PRACTICAL EXERCISE 3: Joining data in QGIS

Introduction

In this exercise you will learn how to join data via attributes and spatially to gain further insight into road traffic accidents across Wales. You will find data for this exercise in the data folder.

This instruction sheet should be read in conjunction with the document *PMIM502_Practical_guide* which gives general information about QGIS and how to find further help.

Learning aim

In this exercise you will join Stats19 road traffic accident data – an open dataset published by the Department for Transport – to census data using two methods. You will use a new open data format called *GeoPackage* which has some advantages over shapefiles and learn about the field calculator. You will experiment with visualising this data to reveal insight about road traffic accidents across Wales.

Instructions

Startup and data preparation

1) Launch QGIS and create a new project. Load the data from the csv folder *csv->acc.csv* into QGIS using the *create layer from a delimited text* file tool 
2) Select the columns for x and y using the Easting and Northing columns to finish loading the data. You should have a point location for all accidents in the UK where the police attended for 2016 (136,621!).
3) Explore the attribute table. You will notice due to the volume of the data it may take a while to load. Click *Abort* and you will be able to view the records that have been loaded.
4) Explore this dataset by panning and zooming and mapping the Stats19 data using some of the techniques you have used in previous practical sessions.

Performing a Spatial Join

5) Although a rich dataset sometimes there is too much information for the map to be easily understood. We therefore need to *aggregate* the data. Load the LSOA_2011 vector outline from the folder *census-> LSOA_2011.gpkg*. This data represents each census boundary in Wales from the 2011 census.
6) Save your Project - Remember to save regularly from now on.
7) There are 32 fields in the Stats19 accidents dataset. This can be overwhelming. We can quickly cut down the fields to those we are interested in by exporting the CSV layer to a *GeoPackage.* GeoPackage is an open data format which has advantages over Shapefiles including to the ability to save longer field names. Right click the CSV accident layer and choose *Save As*.
8) Make sure the format is *GeoPackage* and choose the only the *Number_of_Casualties* field (hint: there is a *Deselect All* button…).
9) Once the new layer has loaded we can join the Stats19 data with number of casualties to a census boundary using a spatial join. Go to the *Processing* menu and search for the '*Join attributes by location'* tool.
10) Fill out the tool form with the target layer as the LSOA layer, the join layer as the newly created Stats19 GeoPackage layer. You can now choose your spatial join method – in this scenario we will use *intersects*. We want to aggregate (or take a summary) the *Number_of_Casualties* field for **all** points which fall within an LSOA boundary – choose the correct option in the Attribute summary drop down. In the joined table choose the option that will keep all the LSOA target records even if there hasn't been an accident in it. Save your joined layer as a new *GeoPackage* layer in the Joined layer box (you may need to scroll down). Once complete there should be a "*Joined Layer"* layer in the layers panel.

Mapping the Data

11) We can no map this data as to reveal spatial patterns of accidents across Wales. Using *Graduated* styling we can create a *Choropleth map* which shades each area in relation to a data variable. Experiment with the mode option – see how the map changes. You can look up the statistical definition of the mode in a search engine or here http://support.esri.com/en/other-resources/gis-dictionary/ - Jenks' Natural Breaks classification has been developed for Choropleth mapping.

12) Do you notice anything strange about the map? Are all areas shaded as you would expect (switching between a single symbol and a categorized symbol might help)? Explore the attribute table and see if you can spot the problem.
13) The LSOA's where no accidents occurred have been assigned a NULL value – we can correct this using the field calculator. Open the *Joined Layer* attribute table and choose the field calculator 
14) To update the *count* field, check the '*Update existing field*' box and enter the following:

    if("count" is NULL, 0, "count")

    This logic code tells QGIS to replace all values in the *count* column when there is a NULL value with a zero, if a value exists just use that value. Repeat this for the remaining columns (not the *fid*, *LSOA11Code*, or *lsoa11name* columns). Once complete toggle the editing button and click save

15) Remap the data and the "missing" LSOAs should now be included in the map.

Joining tables using common attributes

16) One of the issues often seen with *Choropleth* maps is lack of normalisation – failing to give the correct context to the data results in a map delivering the wrong message. For a detailed explanation see https://www.citylab.com/design/2015/06/when-maps-lie/396761/ . In this case we want to normalise the data as the number of casualties per exposed population to highlight areas where there is a higher risk of injury from a road traffic accident. We can use census data to create a basic rate – in this case Mid Year Estimates for resident population from the Office for National Statistics. Add the csv->Wales_MYE_2016.csv file to QGIS as delimited file. There are no coordinates in this file so choose the *No geometry* option for geometry definition.
17) We are now going to join the census data to the aggregated accident data using the LSOA codes which are common to both datasets. Right click the "*Joined Layer"* and click properties. Navigate to the *Joins* tab and click the + icon.
18) Choose the Join Layer as the CSV file that you have just loaded. The *Join Field* and *Target Field* should be common to both attribute tables – explore both tables to see which fields these are. Once you are confident that you have the correct join fields choose *Total_Pop* as the field that you want to join and click OK. Open the attribute table of he *Joined Layer,* if the join has been successful you should see a *Total_Pop (it may have a prefix)* has been added to your attribute table**. *Note: Joins are temporary you will need to save them as new GeoPackage to make them permanent.***
19) We can now use the new population column to create a basic rate of accidents by population. Open the field calculator again, this time we will create a new field with a *Decimal* field type. Fill in the field name (e.g. *Accidents_per_1000_pop*). Go to the expression box and type in the following:

    round(("count" / "Wales_MYE_2016_Total_Pop") * 1000,2)

    This equation creates a rate of accidents per 1000 population - by dividing the number of accidents in an LSOA by the total population multiplied by 1000, rounded to 2 decimal places. You can now save the new joined and updated LSOA layer as a new file

20) Map the normalised field using the same classification technique you settled on before (you may want to duplicate your layer – *right click layer->duplicate*) and compare with the standard counts layer. Explore how the two maps and classifications techniques vary.

Learning outcomes

By the end of this exercise you should:

1) Be able to join data using a *spatial join* and an *attribute join*
2) Be able to use the field calculating to extract and derive new data and meaning
3) Appreciate the different methods for grouping data statistically
4) Understand the importance of normalising data for *Choropleth* mapping