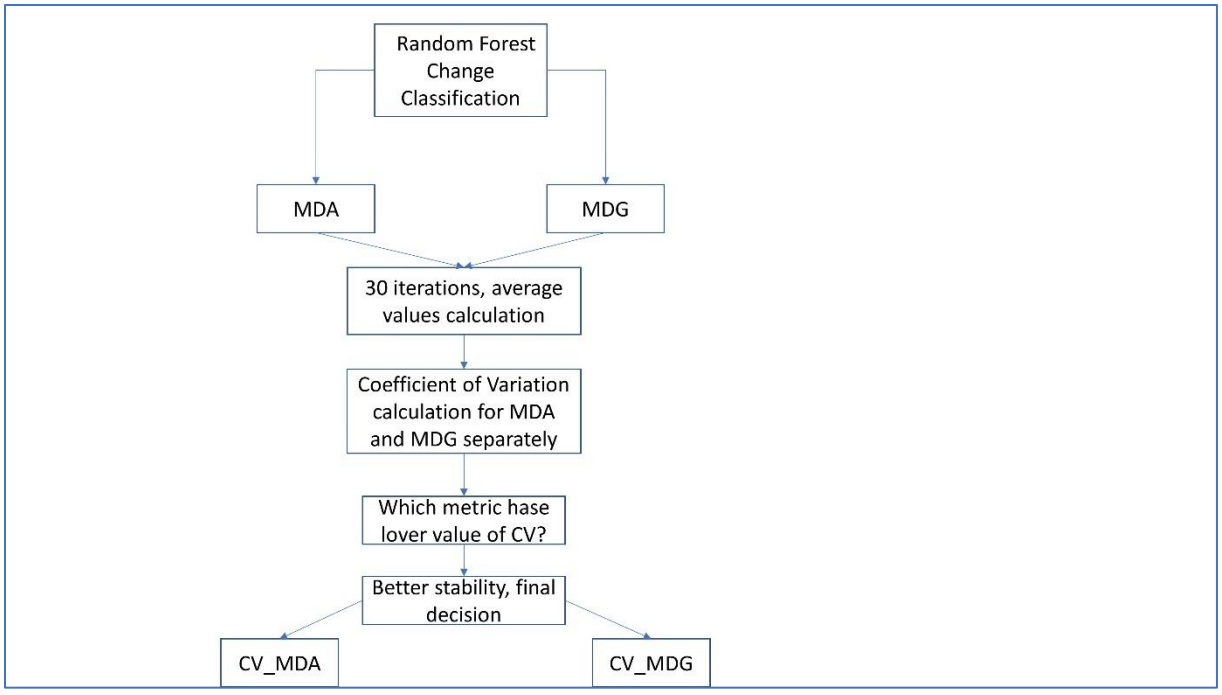
First Run

Requirements for the first run are identical to previous original version of the script, where more detailed instructions are described (root folder of this repository).

Changes in comparison to original version of the script

The original version of the script *RF_VI_TS.R* is still part of this repository so that it can still be freely downloaded an ready to use. The new version now is more automatic especially in the case when user is subjectively forced to decide which Random Forest variable important metric (Mean Decrease Accuracy or Mean Decrease Gini) is better.

Here there is located new vresion of proposed algorithm: *RF_FEATURE_SELECTION_NEW_ALGORITHM_GRASSLANDS.R* for feature selection of time series datasets from satellite data in order to map changes from grasslands to arable lands. New function was implemented that automatically decides which variable importance metric is more appropriate for further feature selection (MDA or MDG). The decision mechanism is very simple. Algorithm in the first step calculates average values for MDA and MDG metrics after 30 iterations (default value, it can be changed by user's requirements). Then coefficients of variations are calculated separately for MDA and MDG metrics. Then it is decided based on lower value of coefficient of variation in favour of MDA or MDG metric. The overall scheme of proposed algorithm is in the picture below:



New Dataset Available

We created new dataset for demonstration purposes that can be downloaded here:

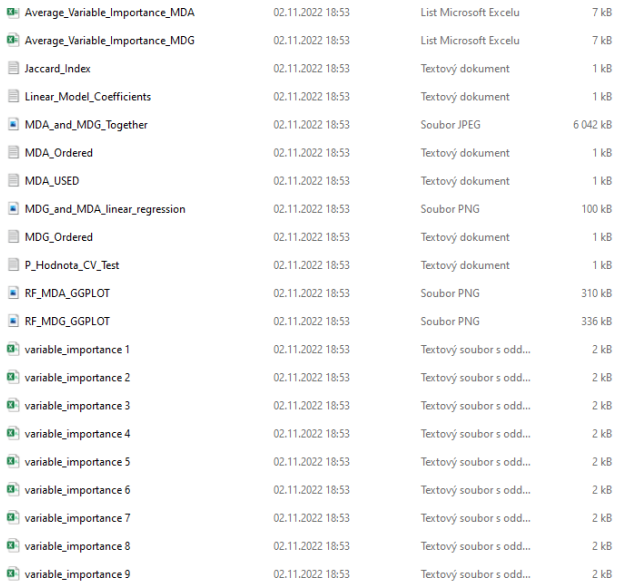
https://data.mendeley.com/datasets/26jhd7ps3v/1

