Due Date: 31/01/2019, midnight
Tutorial date: Wednesday 23/01/2019 (14:00-15:00, Room: IT102)
Email: p.yadav1@nuigalway.ie, piyush.yadav@insight-centre.org

Section 1: Satellite Image Classification

Marks: 30

Assignment:
The objective of this project assignment is to classify two given Landsat Images of different years using QGIS software and identify one major change between two years.

Input Dataset:
Image1: Landsat 7 Image of the year 2003 of the region A.
Image 2: Landsat 8 Image of the year 2018 of the same region(A).

Download Data from below link:
https://nuigalwayie-my.sharepoint.com/:f:/g/personal/p_yadav1_nuigalway_ie/ErlLb3pOOW1OooY2V38AzykBYHirCR3jS4qBqC5YhTjz2Q?e=On45eK

Note:

- Only .tif images are required. There is one _MTL file which contains metadata information which will be required for pre-processing to convert image pixels to reflectance.
- As we see that the Images are from different satellites (Landsat 7 and Landsat 8) as Landsat 8 was launched only in 2013 so we don’t have Landsat 8 images before 2013. Landsat 7 images of 2018 are still available, but due to Scan Line Correction (SLC) issue(discussed in lecture), it's not provided.
- You can also download same data from https://earthexplorer.usgs.gov/ by giving region ‘Dublin’ and date acquired can be seen from metadata file.
- For Image 1(landsat7) use band 1,2,3,4,5 images and for Image 2 (Landsat 8) use band 2,3,4,5,6 images. You can also play with different bands how they look. For viewing them you can use multispec software from Purdue University https://engineering.purdue.edu/~biehl/MultiSpec/

Things to follow for classification:

1: Identify the region. You can easily do that as the Image is geotagged.
2: Classify the image into two classes –

- **Built up class-** This class contains all the concrete area as roads, buildings, houses, pavements etc.
- **Non-Built up class-** other than built up class any other thing like soil, vegetation, trees etc.

Steps to classify Image:

1. Download QGIS Standalone Installer Version from https://qgis.org/en/site/forusers/download.html

2. Install Semi-Automatic classification plugin for QGIS: https://fromgistors.blogspot.com/p/plugin-installation2.html

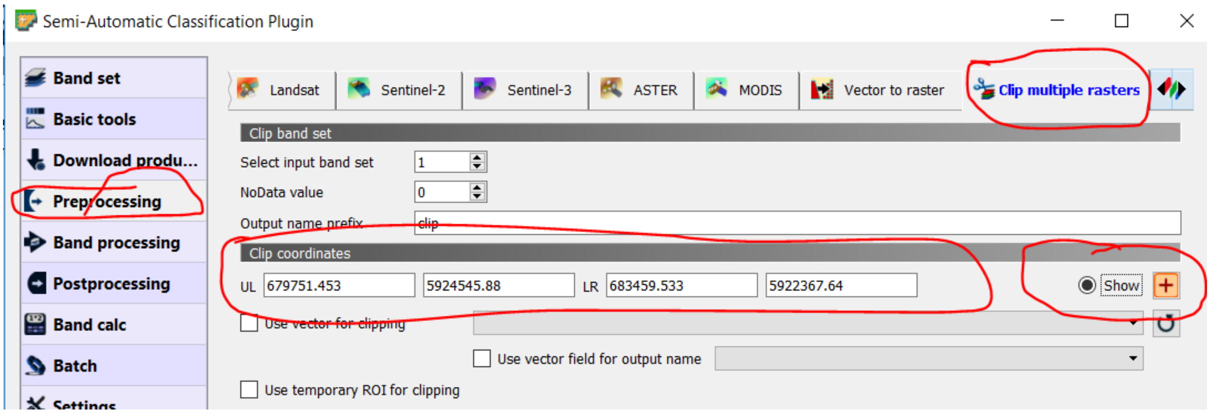
3. Follow the tutorial to classify the given two images **https://fromgistors.blogspot.com/2018/02/basic-tutorial-1.html#more**

   Note: You can directly start from Step 2: Clip the data as the downloaded images have already been provided.

