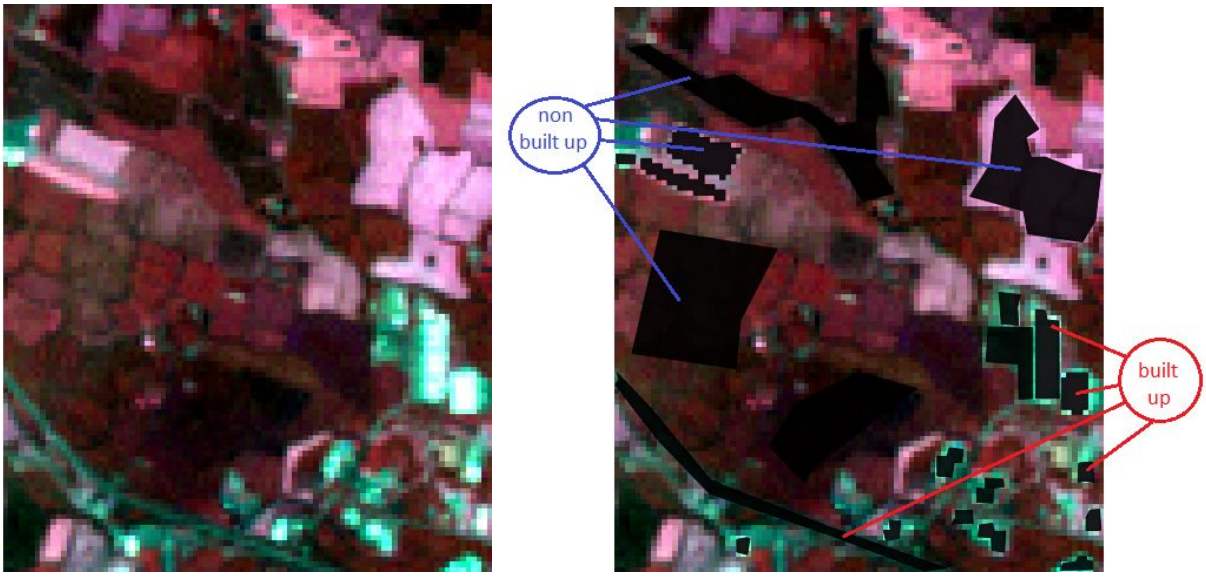
Case study 1
CT5103

Mark Makris
Ross Quinn

Section 1

Region: Clonee, Meath/Dublin border

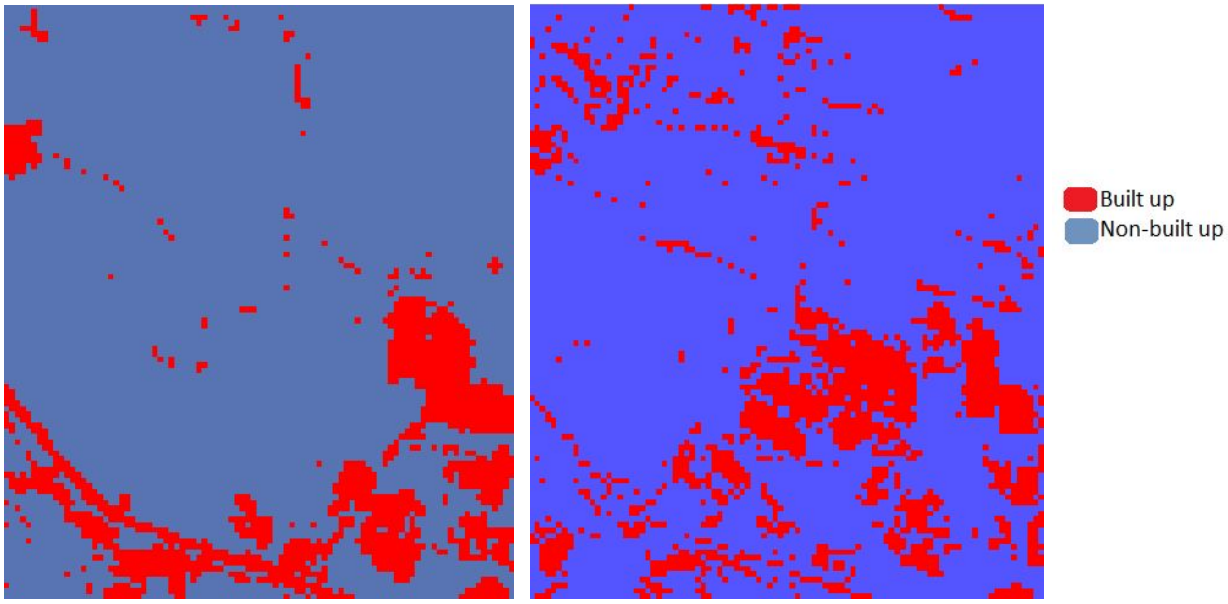
Image of Region: 2009 images



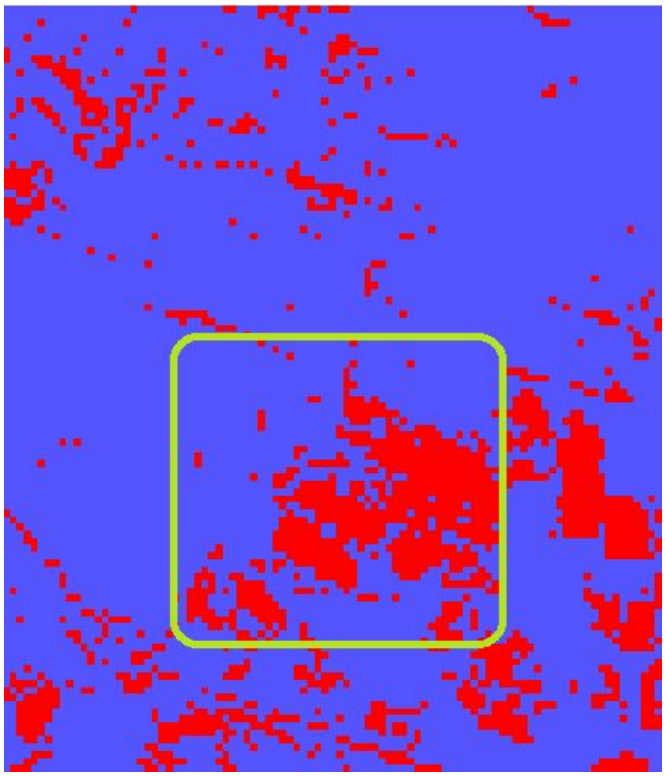
For the non-built up areas, I tried to highlight each distinct colour that indicated that class, using the 4-3-2 RGB setting for easier classification.

Classified images w/ legend:

2009 **2018**



Big change marked:



This change marks the Facebook Datacenter near the IBM building on the right of the image in Clonee.

Section 2

Data observation and modelling:

For the data file selection we chose the csv files from the DCC and DLR (both csv files from DCC were identical) and the DB_TABLE from the third website. The DCC csv had lots of information on the names and leagues of each field while the DLR csv had the longitude and latitude coordinates for pitches with some size information as well. Each table has different missing data and we should use different techniques to change each set as they do not overlap in which pitches are represented. The data table can be transformed into a csv file as well which can then be easily read by python.

The DCC dataset had four columns: park, area, clubs, and leagues, as shown below

PARK	AREA	CLUBNAME	LEAGUE
ALBERT COLLEGE	NORTH WEST	DRUMCONDRA F.C (Snr)	AMATEUR FOOTBALL LEAGUE
ALBERT COLLEGE	NORTH WEST	GLASNEVIN AFC	AMATEUR FOOTBALL LEAGUE
BEECHILL	SOUTH EAST	BALLSBRIDGE FC	AMATEUR FOOTBALL LEAGUE
BELCAMP	NORTH CENTRAL	NEWTOWN CELTIC	AMATEUR FOOTBALL LEAGUE

What it was missing was the geographic coordinates which need to be obtained from other datasets, although thankfully no data was missing within the dataset.