Provided dataset has the same structure as original one (please see HOW TO RUN doccument i in the root directory of this repository) but provided data contain new study area in the Czech republic of Leaf Area Index. The dataset is provided as time series in the form of mutiband raster where each band represents acquisition date of time series based on the table below:

Image Number	Acquisition Date
1	30.08.2015
2	04.08.2016
3	20.06.2017
4	28.09.2017
5	06.04.2018
6	21.04.2018
7	31.05.2018
8	29.08.2018
9	18.09.2018
10	28.09.2018
11	13.10.2018
12	01.04.2019
13	16.04.2019
14	30.06.2019
15	29.08.2019
16	05.04.2020
17	20.04.2020
18	14.07.2020
19	24.07.2020
20	08.08.2020
21	13.08.2020
22	28.08.2020
23	12.09.2020

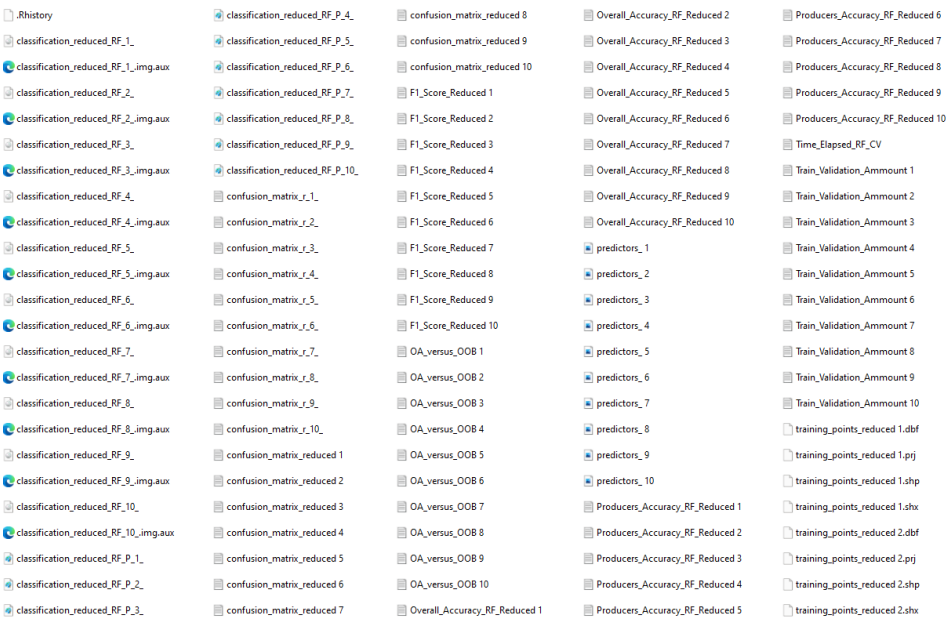
New Structure of produced files in the working directory

There are two major changes on the contrary to original version (RF_VI_TS.R). Now algorithm creates two standalone directories where different results are located. The first directory *Importance_Plots,* where new outputs are located:



Average_Variable_Importance_MDA	02.11.2022 18:53	List Microsoft Excelu	7 kB
Average_Variable_Importance_MDG	02.11.2022 18:53	List Microsoft Excelu	7 kB
Jaccard_Index	02.11.2022 18:53	Textový dokument	1 kB
Linear_Model_Coefficients	02.11.2022 18:53	Textový dokument	1 kB
MDA_and_MDG_Together	02.11.2022 18:53	Soubor JPEG	6 042 kB
MDA_Ordered	02.11.2022 18:53	Textový dokument	1 kB
MDA_USED	02.11.2022 18:53	Textový dokument	1 kB
MDG_and_MDA_linear_regression	02.11.2022 18:53	Soubor PNG	100 kB
MDG_Ordered	02.11.2022 18:53	Textový dokument	1 kB
P_Hodnota_CV_Test	02.11.2022 18:53	Textový dokument	1 kB
RF_MDA_GGPLOT	02.11.2022 18:53	Soubor PNG	310 kB
RF_MDG_GGPLOT	02.11.2022 18:53	Soubor PNG	336 kB
variable_importance 1	02.11.2022 18:53	Textový soubor s odd...	2 kB
variable_importance 2	02.11.2022 18:53	Textový soubor s odd...	2 kB
variable_importance 3	02.11.2022 18:53	Textový soubor s odd...	2 kB
variable_importance 4	02.11.2022 18:53	Textový soubor s odd...	2 kB
variable_importance 5	02.11.2022 18:53	Textový soubor s odd...	2 kB
variable_importance 6	02.11.2022 18:53	Textový soubor s odd...	2 kB
variable_importance 7	02.11.2022 18:53	Textový soubor s odd...	2 kB
variable_importance 8	02.11.2022 18:53	Textový soubor s odd...	2 kB
variable_importance 9	02.11.2022 18:53	Textový soubor s odd...	2 kB

