

CASA0005 Geographic Information Systems and Science

Andy MacLachlan and Adam Dennett

2019-06-19

Contents

Chapter 1

Welcome

Welcome to the CASA0005 Geographic Information Systems and Science online practical handbook. This website is hosted on Github and holds all the practical instructions and data. All data used within the practicals is available online, however occasionally websites can undergo maintenance or be inaccessible due to political factors such as the government shutdowns.

If you need the practical data you can access it from my github repository here: <https://github.com/andrewmaclachlan/CASA0005>

Practical data is divided into the relevant sessions, although sometimes I'll refer to a dataset used within a previous week.

1.1 The world of GIS

Spatial analysis can yield fascinating insights into geographical relationships. However, at times it can be difficult to work with. You will get lots of error messages and have software crash. The academic staff are here to help you work through these practicals but we do not know everything. It's a good idea to become familiar with online sources of help, such as:

- Stack Exchange <https://stackoverflow.com/>
- QGIS documentation <https://docs.qgis.org/3.4/en/docs/index.html>
- R documentation <https://www.rdocumentation.org/>
- ArcGIS help pages <https://support.esri.com/en>

1.2 Getting Started

One of the issues with GIS is that many of the files we will be working with are quite large. Fortunately, in recent years, UCL has seriously beefed up the storage available for students. You now get 100GB of free storage, which should be plenty for the work you will be doing this year! The Bartlett faculty has several gigabytes of storage space available on their central servers, so before we get started, we will connect to our N drive to carry out all of our practical work over the coming weeks.

1.3 Offline viewing

If you are unable to access the internet to view this github website most web browsers allow you to save webpages for offline viewing.

Instructions for Google Chrome are provided here: <https://support.google.com/chrome/answer/7343019?co=GENIE.Platform%3DDesktop&hl=en&oco=1>

Chapter 2

Practical 1 – Geographic Information

2.1 Learning outcomes

By the end of this practical you should be able to:

- Describe and explain GIS data formats and databases
- Source and pre-process spatial data
- Load and undertake some basic manipulation of spatial data in: ArcMap, QGIS and R
- Evaluate the (dis)advantages of each GIS you have used

2.2 The Basics of Geographic Information

Geographic data, geospatial data or geographic information is data that identifies the location of features on Earth. There are two main types of data which are used in GIS applications to represent the real world. **Vectors** that are composed of points, lines and polygons and **rasters** that are grids of cells with individual values.

In the above example the features in the real world (e.g. lake, forest, marsh and grassland) have been represented by points, lines and polygons (vector) or discrete grid cells (raster) of a certain size (e.g. 1 x 1m) specifying land cover type.

2.2.1 Important GIS data formats

There are a number of commonly used geographic data formats that store vector and raster data that you will come across during this course and it's important to understand what they are, how they represent data and how you can use them.

2.2.1.1 Shapefiles

Perhaps the most commonly used GIS data format is the shapefile. Shapefiles were developed by ESRI (<http://www.esri.com/>) – one of the first and now certainly the largest commercial GIS company in the world. Despite being developed by a commercial company, they are mostly an open format and can be used (read and written) by a host of GIS Software applications.

A shapefile is actually a collection of files – at least three of which are needed for the shapefile to be displayed by GIS software. They are:

1. **.shp** - the file which contains the feature geometry
2. **.shx** - an index file which stores the position of the feature IDs in the **.shp** file
3. **.dbf** - the file that stores all of the attribute information associated with the coordinates – this might be the name of the shape or some other information associated with the feature
4. **.prj** - the file which contains all of the coordinate system information (the location of the shape on Earth's surface). Data can be displayed without a projection, but the **.prj** file allows software to display the data correctly where data with different projections might be being used

