Detection of changes from permanent grassland to arable land using Sentinel 2 satellite data in R

The purpose of this project is to demonstrate the detection of changes from permanent grassland to arable land using freely available satellite data and freely available statistical software R. The main goal is to demonstrate the proposed methodological procedure, which is part of the research within the dissertation and planned scientific publication.

The proposed procedure is demonstrated on freely available satellite data from Sentinel 2 and vector data from LPISU (Land Parcel Identification System). Detection of changes takes place in the form of supervised classification. The research aim of the proposed method is the detection of the best predictors using the internal metrics of the Random Forest classifier, which are best suited for the detection of changes from grassland to arable land. It is well known that the large number of redudant predictors in Remote Sensing, whether spectral bands or other derived predictors (such as shape and texture characteristics), reduces the overall accuracy of change detection and increases the total computational time.

The methodical process of selection of suitable predictors is implemented in statistical software R and thus enables easy portability and adaptation based on user needs.

Let’s begin

All data used for demonstrators in this project can be freely downloaded from the Internet. Data from Sentinel 2 satellites are freely available after registration under the Copernicus project, managed by the European Space Agency (ESA). Reference data in the form of soil blocks from the LPISU are also freely available for the territory of the Czech Republic, where the study site is located, which in this case serves as a demonstration site.

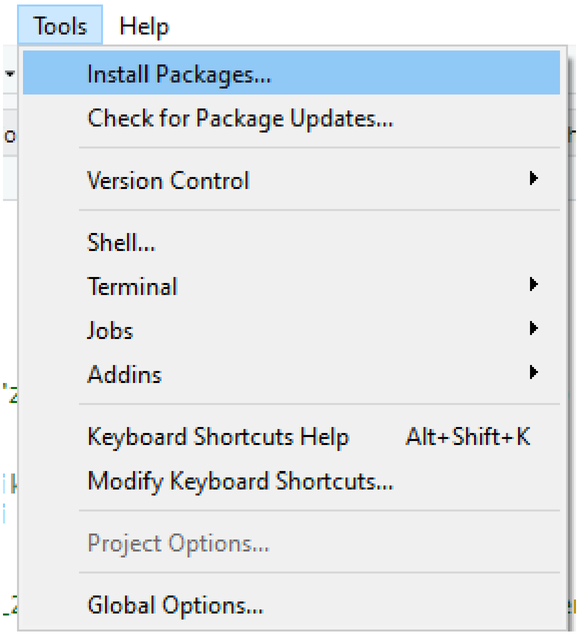
Installation

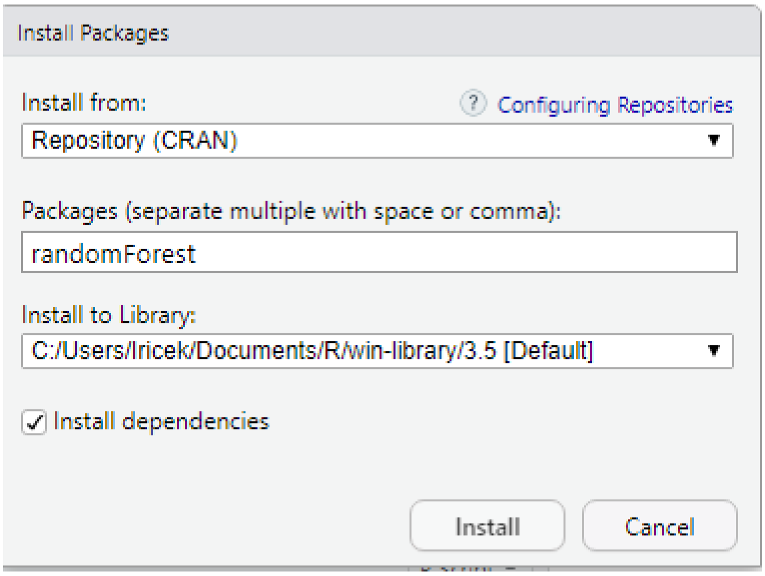
RStudio was used as the development environment for R. The R software itself used an improved version of Microft (Microsoft R Open), which includes various improvements to speed up computations. Version 3.5.3 was used.

You must first install the necessary packages, provided they are not installed in bulk:

```
install.packages(c(“randomForest”, “rgdal”, “raster”, “caret”), dependencies=T)
```

This step is not necessary provided that the affected packages are already installed. An alternative installation method is possible within RStudia via the *Tools - Install Packages* menu.