   Data Coordinates for clipping the region are:

   679751.453
   5924545.88
   683459.533
   5922367.64

   As shown in below screenshot for Image 2:



Clip band set

Select input band set 1

NoData value 0

Output name prefix clip

Clip coordinates

UL 679751.453 5924545.88 LR 683459.533 5922367.64 ● Show

□ Use vector for clipping

□ Use vector field for output name

□ Use temporary ROI for clipping

Figure 1

You need to zoom in to see the selected region

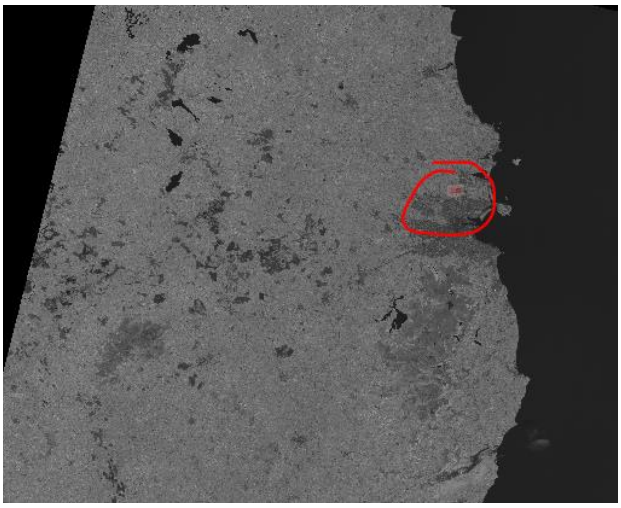


Figure 2

In Step 4 of the tutorial , in Quick wavelength settings select Landsat 8 Oli for Image 2 and Landsat 7 ETM for Image 1.

4. The tutorial used four classes (Water, built up, soil and vegetation), but in assignment, we require only two classes built up and not built up area.

5. There is no requirement to show any spectral calculation in your report as shown in the tutorial.

6. You are free to choose whatever classification algorithm you want. More the number of training region better the classification will be.

7. Classify both the images and point out one major change in built-up area in the classified image. This means there was no built up in 2003 image, but in 2018 image there is new built at that place.

For example:

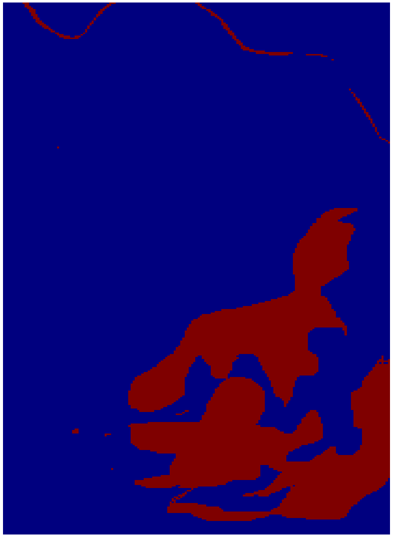


Figure 3 Image 2003

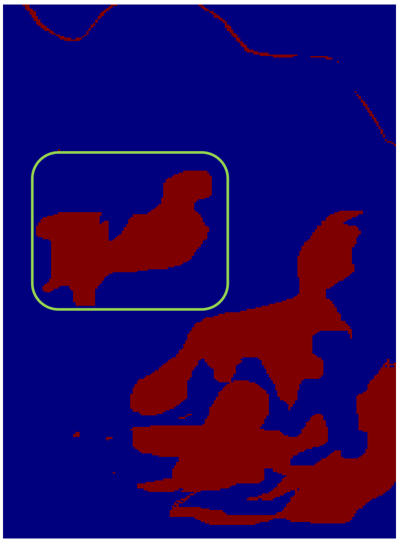
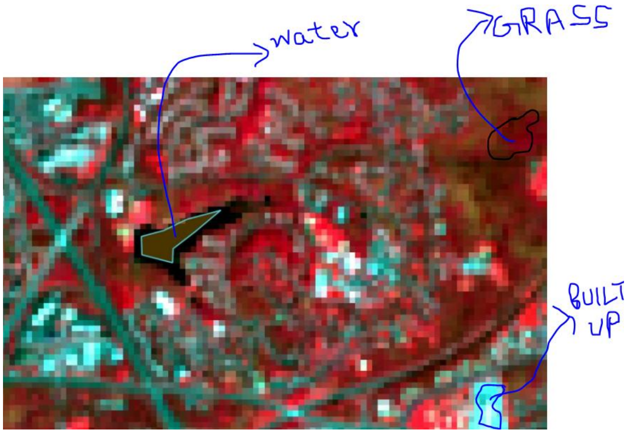


Figure 4 Image 2018

In the above images red means built-up area and blue means non-built up area. In Fig4 there is an extra built up (green rectangle). Similarly you need to show the two images with **ONE major change.**

Deliverables:

1. **Report which includes:**
   a. Name of the region.(Like Its Image from Galway NUIG, Quadrangle)
   b. Image of region of interest selected(Step 5 of tutorial). This will let me know what training samples you have chosen for your classification. An example is shown below:



   c. Both classified Images (as shown in figure 3 and 4). Make sure legends of each class is there in the image.
   Note: In case if you are unable to export the image from QGIS take a screenshot of the classified image and put in your report with legend information.
   d. Marking of one new major built-up area ((as shown in figure 4)).