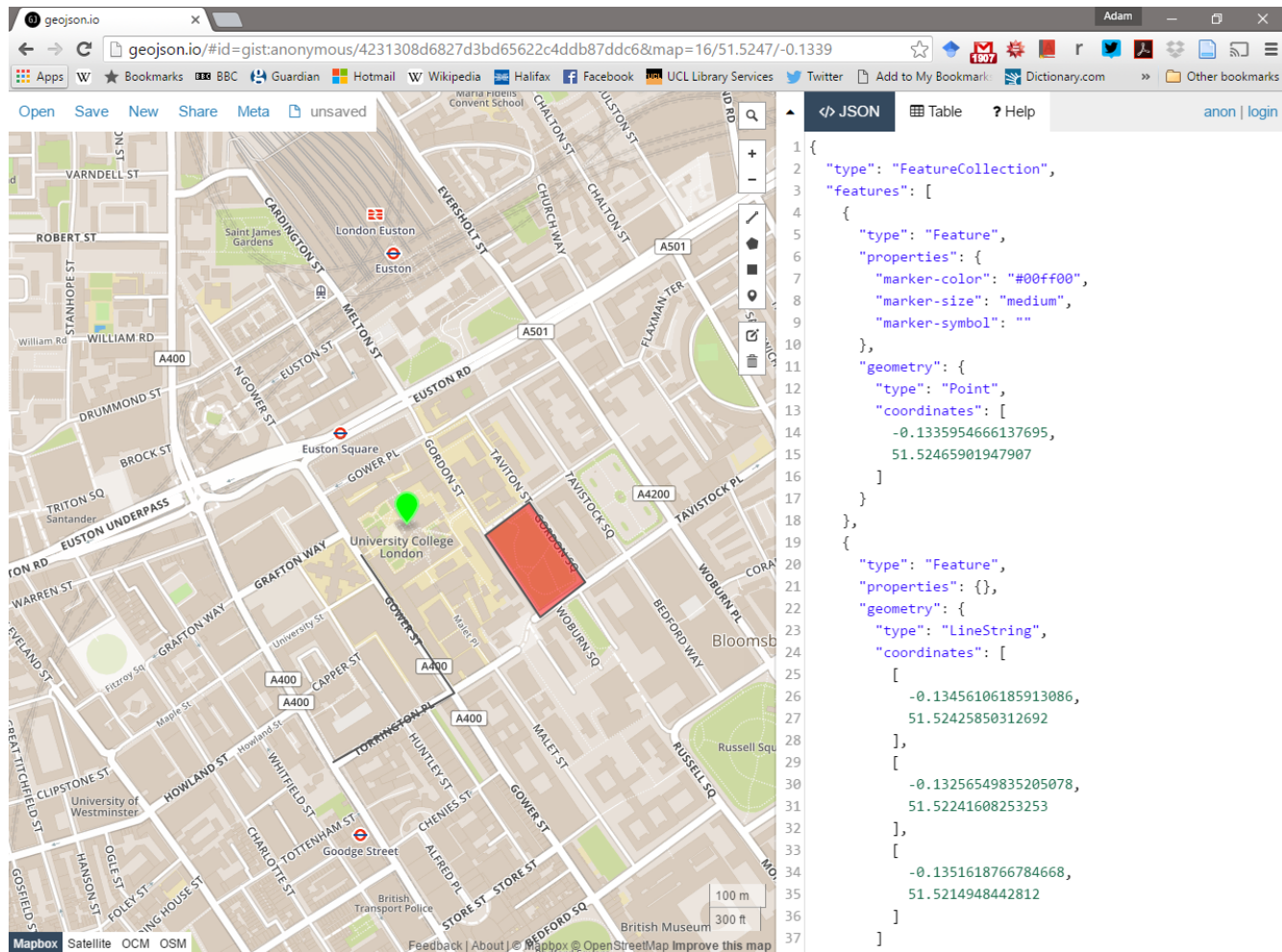
2.2.1.2 GeoJSON

GeoJSON (Geospatial Data Interchange format for JavaScript Object Notation, <http://geojson.org/>) is becoming an increasingly popular spatial data format, particularly for web-based mapping as it is based on JavaScript Object Notation. Unlike a shapefile in a GeoJSON, the attributes, boundaries and projection information are all contained in the same file.