The FCC dataset had columns of id, name, description, and the_geom, as shown below

_id	Name	Description	the_geom
0	Balbriggan Town Park	<table	01010000A0E6100000ACE247B9BABA18C044B425516FCD4A400000000000000000
1	Balheary Reservoir	<table	01010000A0E61000009EA4C4195EE418C0EDD3D7BF81BC4A400000000000000000
2	Town Park	<table	01010000A0E61000007BEB98DFBC7118C0B060AEDBDEC94A400000000000000000
3	St. Mologa's Park	<table	01010000A0E6100000835F63ABE9C118C0627466B80FCF4A400000000000000000
4	Seagrange Park	<table	01010000A0E610000079ABC5A9998A18C0C9CD2600C6B24A400000000000000000
5	Grace O'Malley Park	<table	01010000A0E610000089B13B3D9E4718C06EBDA8471FB14A400000000000000000

The id, name, and geom were actually not valuable to us at all. For instance, the id attribute is just for keeping track of the row in the database. The description has lots of html which has all the information we need like name, location, and coordinates so we needed to extract that data.
The DLR dataset had columns location, number, size, latitude, and longitude, as shown below.

Location	Number	Size	Latitude	Longitude
Kilbogget	1	Snr	53.25724	-6.14067
	2	SSG	53.25761	-6.13988
	3	SSG	53.25784	-6.13927
	4	SSG	53.2571	-6.13909

What we needed was location, latitude, and longitude which is helpful. The ‘number’ and ‘Size’ attributes from this csv are dealt with in the below sections. Finally, the ‘Location’ attribute was changed to mark each pitch entry instead of only the first pitch of that complex.

In order to maintain all the given information and produce a clean data set we first used all the available columns; location(name in FCC dataset), longitude(x), latitude(y), size, clubs, and leagues.