Install Packages

Install from: Repository (CRAN)

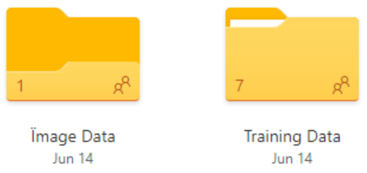
Packages (separate multiple with space or comma): randomForest

Install to Library: C:/Users/lricek/Documents/R/win-library/3.5 [Default]

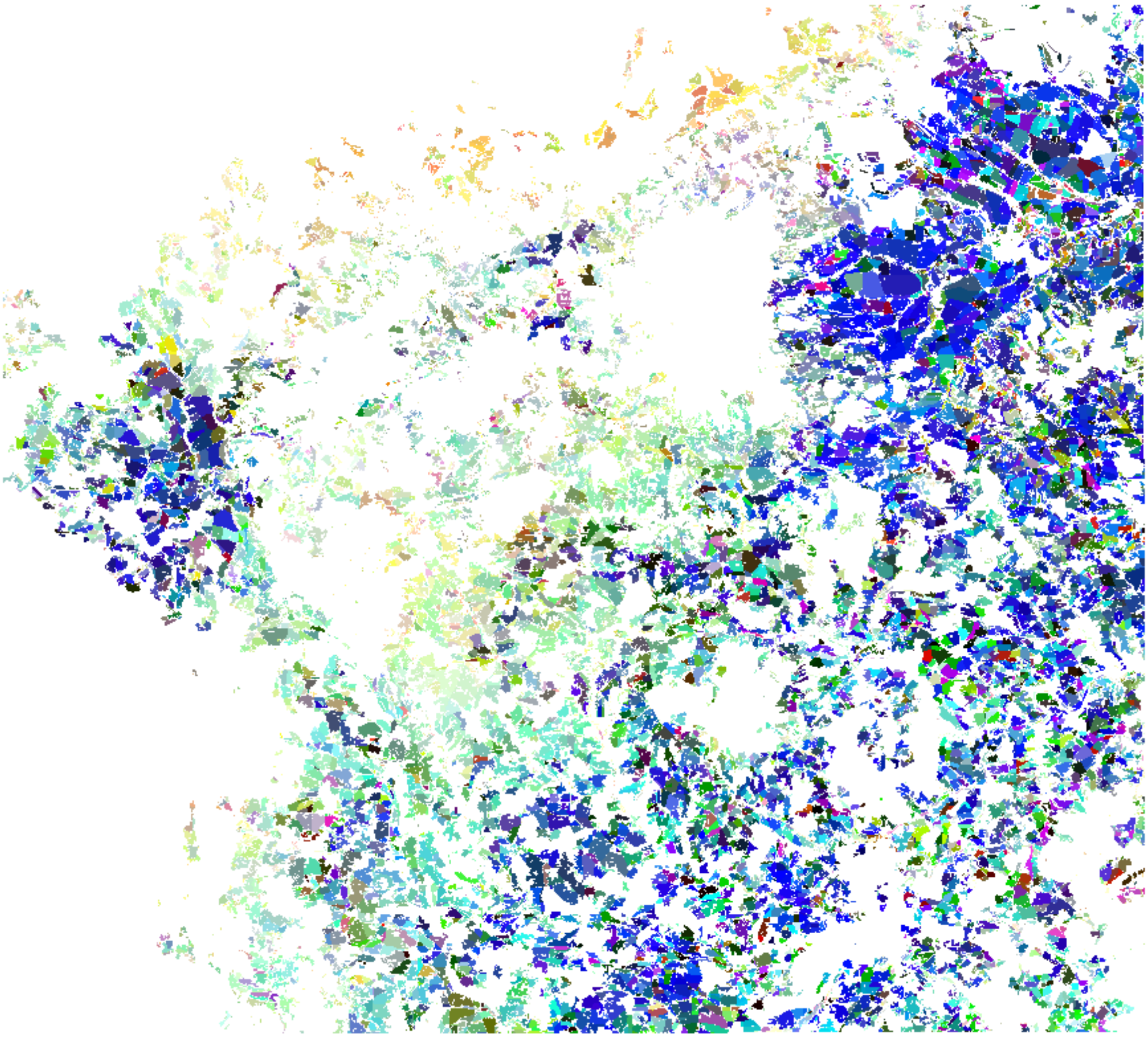
☑ Install dependencies

Input Data and their structure

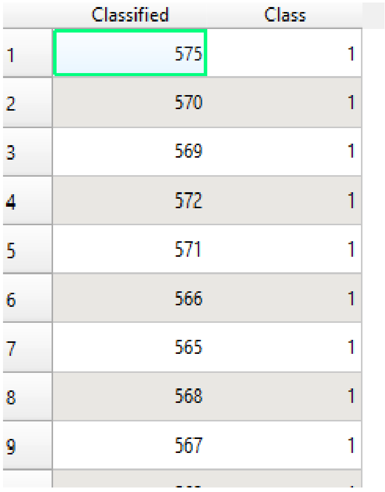
All input data can be found here. There are two separate directories in the folder - *Image Data* and *Training Data*



The *Image Data* folder contains input satellite data already modified for change classification purposes. These are soil blocks with average values of the NDVI vegetation index in raster form forming a multitemporal time series in the form of the monitored locality (see the publication).



The Training Data folder contains reference points that are intended for training the Random Forest classifier used. It is a point layer in shp format (shapefile) with the following attributes:



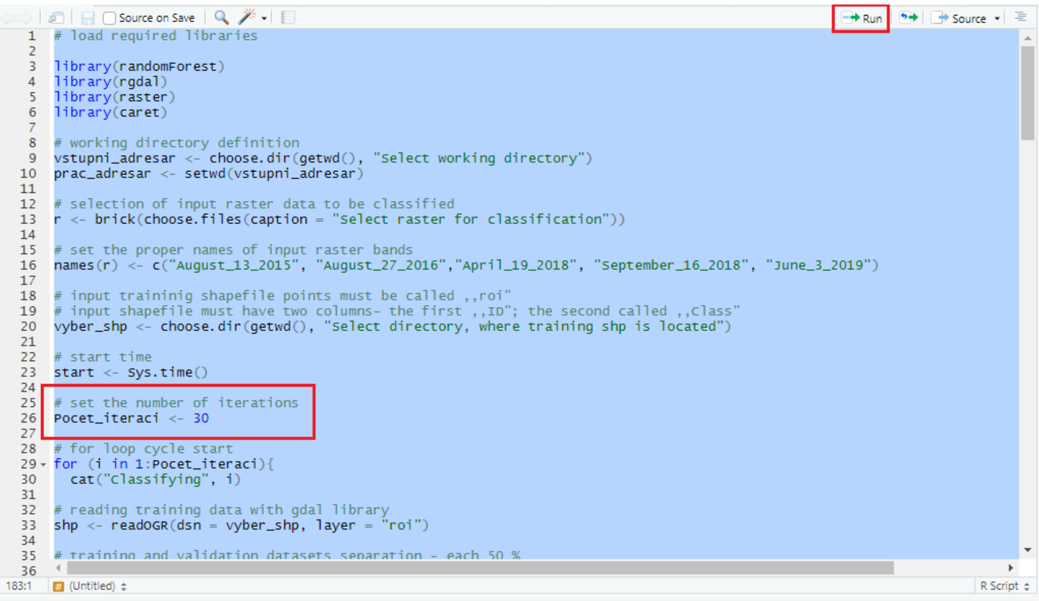
	Classified	Class
1	575	1
2	570	1
3	569	1
4	572	1
5	571	1
6	566	1
7	565	1
8	568	1
9	567	1

It is necessary to mention that the name of the training shapefile must remain "roi.shp", otherwise the script will not work properly. It is necessary to keep the strict names of the above attribute fields, including their exact order.

The entry point layer contains a total of 2000 reference points. The *Classified* attribute column is the identifier of each subpoint. The *Class* attribute column contains the numeric code of the classification classes, which takes the values 1 and 2. The value 1 determines the classification class of changes from permanent grassland to arable land, the value 2 soil blocks for which no changes have taken place.

