**PEM : Random Forest Analysis Notes**

**October 22nd 2018**

Contents

[1. Setting up your machine 1](#_Toc527982541)

[1.1 Folder structure for PEM 1](#_Toc527982542)

[1.2 Field data preparation 2](#_Toc527982543)

[1.3 Layer file prep. 2](#_Toc527982544)

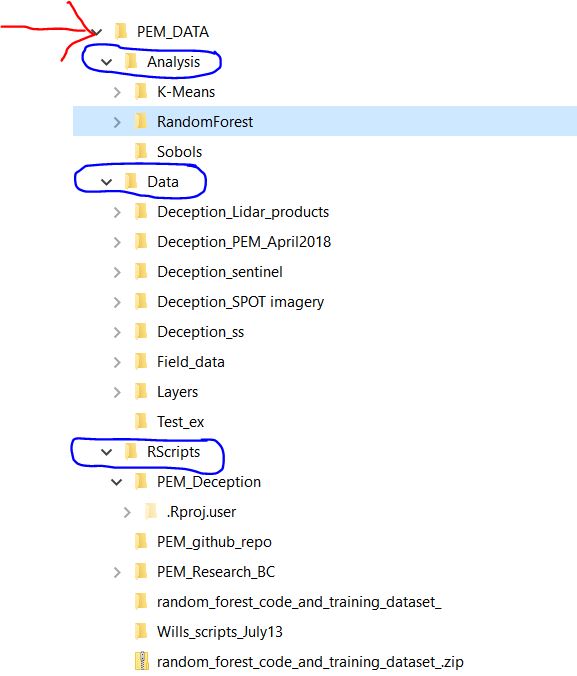
[2. Random Forest Analysis 4](#_Toc527982545)

[References: 7](#_Toc527982546)

# 1. Setting up your machine

## 1.1 Folder structure for PEM

To make it easier to run the code set up your folders in the following structures.



## 1.2 Field data preparation

- Currently there are no points in SBS dk area.

- Need to clean up the site series/ forest and non/forest classifications.

## 1.3 Layer file prep.

|  |  |  |  |
| --- | --- | --- | --- |
| Layer Name | Scale | Location |  |
| AnisotropicHeating.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  | 10m | C:\PEM\_DATA\Data\Layers\Dec\_North\_10m\layers |  |
|  | 5m |  |  |
|  | 1m |  |  |
| Dec\_dem\_BCALbers\_25m.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  | 10m | C:\PEM\_DATA\Data\Layers\Dec\_North\_10m\layers |  |
|  |  |  |  |
| GeneralCurvature.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  | 10m | C:\PEM\_DATA\Data\Layers\Dec\_North\_10m\layers |  |
|  |  |  |  |
| MultiResValleyBottomFlatness.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  | 10m | C:\PEM\_DATA\Data\Layers\Dec\_North\_10m\layers |  |
|  |  |  |  |
| Openness\_Negative.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  | 10m | C:\PEM\_DATA\Data\Layers\Dec\_North\_10m\layers |  |
|  |  |  |  |
|  |  |  |  |
| Openness\_Positive.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  | 10m | C:\PEM\_DATA\Data\Layers\Dec\_North\_10m\layers |  |
| Slope.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  | 10m | C:\PEM\_DATA\Data\Layers\Dec\_North\_10m\layers |  |
| TerrainRuggedness.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  |  |  |  |
| TopographicPosition.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  | 10m | C:\PEM\_DATA\Data\Layers\Dec\_North\_10m\layers |  |
|  |  |  |  |
| TWI.tif | 25m | C:\PEM\_DATA\Data\Layers\Dec\_North\_25m\layers |  |
|  | 10m | C:\PEM\_DATA\Data\Layers\Dec\_North\_10m\layers |  |
|  |  |  |  |
| TRIM |  |  |  |
| LIDAR - |  |  |  |
| Canopy Height Model |  |  |  |
| Sentinel – Wetlands |  |  |  |
| Sentinel - NDVI |  |  |  |
| SPOT |  |  |  |

# 2. Random Forest Analysis

To consider: Points of note (after reading and talking to Peter Ott)

**1)** No need for testing and training split. This is relevant if using a Out of Bag metric as this is effectively a continual subsampling of data with replacement.

**2)** Unbalanced samples

Peter’s comments: The correct terminology for dealing with imbalance seems to be ‘undersampling’ (the common classes) and/or ‘oversampling’ (the rare classes). Other approaches include altering the decision rule so that the cost of false positives and false negatives are not equal. I found the links below helpful, but please don’t feel committed to read all of them:

<https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/>

<https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf>

<https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation>

Here is one explanation (top of page 7) of the required adjustment on the predicted probability to handle undersampling:

<http://www2.sas.com/proceedings/forum2007/073-2007.pdf>

The good news is that randomForest seems to handle undersampling as an argument, and from what I can tell also performs the necessary adjustments automatically. I did not find a simple way to introduce differential costs to weight misclassification in RF. I think undersampling will be the most straightforward way to go.