To best see what data is available and missing we translated the html data into a table like list with all the available information. Putting all the data together into one file in a csv made it even easier to go back and look for what we were missing and compare everything at once without looking through a lot of documents.

Data quality enhancement:

Some of the names and locations in each of the datasets may not have even been pointing to actual pitches, and so these had to be checked manually. Taking the DLR csv as an example, a pitch is said to be located at Mount Albany, though Google maps and streetview give no indication that a pitch exists there at all, nor does it appear on a reference website we were using to check if pitches were open, http://www.pitchcheck.ie/. As such, the pitch was removed from the dataset before creating the combined csv file.

In terms of the quality of the attributes themselves, some of the coordinate data is not perfect as it doesn't point to the centre of the pitch, but rather the sideline or even to a road near the pitch. In the FCC dataset, the coordinates of the sports complex is pasted as the location of each pitch within that complex, meaning multiple pitches are marked as being in the exact same location. However, the amount of research it would take to find out what the exact matching location and coordinates for each pitch centre are would simply be too much to be reasonable for this project. As such, the coordinates of the pitches were checked only for being in the general area of the pitch rather than the exact coordinates. Similarly troubling was the 'Size' attribute in the DLR csv, as no defined standardised values or even a legend appear to exist. A few of the pitches are simply marked as size 11, which we eventually discovered to be 11-a-side soccer pitches, and some not marked at all. We decided that this data was to be included in the combined csv as it does provide possibly important information, and so we decided to try to alter the spelling of a few of the entries to match the majority or, for the many pitches missing this attribute in the combined csv, leave them blank. Another confusion in this attribute is the lack of differentiation between soccer and gaelic grounds. This would be a difficult attribute to classify manually as many of the pitches are unmarked, and so this was left as is.

The FCC dataset contained two basketball courts as shown below

St. Mologa's Park	53.61767	-6.18937		All weather pitches
Seagrange Park	53.39667	-6.13535		Basketball Court
Grace O'Malley Park	53.38377	-6.06994		Basketball Court
Broomfiled	53.43916	-6.14467		GAA Pitches

We decided to remove these two entries from the csv file as they weren't pitches and so didn't fit within the project. Similarly, a tennis court complex in Seagrange park was also removed.

Finally, many of the names are written slightly differently, such as st annes, st annes, and st. annes. We had to find ways to make all of those known as the same location. To do this we would search to see if 'st' and 'annes' was in the name so the extra spaces and extra characters would not matter.

```
    if "ST" in row[0] and "ANNES" in row[0]: #accounts for st annes, st. annes, and st annes spelling mistakes
```

With this done, the 'Park' attribute from the FCC file could then be combined into the 'Location' column along with the other two csv files.

Incomplete data challenge:

There is a lot of missing information in all of the files used. The html was missing some of the location names like the dlr as well and the dcc file had no coordinates. There are a few ways to fill in the missing data. We were given the geocoder to find locations and coordinates for missing data. Given the advice of making sure that the places were actually pitches we decided to look at them on geocoder to find locations or coordinates and see if they were correct. Again taking the DLR file as an example, there was missing longitude and latitude data for a temporary pitch located at Shanganagh castle and a pitch that's under construction in Hudson Road. Using Google maps to scout around the location of the castle, the pitch is easily identified though there is no evidence as to whether it is still in use. However, the site http://www.pitchcheck.ie/ states that the pitch is open and so the coordinates were manually filled in. The same was done for the pitch under construction after we used this site to check if it has been opened and found that it had been. If we used a program to fill in the missing data instead of checking maps manually, we would not be sure if it was even accurate. It would be like throwing darts in the dark hoping that the correct answer was found.

We would have tried to use the geocoder api but it costs money to use now and is not worth the price for this project which is why we have to use other options. We

tried using it without paying but it just returns a list of NONE. We also looked at the OSiNPG data sets but in order to find a very precise location of x and y coordinates it did not have the features required. It did not have enough detail to find very specific locations within the counties.

Location	Number	Size
Kilbogget	1	Snr
	2	SSG
	3	SSG
	4	SSG
	5	SGG
	6	Snr
	7	Snr
	8	Snr
	9	Snr
	10	SSG
	11	11
	A	SNR
	B	JNR
	All weath	Mini
	All weath	Full

Above: The 'number' column that was removed

We did not want to get rid of any information that might be somewhat useful like size and type but we could not reasonably fill in the rest of the missing data for the 'Size' column for example so we left it empty where it was not already given. One column that was removed from the final csv was the 'Number' column from the DLR csv file. The 'Number' attribute appeared to lack any standardised values and varied between the assigned pitch numbers within a sports complex, the type of sport played on the pitch, and notes on the usability of the pitch itself. Coupled with the large amount of blank space it would present in the combined csv, the fact that this attribute wasn't consistent in its information means it would either have to be filled in manually with defined values or removed from the combined csv file altogether, which we decided was the better option. If we had more time it would have been great to look into more of this missing data to try and fill it in but we are limited in this project.