Objectives: The objectives of this exercise were as follows:

- To visually interpret the different satellite images of the project area.
- To estimate forest volume using the random forest method.
- To estimate species-specific volume using the k-nearest neighbor (KNN) method.

Methods: The following methods were used in this exercise:

- Random forest.
- KNN method (neighbor = 5).

Results:

1 Visual interpretation of satellite image:

Comparison between the PALSAR-2 HH and HV against the base map:

As can be observed in Figure 1 (a, b, and c), forest areas of both PALSAR-2 HH and PALSAR-2 HV could be distinguished. The whitish area of the base map (a) representing the forest and the grey color of PALSAR-2 HH (b) and PALSAR-2 HV (c) can be matched well. However, areas other than forest areas are partially differentiated in PALSAR-2 HH as the density of black color representing non-forest areas is relatively low. In contrast, the non-forest areas are more distinguishable in PALSAR-2 HV as more dense blackish colors are present in those areas. Similarly, it can be observed that the density of non-forest areas in the land cover map (d) matches well with those areas of the base map, but the density of non-forest areas is combined lakes with farm areas, and lake areas overlap the farm areas. Therefore, non-forests represent more lake areas and less farm areas. Additionally, it seems that the forest areas are slightly overestimated in the land cover map. Relatively HV has a higher contrast to HH polarization.

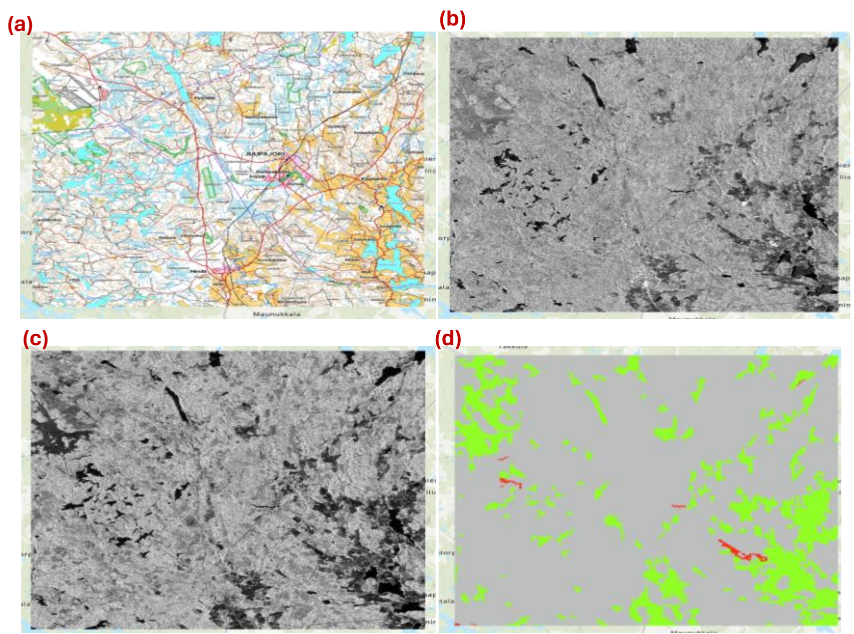

Figure 1:- Figures depicting the differences between different land features: (a) shows the base of Juupajoki area, (b) PALSAR HC, (c) PALSAR HV, and (d) PALSAR-LC. In PALSAR-LC, grey color represents the forest area, green represents the lakes, and red represents the non-forested lands.

2 Forest volume estimate using the random forest method:

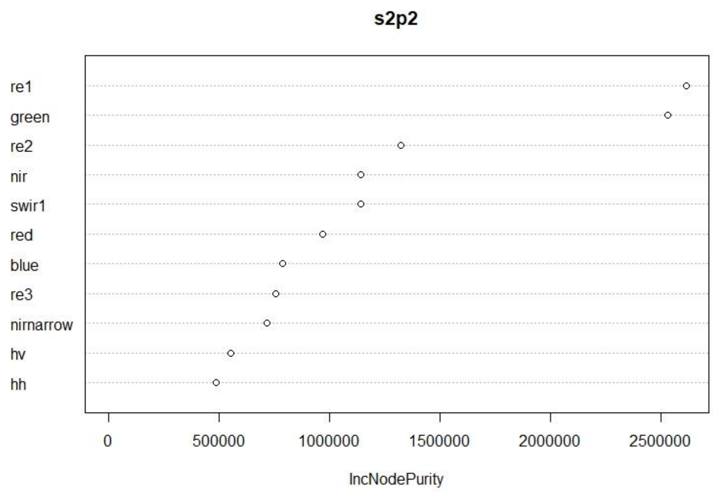

Figure 2:- Variables importance for s2p2 model.