The marks will depend on how good is your classification image. During classification, you will find some areas which are not classified. This will happen because you have not collected enough training data. It’s ok to have some unclassified area as presently we are dependent on software and have not written our code. Try to collect more training regions for better classification. If your classified image has too many unclassified regions, then marks will be deducted.

Section 2: Multi-Source Open Data Quality

Marks: 70

Assignment:
The objective of the project is to create a complete and clean dataset of playing pitches around Dublin region. The resulting dataset should be a csv file containing as much informative columns as possible.

Input datasets:
Dataset are available on https://data.gov.ie

- [DCC] Playing pitches in Dublin City Council: https://data.gov.ie/dataset/dublin-city-council-parks-playing-pitches
- [DLR] Playing pitches in DLR: https://data.gov.ie/dataset/dlr_pitches
- [F] Playing Pitches in Fingal: https://data.gov.ie/dataset/playing-pitches

Data format:
Each of these datasets exhibits resources using various format: csv, xml, kml, html, etc. You should use at least 2 different formats (you will need to manage 3 files with at least 2 different formats).

Marking scheme:

- Report: 50% (35 marks)
- Code/Analytics: 50% (35 marks)

Deliverables required:

- Source code of your solution should be either in Python or Java with detailed comments to facilitate its readability (Using Jupyter notebook is highly recommended).
- A report outlining your decisions for the different challenges of this project
- The resulting dataset
- Optionally you can share your code on github.

Methodology:
You will have to create the resulting dataset by merging three different resources by resolving the following the three-step pipeline introduced in the lecture:

- **Step1: Observation.** Perform an observatory analysis of the datasets and define a data management plan for creating the resulting dataset.
- **Step 2:** **Data modelling**: Create a unified model that should include the resulting fields of each record in your dataset. The minimal fields required are: location, and geographical coordinates (use x and y).
- **Step 3: Data Quality Enhancement:** You will be faced with two main challenges:
  - **Data cleaning challenge**: Merge the input resource and make sure that there are no duplicates, Locations can be created by merging various fields, etc.
  - **Incomplete data challenge:** The first dataset [DCC] does not have any geographical coordinates. A solution to complete this information is required.

Data observation and modelling: (20 points)
The selected datasets do not share the same data model. Attribute names are different, some attributes or values are missing, etc.

You have to define a data model that describes best the resulting dataset without losing any information. That is, if one of dataset has a field that describes the type of playing pitch, it should be part of the resulting dataset.

A key requirement at this step is to make sure that at least three fields are considered: location, x and y (x and y are the geographical coordinates of the playing pitch).

In your report, indicate what issues where encountered: missing field, different field names, missing values, etc. and indicate what strategy can you use to resolve the issue.

Data Quality Enhancement:

- **Data cleaning challenge: (25 points)**

While merging the datasets, you will notice data issues that you need to resolve. For example, in the second dataset [DLR], you will find that the field "location" is empty, it should take the values of the previous line.

Analyse the datasets, identify data issues and propose a solution for each of them.

- **Incomplete data challenge: (25 points)**

The first dataset [DCC] does not contain any geographical coordinates. To complete the dataset, you will need to complete the dataset by the solution that you think is most suitable (google geocoding service, geocoder python package index, etc.).

You are free to use the method that you like with convincing reasons included in your report.

A suggested method is the following: Use another dataset that helps identify the geographical coordinates as follows:

- [OSiNPG] Dataset: https://data.gov.ie/dataset/townlands-osi-national-placenames-gazetteer2fe62
- Find out the best fields from the first dataset [DCC] that help you find a matching entry in [OSiNPG]. Fields can be the name of the park or the club.

Please note that there is not one correct answer, but there are many **convincing** answers. It is critical to justify the approaches used in the report.

Submission Instructions

- Please put your code into a single .zip archive with name "YourName_CaseStudyAssignment1_code.zip", submit via Blackboard
- Include a screenshot of the output of your application for any relevant output.
- Include all source code files (that is, files with name ending .java etc) required to compile and run your code.
- Use comments to explain your source code. Insufficient comments can lead to mark deductions.
- Please put your report into a single .pdf file with name "YourName_CaseStudyAssignment1_report.pdf", submit via Blackboard
- Please note that all submissions (both code and report) will be checked for plagiarism.

Note: Google Colab is not allowed. It’s a good tool, but a person can change his code even after submission, so submit python/java files. The marks will be penalised if any plagiarism is found between groups.