2.2.1.3 Shapefile and GeoJSON

We're now going to explore a shapefile (**.shp**) and GeoJSON (**.geojson**) in action.

Go to: <http://geojson.io/#map=16/51.5247/-0.1339>



1. Using the drawing tools to the right of the map window, create 3 objects: a point, line and a polygon as I have done above. Click on your polygon and colour it red and colour your point green
2. Using the 'Save' option at the top of the map, save two copies of your new data – one in .geojson format and one in .shp format
3. Open your two newly saved files in a text editor such as notepad or notepad++. For the shapefile you might have to unzip the folder then open each file individually. What do you notice about the similarities or differences between the two ways that the data are encoded?

2.2.1.4 Raster data

Most raster data is now provided in GeoTIFF (.tiff) format, which stands for Geostationary Earth Orbit Tagged Image File. The GeoTIFF data format was created by NASA and is a standard public domain format. All necessary information to establish the location of the data on Earth's surface is embedded into the image. This includes: map projection, coordinate system, ellipsoid and datum type.

2.2.1.5 Other data formats

Aforementioned data types and formats are likely to be the ones you predominately encounter. However there are several more used within spatial analysis. There's include:

Vector

- GML (Geography Markup Language – gave birth to KML) - <http://www.opengeospatial.org/standards/gml>

Raster

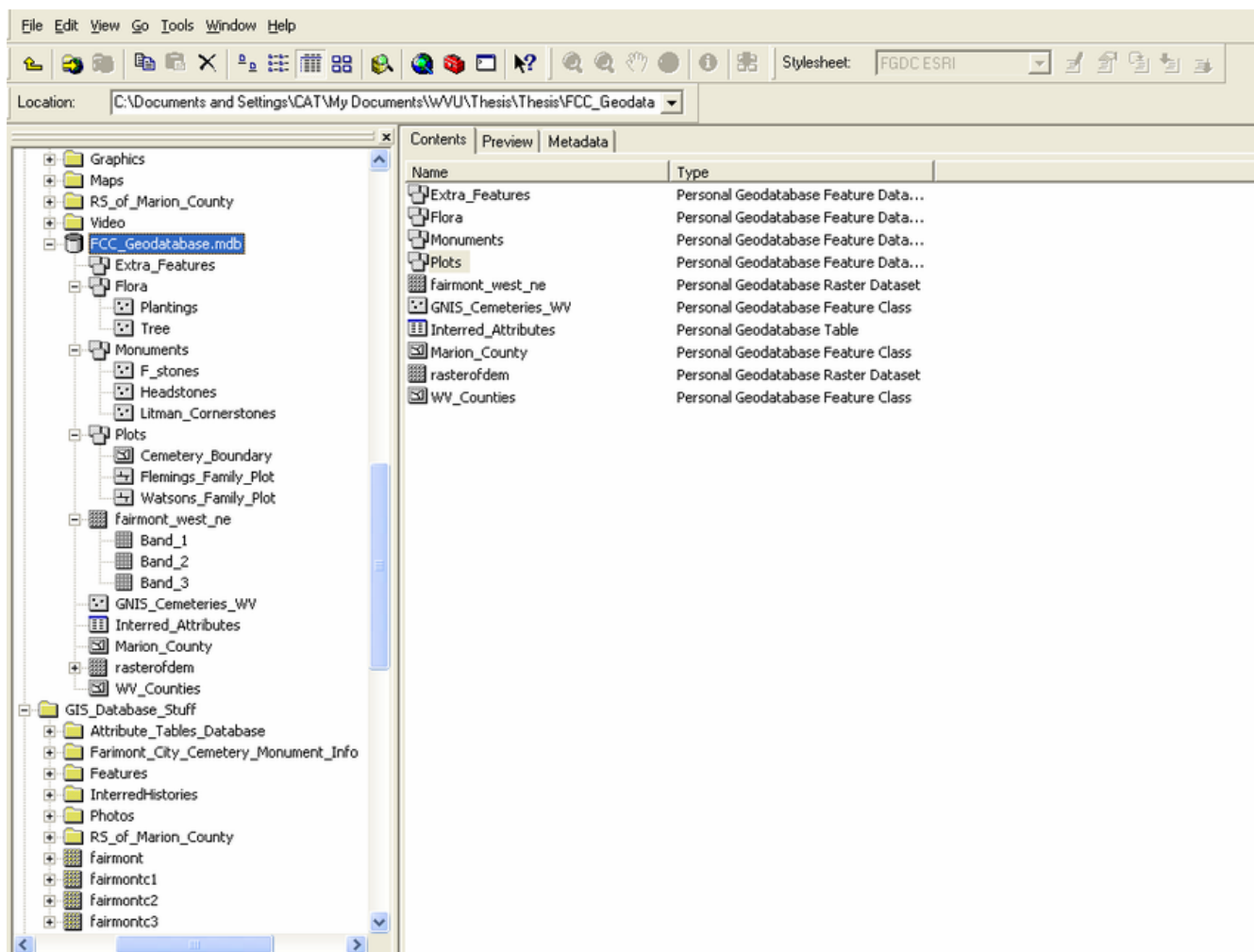
- Band SeQuential (BSQ) - technically a method for encoding data but commonly referred to as BSQ.
- Hierarchical Data Format (HDF)
- Arc Grid