InNodePurity determines the most suitable variables based on their importance for the model. Here higher InNodePurity means the more important variables. Therefore, among other variables re1, and green variables have significantly higher values than others (Figure 2) signifying their importance for the Random Forest algorithm and should be there as predictors in random forest. InNodePurity is obtained as a reduction in node impurities and measured by the Gini Index from one variable and then averaged over all decision trees (Carvajal et al., 2018).

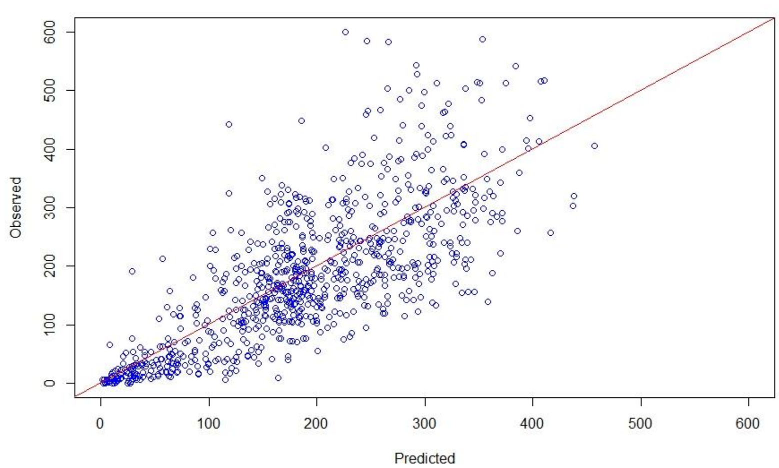

Figure 3:- Predict vs Observed plot of Random Forest.

Random forest predicted well as the residuals of the model aligned with the diagonal line (Figure 3). Model RMSE and relative RMSE are attached to Table 1.

Table 1:- Random Forest RMSE and relative RMSE

RMSE	Relative RMSE
86.51051	0.4640539

Random forest in comparison to our referenced NFI volume, it can be seen that Random forest predicted well for all forests i.e., high numbers of saplings, and middle-aged trees, and low numbers of old-aged and below middle-aged trees (poles). In a real forest, this distribution matches quite well. Other than this, RF prediction and NFI (NFI predicted high for poles) prediction correspond to each other, where RF's mean volume is slightly above than that of NFI (Table 2).

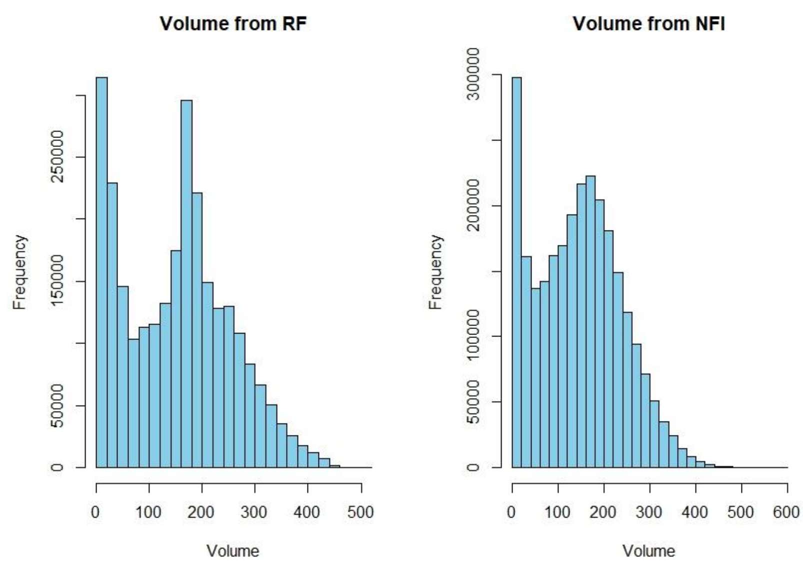

Figure 4:- Figure showing a comparison between volume prediction by Random Forest (left) and referenced volume prediction by NFI (right).

Table 2:- Mean values of Random Forest and NFI.

