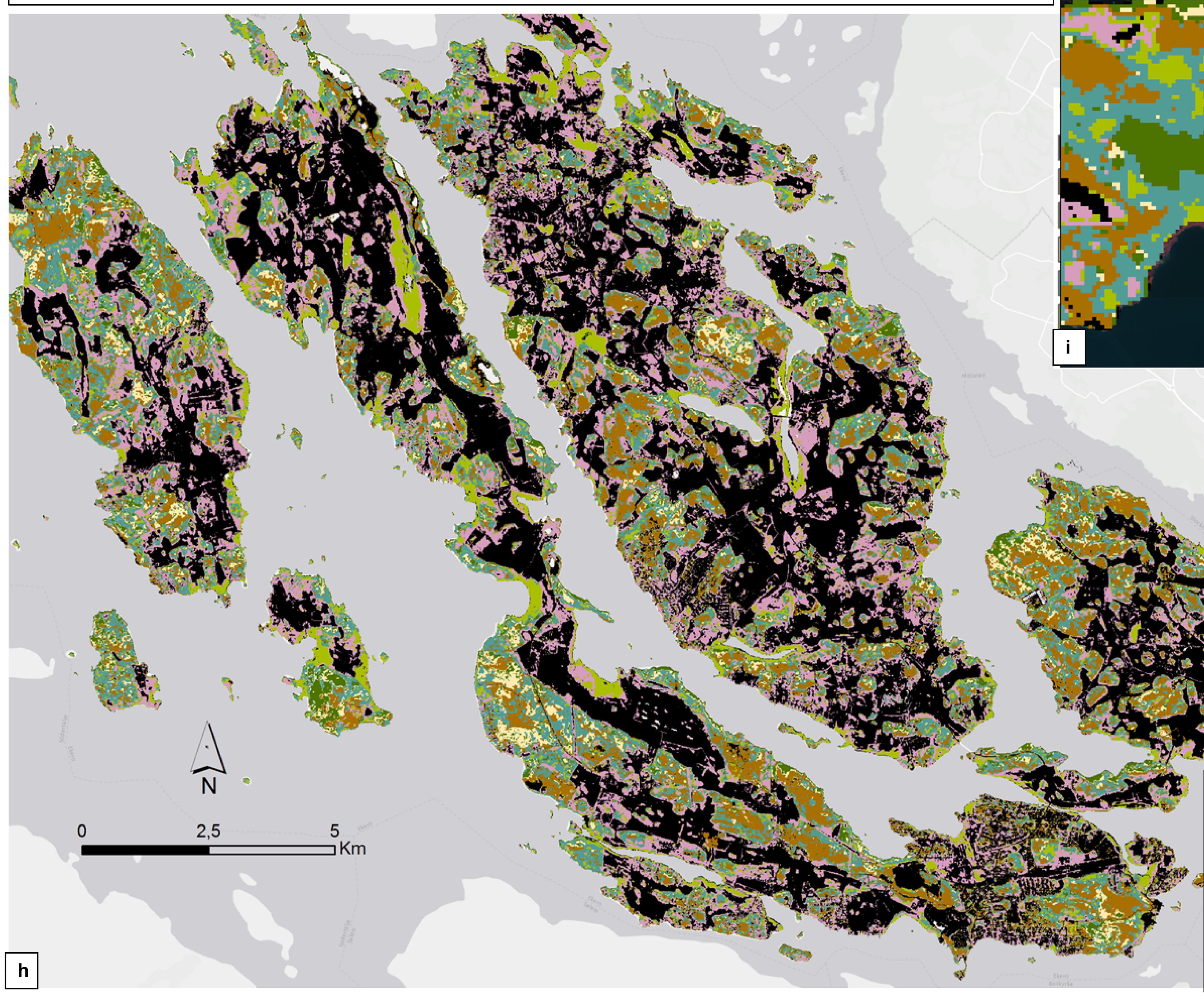
Sentinel-2 imagery from May 2, July 21, and August 28, 2016 were selected for this study as they contain negligible cloud/haze cover and occur within the vegetation period. A modified Topographic Wetness Index, a relative measure of moisture status, was included as an ancillary input - a combination of 2 rasters with different weights: a Depth-to-water (70% weight) containing soil data, and a DEM-derived Topographic Wetness Index (TWI) (30% weight). Together, the combined product called Soil Topographic Index (STI) provides information on both topographic and edaphic conditions, a proxy for soil transmissivity, allowing for better predictions of potentially saturated areas where soils are not uniform. Spectral and ancillary data were extracted to 663 field reference points. A random forest model was then fitted to this data, and the OOB error used to determine optimal input parameters in terms of band combinations, time series, and ancillary data. The fitted model with the lowest OOB error was used to produce a classification raster, verified via 10-fold cross validation.

