ISOLATION FOREST – Time Series Anomaly Detection in Remote Sensing Data

This research project is concentrated on anomaly detection of agricultural crops in order to reveal changes from permanent grasslands to arable lands with help of optical data from freely available Sentinel 2 Satellite. The proposed methodology is implemented in freely available R programming language and it is part of Ph.D. research. For the purpose anomaly detection Isolation Forest algorithm is utilized (Liu et al. 2008).

In terms of mapping changes of vegatated areas it is very hard to distinguish abrupt ones from permanent ones. It is caused due to high phenological variability of agricultural crops. Phenological variability of agricultural crops is strongly correlated with local climate conditions and altitude.

Agricultural areas where human intervention plays major role and it strongly impacts phenological state of each agricultural crop. The main purpose of this research is to propose and demonstrate simple methodology in order to map permanent and abrupt changes of agricultural crops especially changes from permanent grasslands to arable land.

Required software and mandatory packages installation

RStudio was used as the primary software environment for programming language R. All code in attached R script was tested in enhanced R 4.0.2 version by Microsoft known as Microsoft R Open (MRAN: https://mran.microsoft.com).

The first step is to install necessary packages in order to run implemented script correctly. They can be installed in bulk:

install.packages(c("raster", "rgdal", "caret", "h2o", "randomForest", "rJava", "ggplot2", "EBImage", "xlsx"), dependencies=TRUE)

Please make sure you have proper Java installation and set its path to environment variable before installation in your computer. Very few packages require Java as background – such as h2o library (Downloading & Installing H2O — H2O 3.36.0.1 documentation).

However we do not recommend bulk approach because any of required packages might be already installed and their different versions may cause compatibility problems.

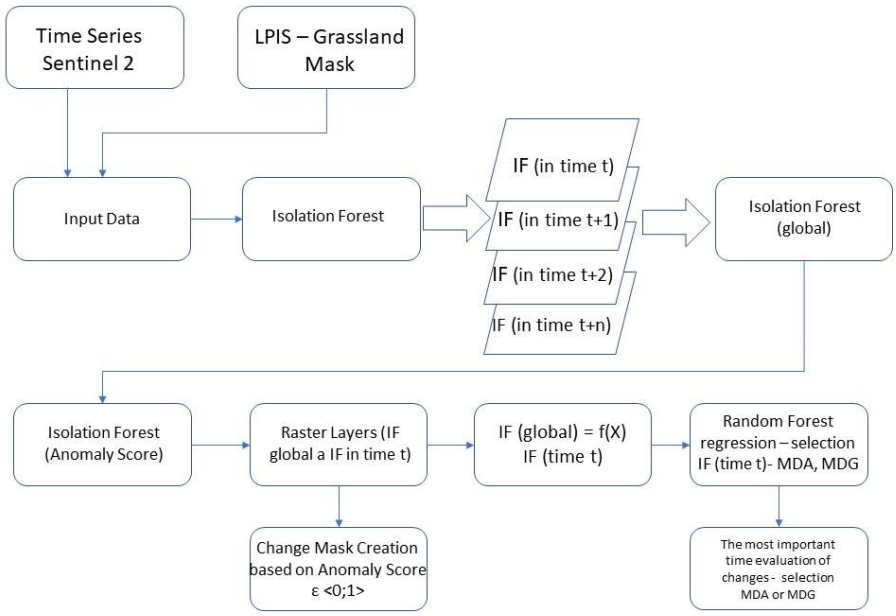
General Overview of Implemented Change Detection Process

In the first steps Sentinel 2 Time Series and LPIS vector (Land Parcel Identification System) Grasland Mask are required. Then Isolation Forest algorithm is applied on Sentinel 2 time series in global form (IF Global) in other words the whole Sentinel 2 time series raster stack. Then Isolation Forest algorithm is applied on each band in time series raster stack which represents acquisition date of original satellite imagery (IF in time t). Isolation Forest algorithm represents detected anomalies in the form of anomaly score in range between 0 and 1. Values greater than 0.5 are considered as anomalies in our case detected changes. Outputs of Isolation Forest calculation in global and local form are rasters with values in range 0 and 1. Each image mapping unit (pixel or object) has its own value of anomaly score and then it can be evaluated for changes.

Then it can be considered that global anomaly score (IF Global) is function of all anomaly scores in time t of the whole time series raster stack (IF in time t):

$$IF(global) = f(x)\ (IF\ in\ time\ t)$$

In order to find functional mathematical relationship between values of global anomaly score and other values in time t Random Forest regression is implemented so as to determine the most important image acquisition date with the highest changes represented by high values of anomaly score. The best acquisition date is determined by internal Random Forest Metrics based on MDA (Mean Decrease Accuracy) and MDG (Mean Decrease Gini) as it was shown in our previous research (Šandera, Štych 2020). Random Forest regression is calculated 30 times and then the average values of MDA and MDG are calculated in order to obtain unbiased results (Šandera, Štych 2020). The overview of implemented anomaly detection scheme is in the picture below.