There are normally valid reasons for storing data in one of these other formats. For example, BSQ are raster data with a separate text header file (.hdr) providing geographic spatial reference information. Earth observation data often monitors the electromagnetic spectrum in bands. Humans see in the visible range of the spectrum and our vision is composed of red, green and blue wavelengths. If we wanted to analyse just the red wavelength the BSQ format would let us *read in* only that data. In comparison a GeoTIFF might come with all the data 'packaged' in one file and when doing analysis over thousands of images would significantly slow things down. That said you can now often find GeoTIFFs separated in a similar format to BSQ and it's fairly straightforward to convert between raster formats.

2.2.1.6 Geodatabase

A geodatabase is a collection of geographic data held within a database. Geodatabases were developed by ESRI to overcome some of the limitations of shapefiles. They come in two main types: Personal (upto 1 TB) and File (limited to

250 - 500 MB), with Personal Geodatabases storing everything in a Microsoft Access database (.mdb) file and File Geodatabases offering more flexibility, storing everything as a series of folders in a file system. In the example below we can see that the FCC_Geodatabase (left hand pane) holds multiple points, lines, polygons, tables and raster layers in the contents tab.

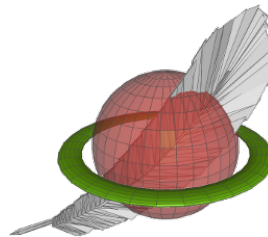


2.2.1.7 GeoPackage



A GeoPackage is an open, standards-based, platform-independent, portable, self-describing, compact format for transferring geospatial data. It stores spatial data layers (vector and raster) as a single file, and is based upon an SQLite database, a widely used relational database management system, permitting code based, reproducible and transparent workflows. As it stores data in a single file it is very easy to share, copy or move.

2.2.1.8 SpatiaLite



SpatiaLite is an open-source library that extends SQLite core. Support is fairly limited and most software that supports SpatiaLite also supports GeoPackage, as they both build upon SQLite. It doesn't have any clear advantage over GeoPackage, however it is unable to support raster data.

2.2.1.9 PostGIS



PostGIS is an opensource database extender for PostrgeSQL. Essentially PostgreSQL is a database and PostGIS is an add on which permits spatial functions. The advantages of using PostGIS over a GeoPackage are that it allows users to

access the data at the same time, can handle large data more efficiently and reduces processing time. In this example (<https://medium.com/@GispoLearning/learn-spatial-sql-and-master-geopackage-with-qgis-3-16b1e17f0291>) calculating the number of bars per neighbourhood in Leon, Mexico the processing time reduced from 1.443 seconds (SQLite) to 0.08 seconds in PostGIS. However, data stored in PostGIS is much harder to share, move or copy.