RESULTS

An overall map accuracy of 86.0% was acheived. Producer's accuracy varied between 90.8% for deciduous hardwood forest and 81.7% for mixed coniferous forest, and User's accuracy varied between 93.2% for deciduous hardwood and 81.8% for mixed coniferous-deciduous forest. Classifications performed using 2 images were on average over 5.3% more accurate than single-date acquisitions. It was determined that the May 2 / July 21 combination yielded the highest map accuracies. The addition of the 3rd, late summer (August 28) image reduced overall accuracy slightly. The inclusion of the modified STI increased overall accuracy in multitemporal models by nearly 2%, and higher in deciduous classes (up to 3.3% avg increase).

To determine variable importance, 2 variations of a Recursive Feature Elimination approach were evaluated. First, random forest's own variable importance ranking was tested. Second, variables were manually removed stepwise on the basis of their OOB error, averaged over 5 trials. Both methods produced similar results. The lowest error rates were achieved with a 4-band combination: 1 in the visible range, 2 in the red edge range, and 1 shortwave infrared. The specific bands amongst these producing the best results were the visible red (b4), red edge b6 and b7, and SWIR2 (b12).





DISCUSSION

As different plant species respond to biological processes such as pigmentation and senescence in unique ways and at different rates, it follows that using multitemporal imagery to capture differences in foliar presentation would help separate forest classes that may be spectrally similar in any 1 single acquisition image. Although a time series generally improves accuracy, there appears to be a limit in certain instances where once reached, there is a tradeoff between information efficacy and redundancy, where additional scenes contain irrelevant details that may effect the ability to discriminate between classes. Similarly, excluding highly-correlated bands in the model increased accuracy by removing redundant information. The correlation coefficient matrix helps explain the value of a combination of visible, red edge, and shortwave infrared bands for forest classification by showing which spectral information is redundant to the classifier.

The results confirm that springtime acquisitions were most critical in the separation of the classes. This finding corroborates past studies which use acquisitions from the start or the end of the growing season. However, trees during autumn senescence are more susceptible to frosts and wind that may contribute to premature leaf removal. Furthermore, Sweden's location at higher latitudes can introduce noise related to different canopy illumination angles and intensity, weakening the discriminatory power of the model. As solar zenith angles increase as position moves further away from the equator and further away in time from the summer solstice, the effects of shadows increase, a reduction in the signal to noise ratio occurs, and the atmospheric path is longer, affecting the spectral distribution of the irradiance. An early-to-mid May acquisition, the approximate time of leaf foliation in Ekerö, is ~6 weeks away from the summer solstice. A mid-autumn image on the other hand, could be 14+ weeks after the summer solstice. Therefore, while a mid-autumn acquisition may have been suitable at lower latitudes, one should exercise caution when performing analysis on acquisitions at higher latitudes where these effects may affect the result.

CONCLUSIONS

The use of multitemporal data is helpful in classifying forest types that may be spectrally similar in any single time frame, though one should be careful to select the appropriate timing of the images to maximize phenological differences. Reducing the number of bands via a feature reduction approach to exclude highly-correlated spectral information may increase overall accuracy. Determining relative variable importance and calculating a correlation matrix can help to make sense of the findings. Ancillary data including topographic information may complement the spectral information in a model to produce better results, but the scale and physical characteristics of the study site should be taken into consideration. If soil data is available, calculating a Soil Topographic Index may prove useful. Other recommendations to improve accuracy in future studies include the incorporation of fuzzy logic for mixed forest classes, high resolution LiDAR data to assess stand characteristics, and textural analysis. Lastly, the random forest machine learning algorithm proved to be a simple yet powerful method for land cover classification. As it is resistant to overfitting and doesn't rely on assumptions of data distribution, it appears to be well suited to handle future remote sensing data that continues to increase in dimensionality, and is easily implemented in widely-available open source platforms such as Python and R.

MAJOR REFERENCES

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Hill, R.A., Wilson, A.K., George, M. and Hinsley, S.A., 2010. Mapping tree species in temperate deciduous woodland using time series multi spectral data. Applied Vegetation Science, 13(1), pp.86-99.

Fassnacht, F.E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L.T., Straub, C. and Ghosh, A., 2016. Review of studies on tree species classification from remotely sensed data. Remote Sensing of Environment, 186, pp.64-87.

Immitzer, M., Vuolo, F. and Atzberger, C., 2016. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. Remote Sensing, 8(3), p.166.

Verrelst, J., Rivera, J.P., Veroustraete, F., Muñoz-Marí, J., Clevers, J.G., Camps-Valls, G. and Moreno, J., 2015. Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods–A comparison. ISPRS Journal of Photogrammetry and Remote Sensing, 108, pp.260-272.

Waske, B. and Braun, M., 2009. Classifier ensembles for land cover mapping using multitemporal SAR imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 64(5), pp.450-457.