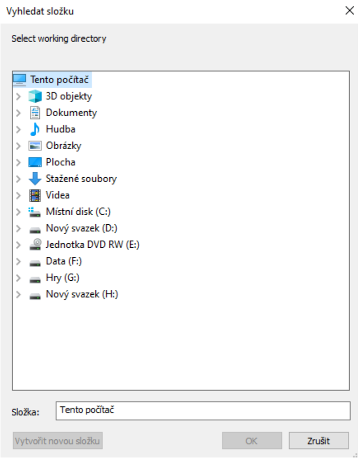
Launching the RF_VI_TS.R Script

First of all, it is necessary to mention that all comments and comments in the script are marked with a "#". Before the first run of the script, it is advisable to first set the required number of iterations, ie the number of repetitions, how many times the classification of changes from permanent grassland to arable land should be performed. This option is indicated by the variable,,*Počet_iteraci*", which is set to 30 in the basic settings. If it is desired to classify changes more or less, here everything can be set according to the user's needs.



```
# load required libraries

library(randomForest)
library(rgdal)
library(raster)
library(caret)

# working directory definition
vstupni_adresar <- choose.dir(getwd(), "Select working directory")
prac_adresar <- setwd(vstupni_adresar)

# selection of input raster data to be classified
r <- brick(choose.files(caption = "Select raster for classification"))

# set the proper names of input raster bands
names(r) <- c("August_13_2015", "August_27_2016","April_19_2018", "September_16_2018", "June_3_2019")

# input traininig shapefile points must be called ,,roi"
# input shapefile must have two columns- the first ,,ID"; the second called ,,Class"
vyber_shp <- choose.dir(getwd(), "Select directory, where training shp is located")

# start time
start <- Sys.time()

# set the number of iterations
Pocet_iteraci <- 30

# for loop cycle start
for (i in 1:Pocet_iteraci){
  cat("Classifying", i)

# reading training data with gdal library
shp <- readOGR(dsn = vyber_shp, layer = "roi")

# training and validation datasets separation - each 50 %
```

Run RStudio and open the appropriate script - *RF_VI_TS.R*. Then use the keyboard shortcut Ctrl + A to select all the code and run the script by clicking the Run button. After successful start-up, a sequence of dialog boxes will follow, where it will be necessary to enter the input data. In the first case, you must enter the path to the working directory. This is a folder in which all the results will be saved in the final after the script is completed.

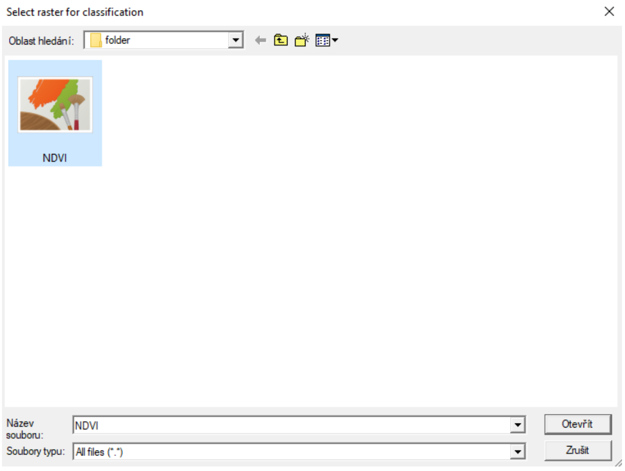


Vyhledat složku

Select working directory

Složka: Tento počítač

The following is a selection of input data that are subject to classification. This is nothing more than a multitemporal raster set of the NDVI vegetation index from the Image Data folder.



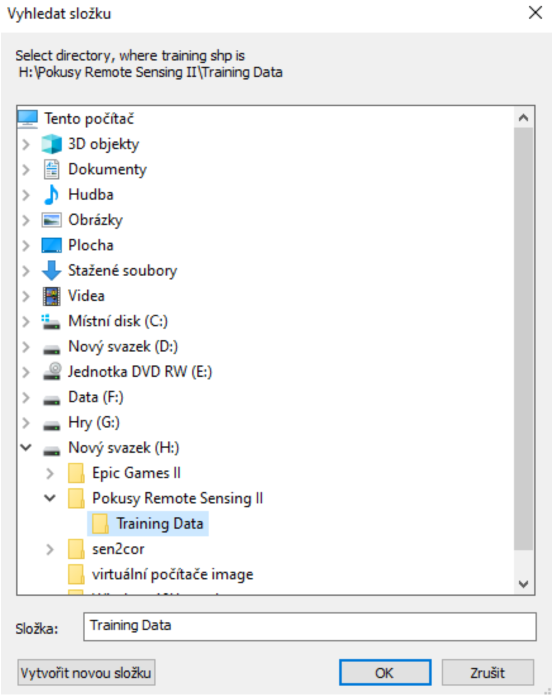
Select raster for classification

Oblast hledání: folder

Název souboru: NDVI

Soubory typu: All files (*.*)

In the last step, it is necessary to search the directory with reference points.