2.2.1.10 What will I use

The variety of data formats can seem a bit overwhelming. But don't worry, most of the time you'll be using shapefiles, GeoPackages or raster data.

2.3 Data

The volume of geographic information which is freely available for use in the UK is increasing exponentially and spatially referenced data can often be found in many different places. In this practical we're going to use data from the London data store — a free and open data-sharing portal provided by the Greater London Authority (GLA), also known as City Hall that is the devolved regional governance body of London.

We are going to get spatial data of the London boroughs and join flytipping (the illegal deposit of waste, commonly on road verges) data that is provided as a `.csv` file. `.csv` stands for comma-separated values (CSV) — it uses a comma to separate each value.

At the end of this document I'll also run through some common sources of data that will stand you in good stead (be advantageous) for the rest of the course.

2.3.1 File paths

In your N drive: create a new folder called GIS and within this a subfolder called wk1. It is up to you how you organise your files. Make sure you change the file paths within where appropriate to your own.

2.3.2 Data download

Firstly we need to get a spatial outline of the London boroughs. The geographic boundaries that are used in the UK are a complex, often inter-related, but ever changing mass of areas. For anyone new to the UK (or indeed not a trained quantitative geographer), it can be quite a daunting task to attempt to understand all of the boundaries that are in use. Fortunately the Office

for National Statistics (ONS) has an online beginners guide to UK geography. If you need more information on the vast array of different UK geographies, this is the place to start: <http://geoportal.statistics.gov.uk/datasets/a-beginners-guide-to-uk-geography-2018-v1-0>

- Spatial Data

1. To get the data go to: <https://data.london.gov.uk/>
2. Search for Statistical GIS Boundary Files for London
3. Download the statistical-gis-boundaries-london.zip
4. Unzip the data and save it to your wk1 folder.

- CSV data

1. On the same website search for fly-tipping incidents
2. Download the .csv file

2.3.3 Data pre-processing

Question Open the .csv in Excel, what do you notice about how the data is stored?

Answer The year is a column and for each area the values are repeated for different years. In our analysis it is easier to have the different years as a column and populated for each area. So, we want to go from this...