The most important files in this directory are *MDA_Ordered.txt* or *MDG_Ordered.txt* which inform the user if MDA or MDG metric has been used for further feature selection. Another directory called *Reduced_Time_Series* that is subdirectory of *Importance_Plots* and cotains results of reduced time series dataset classification:



.Rhistory	classification_reduced_RF_P_4_	confusion_matrix_reduced 8	Overall_Accuracy_RF_Reduced 2	Producers_Accuracy_RF_Reduced 6
classification_reduced_RF_1_	classification_reduced_RF_P_5_	confusion_matrix_reduced 9	Overall_Accuracy_RF_Reduced 3	Producers_Accuracy_RF_Reduced 7
classification_reduced_RF_1_.img.aux	classification_reduced_RF_P_6_	confusion_matrix_reduced 10	Overall_Accuracy_RF_Reduced 4	Producers_Accuracy_RF_Reduced 8
classification_reduced_RF_2_	classification_reduced_RF_P_7_	F1_Score_Reduced 1	Overall_Accuracy_RF_Reduced 5	Producers_Accuracy_RF_Reduced 9
classification_reduced_RF_2_.img.aux	classification_reduced_RF_P_8_	F1_Score_Reduced 2	Overall_Accuracy_RF_Reduced 6	Producers_Accuracy_RF_Reduced 10
classification_reduced_RF_3_	classification_reduced_RF_P_9_	F1_Score_Reduced 3	Overall_Accuracy_RF_Reduced 7	Time_Elapsed_RF_CV
classification_reduced_RF_3_.img.aux	classification_reduced_RF_P_10_	F1_Score_Reduced 4	Overall_Accuracy_RF_Reduced 8	Train_Validation_Ammount 1
classification_reduced_RF_4_	confusion_matrix_r_1_	F1_Score_Reduced 5	Overall_Accuracy_RF_Reduced 9	Train_Validation_Ammount 2
classification_reduced_RF_4_.img.aux	confusion_matrix_r_2_	F1_Score_Reduced 6	Overall_Accuracy_RF_Reduced 10	Train_Validation_Ammount 3
classification_reduced_RF_5_	confusion_matrix_r_3_	F1_Score_Reduced 7	predictors_ 1	Train_Validation_Ammount 4
classification_reduced_RF_5_.img.aux	confusion_matrix_r_4_	F1_Score_Reduced 8	predictors_ 2	Train_Validation_Ammount 5
classification_reduced_RF_6_	confusion_matrix_r_5_	F1_Score_Reduced 9	predictors_ 3	Train_Validation_Ammount 6
classification_reduced_RF_6_.img.aux	confusion_matrix_r_6_	F1_Score_Reduced 10	predictors_ 4	Train_Validation_Ammount 7
classification_reduced_RF_7_	confusion_matrix_r_7_	OA_versus_OOB 1	predictors_ 5	Train_Validation_Ammount 8
classification_reduced_RF_7_.img.aux	confusion_matrix_r_8_	OA_versus_OOB 2	predictors_ 6	Train_Validation_Ammount 9
classification_reduced_RF_8_	confusion_matrix_r_9_	OA_versus_OOB 3	predictors_ 7	Train_Validation_Ammount 10
classification_reduced_RF_8_.img.aux	confusion_matrix_r_10_	OA_versus_OOB 4	predictors_ 8	training_points_reduced 1.dbf
classification_reduced_RF_9_	confusion_matrix_reduced 1	OA_versus_OOB 5	predictors_ 9	training_points_reduced 1.prj
classification_reduced_RF_9_.img.aux	confusion_matrix_reduced 2	OA_versus_OOB 6	predictors_ 10	training_points_reduced 1.shp
classification_reduced_RF_10_	confusion_matrix_reduced 3	OA_versus_OOB 7	Producers_Accuracy_RF_Reduced 1	training_points_reduced 1.shx
classification_reduced_RF_10_.img.aux	confusion_matrix_reduced 4	OA_versus_OOB 8	Producers_Accuracy_RF_Reduced 2	training_points_reduced 2.dbf
classification_reduced_RF_P_1_	confusion_matrix_reduced 5	OA_versus_OOB 9	Producers_Accuracy_RF_Reduced 3	training_points_reduced 2.prj
classification_reduced_RF_P_2_	confusion_matrix_reduced 6	OA_versus_OOB 10	Producers_Accuracy_RF_Reduced 4	training_points_reduced 2.shp
classification_reduced_RF_P_3_	confusion_matrix_reduced 7	Overall_Accuracy_RF_Reduced 1	Producers_Accuracy_RF_Reduced 5	training_points_reduced 2.shx

Reduced dataset provides only three the most important variables (acquisition dates) in raster time series. In our case it means that there were 23 images in original time series

dataset and reduced dateset only contains 3 of them. Other files present in this directory are the same as previous ones.