Vyhledat složku

Select directory, where training shp is
H:\Pokusy Remote Sensing II\Training Data

Složka: Training Data

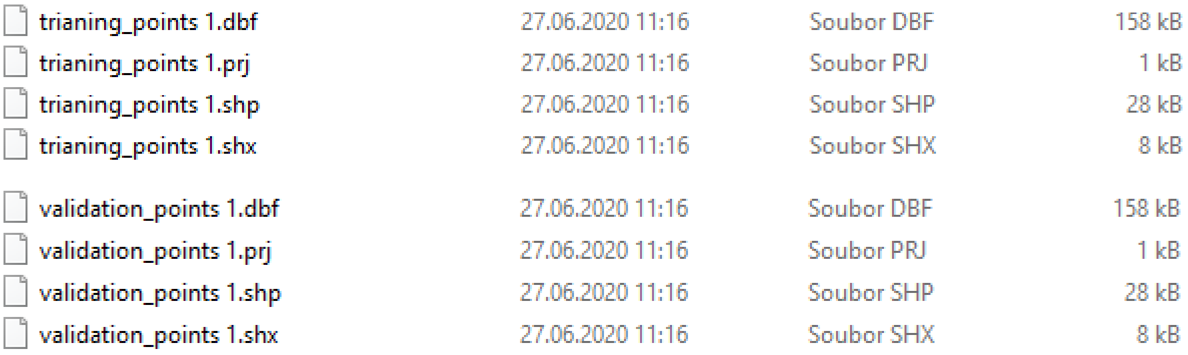
If everything was entered correctly in the previous steps, the script should run without further error messages and just wait for the calculation to finish.

Implemented script functions and their description

In summary, the implemented script functions can be summarized in individual points:

1) The division of reference points into training and validation, including their export to the hard disk in shp format
2) Optional number of classification iterations
3) Change classification itself using the Random Forest algorithm
4) Prediction of the model in the form of a categorical classification, according to the attribute column *Class* of the training data
5) Prediction of the model in the form of probabilities of classification classes
6) Export of the error matrix derived according to the metric "Out of the Bag" in .txt format to the hard drive
7) Export of predictor relevance according to MDA (Mean Decrease Accuracy), MDG (Mean Decrease Gini) in .jpg format
8) Export of numerical values of MDA and MDG in .csv format
9) Implementation of standard accuracy evaluation (Congalton et al. 2008) - error matrix, overall, user and processing accuracy, export to hard disk in .txt format
10) Merging the categorical classification - all classification results for each iteration, which were exported to the hard disk in one raster in one raster
11) Time record in the form of the total calculation time - the result is exported to the hard drive in the form of a separate .txt file after the calculation

Add 1) The division of the original set of reference points into training and validation occurs at each iteration with the appropriate names "*training_points*" and "*validation_points*" of the output shp file on the hard drive.

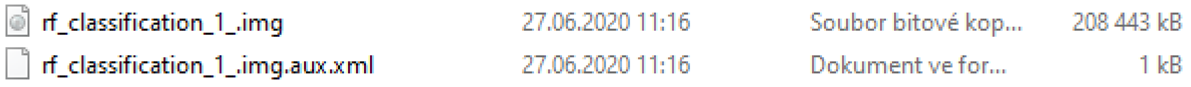


trianing_points 1.dbf	27.06.2020 11:16	Soubor DBF	158 kB
trianing_points 1.prj	27.06.2020 11:16	Soubor PRJ	1 kB
trianing_points 1.shp	27.06.2020 11:16	Soubor SHP	28 kB
trianing_points 1.shx	27.06.2020 11:16	Soubor SHX	8 kB
validation_points 1.dbf	27.06.2020 11:16	Soubor DBF	158 kB
validation_points 1.prj	27.06.2020 11:16	Soubor PRJ	1 kB
validation_points 1.shp	27.06.2020 11:16	Soubor SHP	28 kB
validation_points 1.shx	27.06.2020 11:16	Soubor SHX	8 kB