Mean RF (m^3/ha)	Mean NFI (m^3/ha)
151.84	145.65

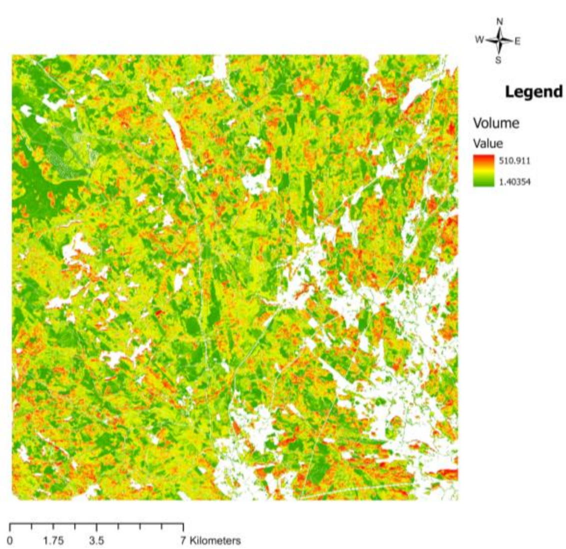

Figure 5:- Random Forest volume prediction in the project area.

3 Species-specific volume estimation with the k nearest neighbor method:

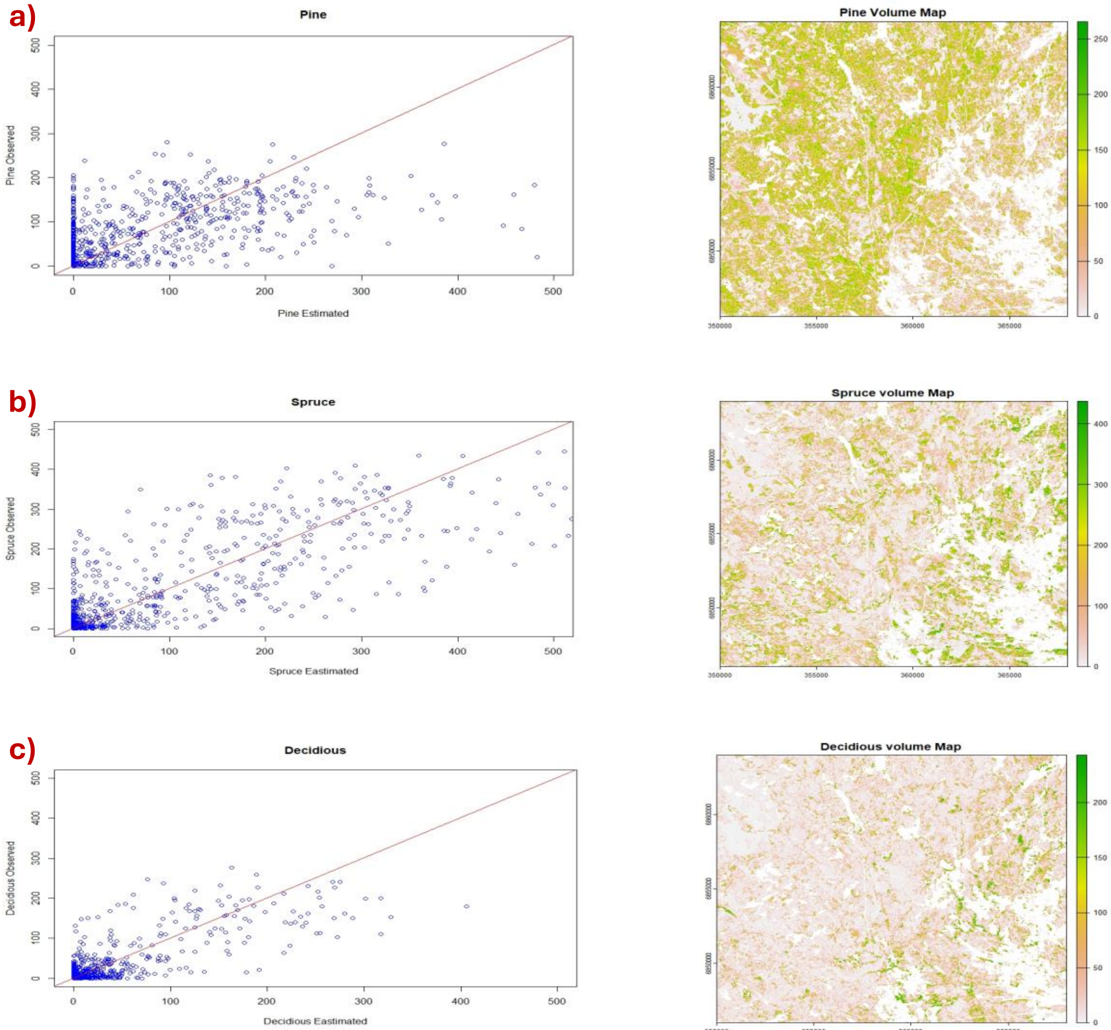

Figure 6:- Figure showing: (a) Pine prediction vs Observed volume (left), and pine volume map (right), (b) Spruce prediction vs Observed volume (left), spruce volume map (right) and (c) Deciduous prediction volume (left), and deciduous volume map (right). The model used K = 5 neighbors for the prediction of these.