| | A | B | C | D | E | F | G | H | I | |
|----|----------|------------------------|---------|--------|--------|-------|-------|-------|---|-----|
| 7 | E0900000 | Bromley | 2011-12 | 2,222 | 306 | 16 | 9 | 16 | | 2 |
| 8 | E0900000 | Camden | 2011-12 | 6,679 | 5,541 | 465 | 31 | 55 | | 1 |
| 9 | E0900000 | Croydon | 2011-12 | 2,740 | 31 | 17 | 6 | 4 | | 0 |
| 10 | E0900000 | Ealing | 2011-12 | 4,964 | 6,727 | 3,852 | 82 | 193 | | 4 |
| 11 | E0900001 | Enfield | 2011-12 | 19,486 | 7,262 | 191 | 4,483 | 478 | | 5 |
| 12 | E0900001 | Greenwich | 2011-12 | 5,120 | 6,344 | 19 | 974 | 36 | | 1 |
| 13 | E0900001 | Hackney | 2011-12 | 1,030 | 2,332 | 22 | 18 | 171 | | 0 |
| 14 | E0900001 | Hammersmith and Fulham | 2011-12 | 5,700 | 6,944 | 0 | 389 | 3,300 | | 2 |
| 15 | E0900001 | Haringey | 2011-12 | 15,713 | 8,817 | 1,455 | 280 | 796 | | 18 |
| 16 | E0900001 | Harrow | 2011-12 | 8,037 | 1,053 | 177 | 0 | 6 | | 6 |
| 17 | E0900001 | Havering | 2011-12 | 2,972 | 13,970 | 7,354 | 657 | 1,264 | | 0 |
| 18 | E0900001 | Hillingdon | 2011-12 | 4,398 | 5,120 | 60 | 2,208 | 144 | | 0 |
| 19 | E0900001 | Hounslow | 2011-12 | 17,059 | 1,234 | 8 | 0 | 0 | | 0 |
| 20 | E0900001 | Islington | 2011-12 | 2,048 | 6,959 | 1,119 | 3,595 | 610 | | 0 |
| 21 | E0900002 | Kensington and Chelsea | 2011-12 | 5,482 | 4,773 | 0 | 120 | 0 | | 0 |
| 22 | E0900002 | Kingston Upon Thames | 2011-12 | 239 | 0 | 0 | 0 | 0 | | 0 |
| 23 | E0900002 | Lambeth | 2011-12 | 1,696 | 598 | 0 | 212 | 31 | | 0 |
| 24 | E0900002 | Lewisham | 2011-12 | 15,757 | 789 | 164 | 101 | 22 | | 15 |
| 25 | E0900002 | Merton | 2011-12 | 2,526 | 1,279 | 171 | 255 | 42 | | 2 |
| 26 | E0900002 | Newham | 2011-12 | 40,499 | 9,593 | 1 | 8,876 | 234 | | 0 |
| 27 | E0900002 | Redbridge | 2011-12 | 5,110 | 6,114 | 3,172 | 66 | 286 | | 0 |
| 28 | E0900002 | Richmond Upon Thames | 2011-12 | 3,208 | 2,610 | 645 | 16 | 1 | | 1 |
| 29 | E0900002 | Southwark | 2011-12 | 16,823 | 0 | 0 | 0 | 0 | | 0 |
| 30 | E0900002 | Sutton | 2011-12 | 1,479 | 1,255 | 65 | 120 | 112 | | 0 |
| 31 | E0900003 | Tower Hamlets | 2011-12 | 6,184 | 3,251 | 790 | 1,079 | 297 | | 0 |
| 32 | E0900003 | Waltham Forest | 2011-12 | 4,456 | 17,182 | 252 | 991 | 6,031 | | 0 |
| 33 | E0900003 | Wandsworth | 2011-12 | 1,561 | 8,226 | 4,609 | 255 | 0 | | 366 |
| 34 | E0900003 | Westminster | 2011-12 | 14,694 | 8,367 | 1,943 | 5,622 | 0 | | 0 |
| 35 | E0900000 | City of London | 2012-13 | 449 | 1,492 | 658 | 245 | 3 | | 0 |
| 36 | E0900000 | Barking and Dagenham | 2012-13 | 2,417 | 2,122 | 123 | 49 | 56 | | 0 |
| 37 | E0900000 | Barnet | 2012-13 | 1,697 | 1,828 | 264 | 67 | 0 | | 0 |
| 38 | E0900000 | Bexley Council | 2012-13 | 1,038 | 627 | 38 | 47 | 0 | | 0 |
| 39 | E0900000 | Brent | 2012-13 | 6,911 | 2,232 | 182 | 184 | 268 | | 12 |

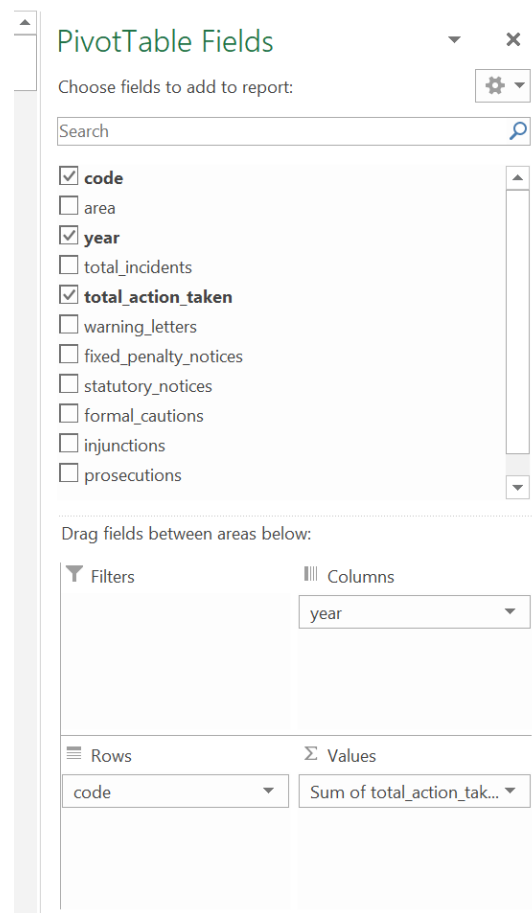
To this...