Add 2) The optional option to set the total number of iterations has been described above

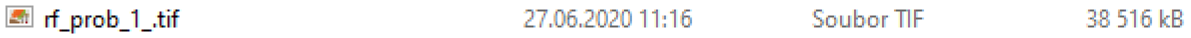
Add 3) Classification of changes is implemented using the "*randomForest*" function of the library of the same name

Add 4) The prediction of the model in the form of a categorical classification includes the classical classification of land cover, in this case the classification of changes from grassland to arable land. The result is in raster form, containing integer values in .img format (Erdas Imagine)



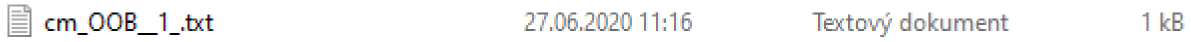
rf_classification_1_.img	27.06.2020 11:16	Soubor bitové kop...	208 443 kB
rf_classification_1_.img.aux.xml	27.06.2020 11:16	Dokument ve for...	1 kB

Add 5) Prediction of the probability of classification classes is given by the ability of the Random Forest algorithm, which determines the probability of occurrence of a given class within a pixel or objects (or other basic image units). The result is again in raster format .tif with the appropriate name



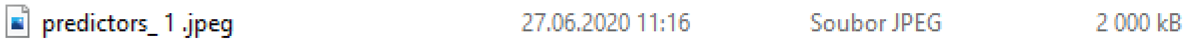
rf_prob_1_.tif	27.06.2020 11:16	Soubor TIF	38 516 kB

Add 6) Based on its internal structure, Random Forest contains the "Out of the Bag Error" metric, which is able to determine the potential accuracy of future classification - details can be found in the original text of Breiman (2001). This metric is exported in a .txt file named "*cm_OOB.txt*" and always at the end with the appropriate number of the iteration



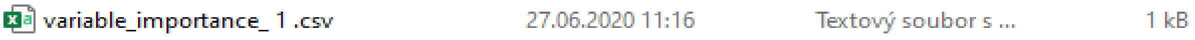
cm_OOB__1_.txt	27.06.2020 11:16	Textový dokument	1 kB

Add 7) The script offers the export of a simple graph for easy evaluation and visualization of the most important predictors within a partial iteration for both MDA and MDG metrics. A graph with a raster file in the .jpeg format has the name "*predictors* ", where at the end of its name is always the number of a specific iteration



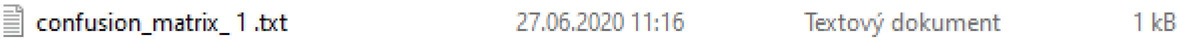
predictors_ 1 .jpeg	27.06.2020 11:16	Soubor JPEG	2 000 kB

Add 8) The numeric values of both MDA and MDG metrics can be found in a file named "*variable_importance*" in .csv format, where a number indicating the appropriate iteration is present at the end of the name.



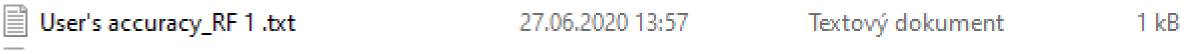
variable_importance_ 1 .csv	27.06.2020 11:16	Textový soubor s ...	1 kB

Add 9) This is an export of a standard error matrix and its basic metrics (overall, user and processing accuracy). The implemented Accuracy Assessment procedure is according to the methodologies of Congalton (1991) and Congalton and Green (2008). The complete error matrix is stored in a file named "*confusion_matrix*" with the appropriate iteration number. The corresponding metrics are also stored in the appropriate file names in .txt format for each iteration separately



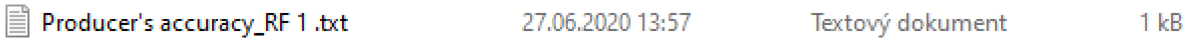
confusion_matrix_ 1 .txt	27.06.2020 11:16	Textový dokument	1 kB

User’s accuracy is located in the file called ,,*User’s accuracy_RF“*



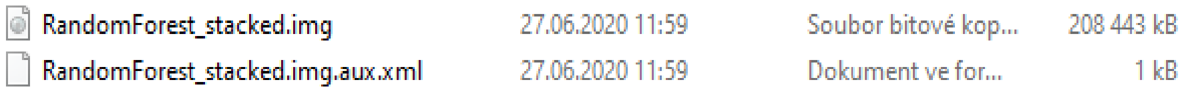
User's accuracy_RF 1 .txt	27.06.2020 13:57	Textový dokument	1 kB