For the KNN method, I used nearest neighbors (5) for modeling, and the variables used *were* ***green, nir, nirnarrow, swir1, and swir2*.**

Table 3:- Each tree model accuracy.

Species	RMSE	Relative RMSE
Pine	74.04731	1.111973
Spruce	81.26861	0.88539
Deciduous	36.68648	1.228463

Considering all model accuracy model parameters, it can be concluded that spruce has the lowest relative RMSE than that of both other species. This indicates that KNN (K = 5) predicted a good fit for spruce.

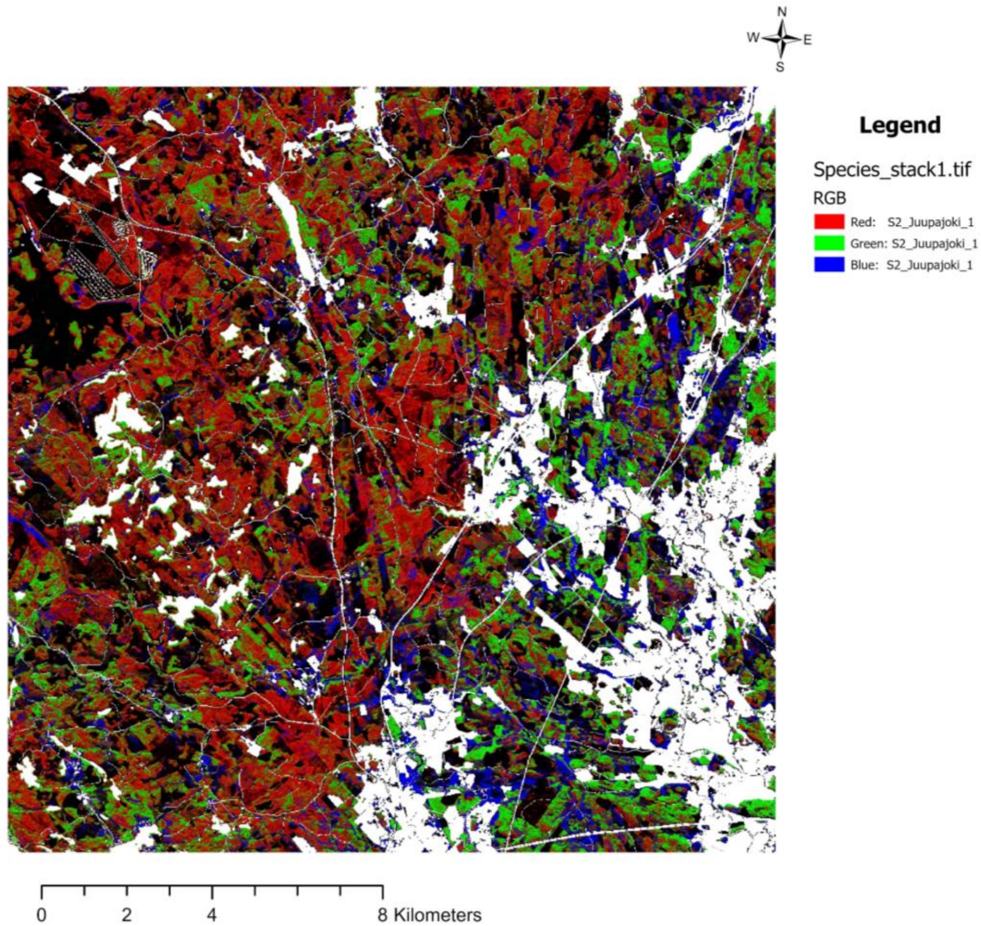

Figure 7:- Figure showing species stacked volumes in the project area. Red color represents Pine, Green color represents Spruce, and blue color represents Deciduous.

Discussion:

I also computed KNN with different neighbors (K = 8, and 10) where the variable changed to red, re1, nir, swir1, and swir2. When neighbors' numbers are increased spruce volume increases in the water bodies areas when comparing the GIS maps. However, model accuracy did not vary so much more than that of neighbors (k = 5), but it is quite unclear how many neighbors there should be for determining the right variables for the right models. As with KNN, selecting the neighbors' numbers seems challenging.

References:

Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D. M., & Watanabe, K. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC infectious diseases*, *18*(1), 183.https://doi.org/10.1186/s12879-018-3066-0