**3)** Small sample size of site series (rare).

There are 33 siteseries (ss) with 1 sample and 64 ss with less than 5. Total of 119 Site series (ESSF/SBS). In order to predict these surfaces we may need to assess whether these are included at all? Unlikely that model can predict based on single surface. Something to thing about…

Possible ways to deal with this

- Aggregate up to broader groups

- supplement field points with “spatially selected sites” based on imagery.

## 2.1 Model fit and assessment:

**1)** Out of Bag omission error. This is 1- proportion of incorrect assigned classes.

- proportion of correctly classified units based on OOB test data (i.e. = 1 - OOB error rate).

https://stats.stackexchange.com/questions/30691/how-to-interpret-oob-and-confusion-matrix-for-random-forest

Overall Percent Correctly Classified (PCC)

Commission Rate = predicted to be in a location where not observed (FP)

Omission Rate = Observed in a location but not predicted (FN)

A number of other metrics include

The accuracy can be defined as the percentage of correctly classified instances (TP + TN)/(TP + TN + FP + FN). where TP, FN, FP and TN represent the number of true positives, false negatives, false positives and true negatives, respectively.

also you can use standard performance measures:

Sensitivity = TP / TP + FN

Specificity = TN / TN + FP

Precision = TP / TP + FP

True-Positive Rate = TP / TP + FN

False-Positive Rate = FP / FP + TN

True-Negative Rate = TN / TN + FP

False-Negative Rate = FN / FN + TP

For good classifiers, TPR and TNR both should be nearer to 100%. Similar is the case with precision and accuracy parameters. On the contrary, FPR and FNR both should be as close to 0% as possible.

- Kappa co-efficient (+SD)

This measures the agreement between classification and observed. A kappa value of 1 represents perfect agreement, while 0 represents no agreement.

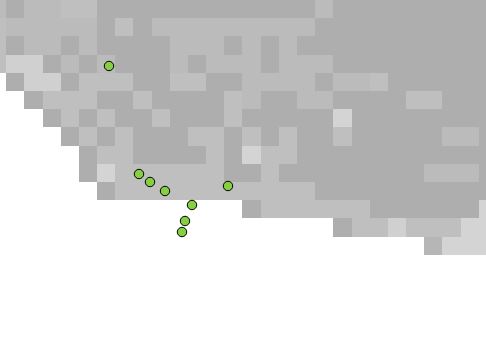
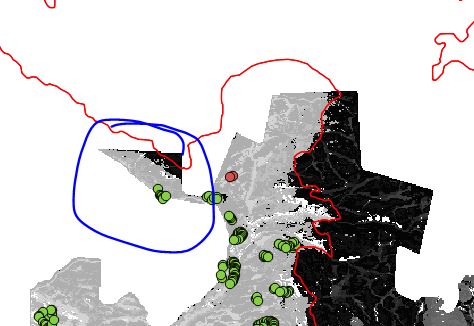
- MAUC (contribution of multiple AUC … cant get this to work properly)

**Notes to date:**

- smaller PCC does not mean it is overall a better model. For example running a subset of the model with experience not = 1. This increases the PCC value although the overall map is poorer fit as the 1 catergory only sampled in some BCGs.

- map comparisons will only work when the map\_keys are the same (ie. Subtract raster 1 from raster 2). If the subsets of SS classifications are not matching this wont line up. In progress of fixing this problem by ensuring maps have common key.

- three points are located outside of the map

These will need to be dropped.

Other metrics / options to consider

The ROC is a plot of sensitivity (true positive rate) against specificity (false positive rate). Poor model performance (i.e., where predictive ability is essentially random) returns a near-diagonal ROC plot (true positive rate equal to false positive rate). The area under ROC curve ranges from 0.5 (poor) up to 1. Producer’s accuracy (or omission error, one minus producer’s accuracy) is the proportion of a land cover class on the ground (i.e., reference) that is correctly classified in the map (prediction).

- unbalanced datasets:

https://stats.stackexchange.com/questions/30691/how-to-interpret-oob-and-confusion-matrix-for-random-forest

**Most important variables**

Landsat TM band 2 was the most important predictor variable, followed closely by band 5, based on the mean decrease Gini (MDG) measure (calculated for each predictor variable as the cumulative increase in data purity associated with each decision tree node split). Bands 3 and 4 were the next most important variables, followed by NDVI variance and band 7. In comparing the variable importance ranks between the two measures, MODIS NDVI variance was ranked 7 places higher in the MDG measure compared to MDA and band 4 (near-infrared), six places higher. These bands can be considered more important with respect to increasing the purity of training data samples after splitting at decision tree nodes, but less important based on the mean decrease accuracy.

# References:

https://cran.r-project.org/web/packages/ModelMap/vignettes/VModelMap.pdf

https://pdfs.semanticscholar.org/1318/8c6e24a2d704c31f69553a2b6398ae0c8958.pdf