Producer’s accuracy is located in the file called ,,*Producer’s accuracy“*



Producer's accuracy_RF 1 .txt	27.06.2020 13:57	Textový dokument	1 kB

Add 10) All outputs of categorical classifications for each iteration (land cover classification) are combined into one raster file at the end of the calculation. This function was implemented due to the potential implementation of post-classification adjustments - for example, the application of a median filter. File with the appropriate name "*RandomForest_stacked.img*"



RandomForest_stacked.img	27.06.2020 11:59	Soubor bitové kop...	208 443 kB
RandomForest_stacked.img.aux.xml	27.06.2020 11:59	Dokument ve for...	1 kB

Add 11) A text file named "*Overall_Time.txt*" provides information about the total calculation time

Overall_Time.txt	27.06.2020 13:57	Textový dokument	1 kB

Launching Script VI_PLOT_TOOL.R

The script calculates the average MDA and displays it in a clear graph according to the importance of individual predictors.

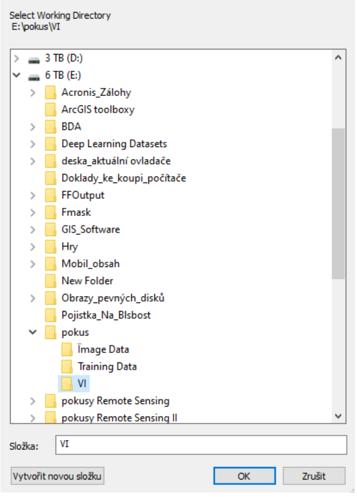
To run this script, you must first install the ggplot 2 library, provided it has not been installed before:

```
install.packages(“ggplot2”, dependencies=T)
```

Before running the script for the first time, you must first copy to a separate directory all the "variable_importance" files that were obtained by running the "*RF_VI_TS.R*" script, depending on the number of iterations required.

variable_importance_ 1 .csv	27.06.2020 11:16	Textový soubor s ...	1 kB

The script is launched in the same way as described above in the section "*Running the Script RF_VI_TS.R*". After successful startup, you will be prompted to enter the path to the working directory. This is nothing more than the folder to which the "*variable importance*" files were copied in the previous step.



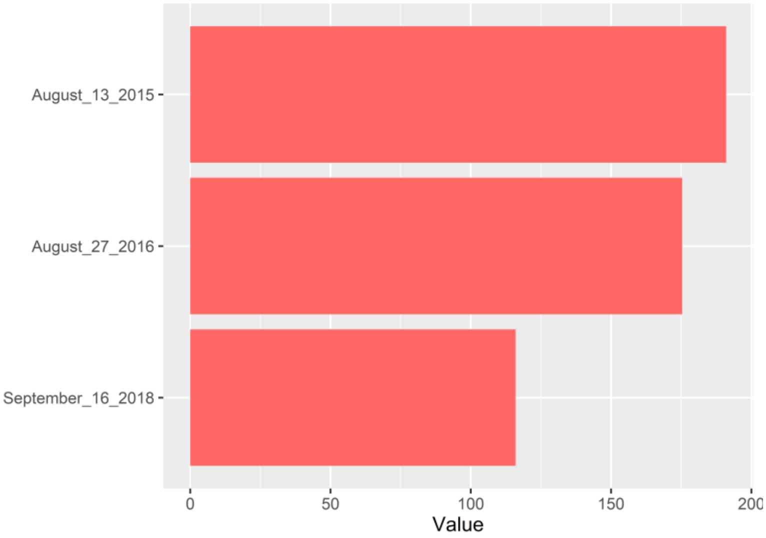
Select Working Directory
E:\pokus\VI

Složka: VI

After successful completion of the calculation, a file named "*RF_MDA_GGPLOT.png*" will appear in the selected folder.

RF_MDA_GGPLOT.png 27.06.2020 13:00 Soubor PNG 159 kB

The resulting graph shows the relevance of the individual predictors that were the subject of the classification:



Contacts

Contact: sanderajiri@outlook.com

Project Link: https://github.com/Iricek/Remote-Sensing-in-R

References

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

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, *37*(1), 35-46.

Congalton, R. G., & Green, K. (2008). *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press