| | A | B | C | D | E | F | G | H |
|----|------------------------|---------|---------|---------|---------|---------|---------|---------|
| 1 | Sum of total_incidents | | | | | | | |
| 2 | Row Labels | 2011-12 | 2012-13 | 2013-14 | 2014-15 | 2015-16 | 2016-17 | 2017-18 |
| 3 | E09000001 | 527 | 449 | 530 | 369 | 627 | 1731 | 1820 |
| 4 | E09000002 | 3665 | 2417 | 1282 | 2564 | 2361 | 2423 | 2620 |
| 5 | E09000003 | 1450 | 1697 | 1779 | 11615 | 5999 | 7029 | 6150 |
| 6 | E09000004 | 980 | 1038 | 1078 | 1162 | 1110 | 1480 | 2100 |
| 7 | E09000005 | 7272 | 6911 | 7001 | 12912 | 13198 | 17340 | 18600 |
| 8 | E09000006 | 2222 | 2495 | 2809 | 3377 | 3343 | 3246 | 3060 |
| 9 | E09000007 | 6679 | 11477 | 10950 | 8308 | 7268 | 6778 | 12170 |
| 10 | E09000008 | 2740 | 11150 | 15113 | 18560 | 0 | 24797 | 19190 |
| 11 | E09000009 | 4964 | 6352 | 5765 | 7257 | 7032 | 14270 | 13610 |
| 12 | E09000010 | 19486 | 17871 | 31692 | 50121 | 70930 | 75614 | 39000 |
| 13 | E09000011 | 5120 | 5170 | 12765 | 14324 | 10712 | 7960 | 8410 |
| 14 | E09000012 | 1030 | 599 | 7635 | 6008 | 6917 | 3267 | 9820 |
| 15 | E09000013 | 5700 | 7079 | 9011 | 9208 | 10829 | 14870 | 18650 |
| 16 | E09000014 | 15713 | 12398 | 31045 | 25709 | 34975 | 33333 | 23540 |
| 17 | E09000015 | 8037 | 6228 | 8429 | 7072 | 8462 | 6835 | 9620 |
| 18 | E09000016 | 2972 | 2842 | 3620 | 2914 | 3726 | 4061 | 4650 |
| 19 | E09000017 | 4398 | 4506 | 1995 | 1817 | 7267 | 7766 | 7180 |
| 20 | E09000018 | 17059 | 13934 | 15864 | 16282 | 19809 | 22973 | 17060 |
| 21 | E09000019 | 2048 | 2231 | 2634 | 3166 | 4174 | 3011 | 2580 |
| 22 | E09000020 | 5482 | 6075 | 6934 | 6603 | 6105 | 9029 | 9850 |
| 23 | E09000021 | 239 | 126 | 339 | 956 | 1263 | 1528 | 1340 |
| 24 | E09000022 | 1696 | 1454 | 1206 | 4056 | 2493 | 3427 | 3430 |
| 25 | E09000023 | 15757 | 10604 | 9152 | 6747 | 5514 | 1931 | 3560 |
| 26 | E09000024 | 2526 | 3098 | 3064 | 3819 | 2816 | 3113 | 8870 |
| 27 | E09000025 | 40499 | 28443 | 67930 | 70192 | 32718 | 19917 | 15200 |
| 28 | E09000026 | 5110 | 6986 | 8939 | 11066 | 10864 | 12461 | 0 |
| 29 | E09000027 | 3208 | 2608 | 2871 | 2903 | 3723 | 5253 | 4700 |
| 30 | E09000028 | 16823 | 25894 | 26638 | 25583 | 21359 | 17131 | 15740 |
| 31 | E09000029 | 1479 | 1512 | 1264 | 1575 | 1478 | 2296 | 2680 |
| 32 | E09000030 | 6184 | 3732 | 5201 | 6671 | 4555 | 6287 | 7460 |
| 33 | E09000031 | 4456 | 3951 | 4723 | 6127 | 6798 | 6772 | 7670 |

As we are going to use this dataset in ArcMap, QGIS and R I've done it in Excel using a pivot table.