Implemented functions and their description

1) Isolation Forest Anomaly Detection in global form
2) Isolation Forest Anomaly detection in time t
3) Random Forest Regression
4) Jaccard Index
5) Variable Importance Plots based on MDA and MDG
6) Linear Regression Model between MDA and MDG

Add 1) The Global Anomaly Detection for time series raster stack is represented by exported raster called ,,Anomalies_IsolationForest_H2O.img“

Add 2) Implementation of Isolation Forest anomaly detection in time t is represented by raster files called ,,anomalies_if_1.img“. The number in the file name represents connected image acquisition date in time series raster stack

Add 3) Results of Random Forest regression are connected with files: ,,RMSE_1“.txt – this file contains Root Mean Square Error metric and determines accuracy of Random Forest regression model. Numbers in the file name collocate with number of iteration that can be set manually before calculation process of the script (more information further)

Add 4) Jaccard index (Jaccard 1912) is implemented in order to compare dependency Random Forest metrics: MDA or MDG – final result is associated with filename ,,Jaccard Index.txt“

Add 5) Variable Importance Plots represent average non biased values after 30 iteratations (default) of Random Forest regression models for MDA and MDG – exported plots can be found in directory ,,Importance Plots“ with associated file names: ,,RF_MDA_GGPLOT.png“ and ,,RF_MDG_GGPLOT.png“ and both importance plots together: ,,MDA_and_MDG_together.png“

Add 6) Linear regression model is implemented between Random Forest importance metrics (MDA and MDG). Values of MDA and MDG are scaled between 0 and 1 in order to properly compare both metrics. Results are exported into text file ,,Linear_Model_Coefficients.txt“

How to run a script Isolation Forest_Anomaly Detection in Time Series of Remote Sensing Data.R

The script itself has been implemented in interactive form. It can be launched directly from R console or RStudio, however it can be launched with double click in case if the Rterm.exe located in R instalation directory is set as environment variable. This is only possible in Microsoft Windows not in Linux or Apple Macintosh.

In case all required packages are installed properly there should be no errors and script should require all necessary inputs which are: Time Series Raster stack with chronological oredered bands based on their time stamps (for example NDVI vegetation index). Proper .csv file associated with dates of original satellite data acquisition. If everything is correct R script should start calculations based on functions decribed above.

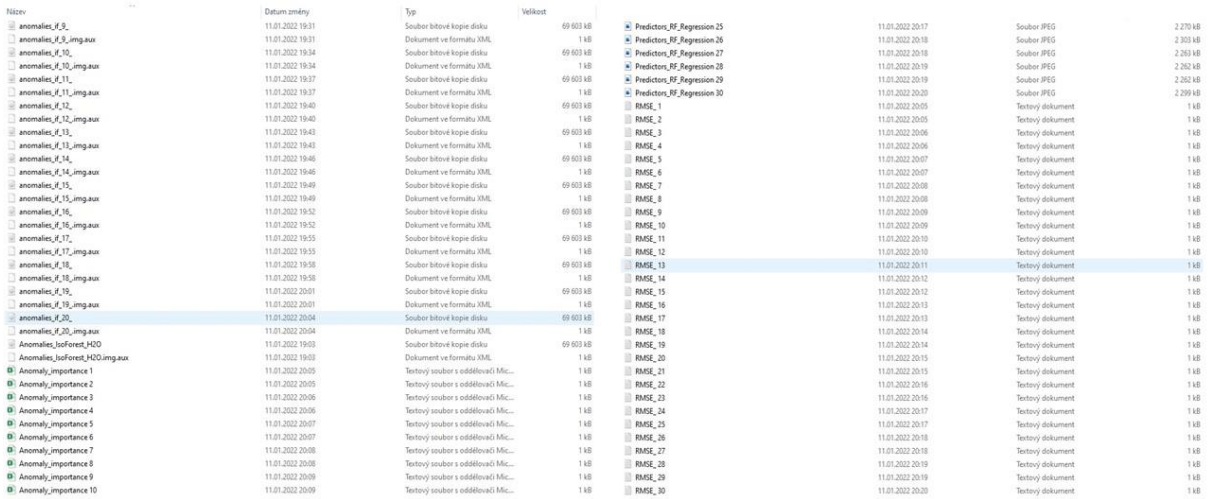
Sample Data Description

All sample data can be found here. The sample data is biological variable FAPAR from Sentinel 2 Time Series raster stack with 20 m spatial over small area located in Czech republic. FAPAR raster stack contains land parcels with permanent grasslands from year 2018 that have been masked from original FAPAR raster images with help of LPIS (Land Parcel Identification Systém – more info here) vector polygons that can be freely downloaded. The .csv file located in the same directory ,,Names of Time Series Raster Bands.csv“ contains time stamps of the time acquisition of original images which are present as associated raster band names in FAPAR time series raster stack. This file is mandatory in order to run attached R script properly

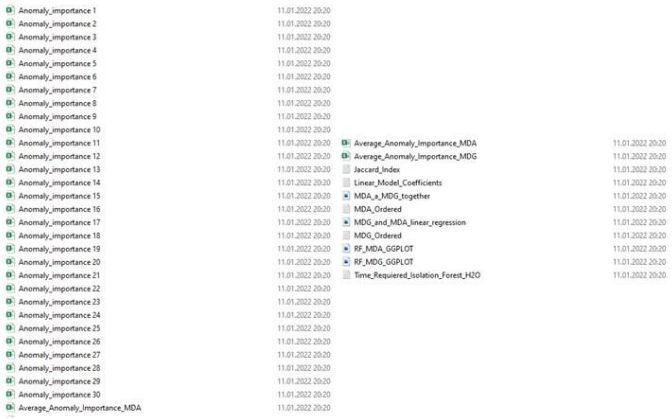
Structure of Exported Results

Two images followed down below show overall structure of exported results in primary working directory and ,,Importance_Plots directory.

Structure of the primary working directory:



Structure of ,,Importance_Plots directory:



References

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.

Šandera, J., & Štych, P. (2020). Selecting Relevant Biological Variables Derived from Sentinel-2 Data for Mapping Changes from Grassland to Arable Land Using Random Forest Classifier. *Land*, *9*(11), 420.

Jaccard, P. (February 1912). "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1". *New Phytologist*. **11** (2): 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x. ISSN 0028-646X.

Contacts:

Contact: sanderajiri@outlook.com

Project link: github.com/jsandera/Isolation-Forest--Time-Series-Anomaly-Detection-in-Remote-Sensing-Data