1. Go to Insert > PivotTable
2. Select the original table and create a PivotTable in a new worksheet
3. The PivotTable Field box will appear, experiment with the different fields in each of the areas

I've used the following:



Note how I've altered the `total_action_taken` to the sum of... as the original was displaying incorrectly, to do so:

1. Click on drop down button for total_action_taken > Value Field Settings > select Count

It's important to think about what data we actually need in the next step and it's good practice to avoid data redundancy where possible.

Spoiler The spatial data we have downloaded already contains borough name, so we don't need it twice. However, we do need a field to link the two datasets on. You could use borough name, but when using text fields sometimes input variations can affect joins. For example, you had the University of Manchester in one dataset and Manchester University in another the join would fail. Consequently it's usually best to join datasets on a code field.

Now save the Excel sheet that contains the pivot table as a new .csv. Make sure that the first row of data holds the column titles. Remove all empty rows.

When saving the file also avoid any special characters (e.g. -) and spaces, use an underscore instead of spaces.

Warning Spatial software (especially ArcGIS) does not like file names with spaces or special characters.

2.3.4 Data loading

Now it's time to load, inspect and do some basic manipulation of this data. As mentioned in the lecture there are several GIS software 'types', here we will repeat the same process across ArcGIS, QGIS and R. Each system has specific benefits, but in general there has been a recent shift towards the use of QGIS and R, both being opensource. ArcGIS was the first major spatial analysis software produced by the Environmental Systems Research Institute, Inc. (Esri), founded in 1969 by Jack Dangermond. Due to its high cost and lack of customisation it is now less commonly used within the research community.



2.3.5 ArcGIS

2.3.5.1 Basics

ArcGIS should be installed as a standard programme in the UCL desktop and you can navigate to it from the Windows start button.

2.3.5.1.1 Installing ArcGIS on your own computer

As a UCL student, you can install ArcGIS on your own computer. This is easy if you have a PC, but if you have a Mac this can be trickier as Arc will only run in a PC environment. If you have a Mac, the options open to you are either to:

- a) Run ArcGIS through the Desktop@UCL application - <http://www.ucl.ac.uk/isd/services/computers/remote-access/desktop>
- b) Dual boot your machine using bootcamp, install Windows (7 or 8 is fine) and then install Arc onto the Windows partition.
- c) Install some kind of virtualisation software such as Parallels (<http://www.parallels.com/ca/products/desktop/>) or VMware (<http://www.vmware.com/products/fusion/>), and run Arc on a virtual windows machine

If you can, it is preferable to run Arc on Bootcamp as virtualisation software can be slow, but the Desktop@UCL facility should suffice for this course. ArcGIS (Version 10.6 is the latest at time of writing, but may have already been superseded) can be downloaded from the UCL Software Database for free — <https://swdb.ucl.ac.uk/>.

2.3.5.1.2 Getting Help

ArcGIS is a huge and complex piece of software, but thankfully it has an excellent help system – depending on the version you are using (they are all quite similar anyway) you can access the online help system here:

- <http://resources.arcgis.com/en/help/main/10.2/>
- <http://resources.arcgis.com/en/help/main/10.1/>

2.3.5.1.3 ArcGIS

ArcGIS is actually a whole suite of software built and maintained by ESRI (<http://www.esri.com/software/arcgis>). Within the ArcGIS for Desktop suite you will find the following programmes:

- **ArcCatalog** — Similar to Windows Explorer, ArcCatalog allows you to manage your GIS files, folders and geodatabases