Geographic Information for Vector Surveillance

Day 3 of a 3 day course with Malaria examples

Getting your own data into QGIS

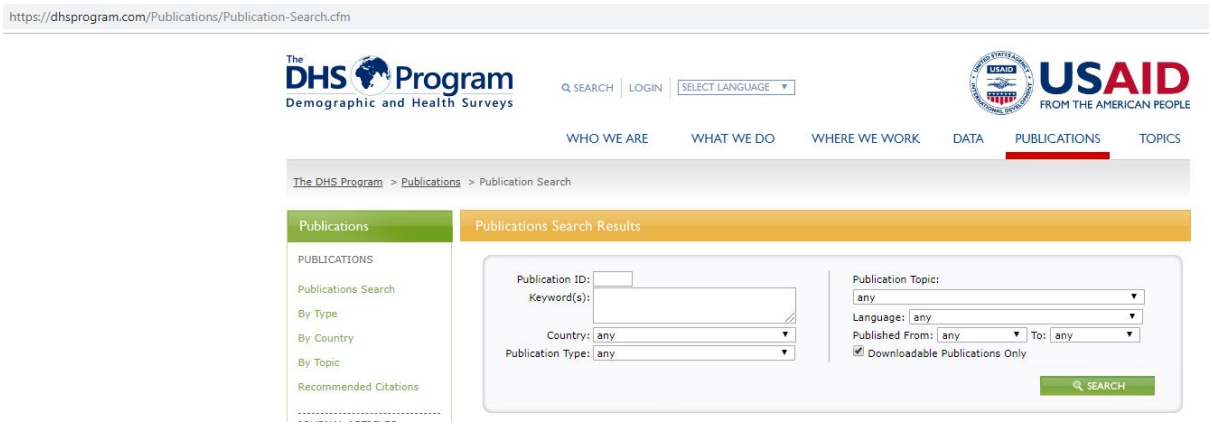
Learning objectives

- be able to join data from an Excel spreadsheet to a shapefile in QGIS
- be able to add point-level data from an Excel spreadsheet into QGIS
- be able to create maps using your own data

Length : 1 day
QGIS version : 3.2
Authors : Michelle Stanton, Sophie Dunkley, Andy South
Updated : October 2018

Joining data from an Excel spreadsheet

Recall that shapefiles, in addition to containing geographical information often contain additional attributes associated with the geography. You can view the extra information contained in the shapefile by looking at the *Attributes* table (right-click on the layer, select Open Attributes Table). This allows us to make maps based on these attributes. It is often the case that the information we want to map is contained in a separate file e.g. an Excel file. In this section we will show you how to join information from an Excel file to a shapefile to allow you to produce a map of these data. We will use province-level data from Zambia to demonstrate this. Specifically, we have obtained information relating to the percentage of households with at least one insecticide-treated net (ITN), and the percentage of under five year olds who had a fever in the two weeks prior to the survey. These data have been extracted from the most recent Demographic and Health survey (DHS) conducted in 2013-14. You can learn more about the DHS on their website https://dhsprogram.com, and search to see what data they have available on your own country via the Publications Search (https://dhsprogram.com/Publications/Publication-Search.cfm).



The joining process involves linking two or more datasets using a unique identification variable that needs to be present in all datasets being joined. For example, suppose I have two spreadsheets as below. One spreadsheet has two columns named country and climate, the other has two columns named nation and favourite sport (this is a very simple example and you may not agree with the data!).

country	climate
UK	cold
Colombia	hot
USA	mixed

nation	favourite sport
Colombia	football
USA	baseball
UK	football

If we join these two spreadsheets using the first column as the unique joining variable, then we obtain a single spreadsheet as seen below.

country	climate	favourite sport
UK	cold	football
Colombia	hot	football
USA	mixed	baseball

There are a number of points to note:

~ the joining columns do not have to have the same name

~ the rows do not have to be in the same order

~ the data columns can contain repeated values (e.g. football in this example)

~ the join columns (country and nation in this case) must not contain repeated values

~ only values that are identical will be joined

We can do the same thing with spatial data. For example, we have administrative boundary data which contains spatial information but no malaria-related information, and we have an Excel file which contains malaria-related information but no spatial information. Both data sources however contain a unique ID variable which allows us to 'join' the spatial information and the malaria-related information into one file. Note, that when joining two or more sources, the joining variable does not necessarily have to have the same column name, but must contain at least some of the same values.

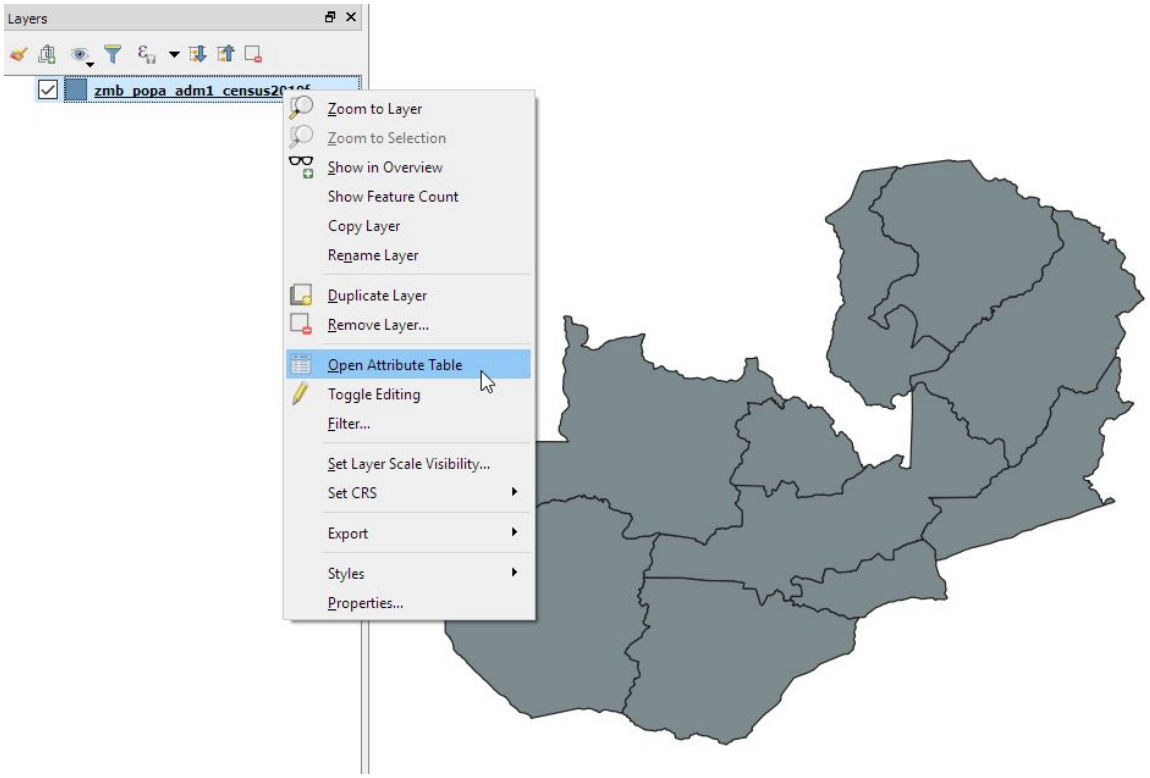
To begin this joining process, we first need to add the province-level polygons for Zambia to QGIS.

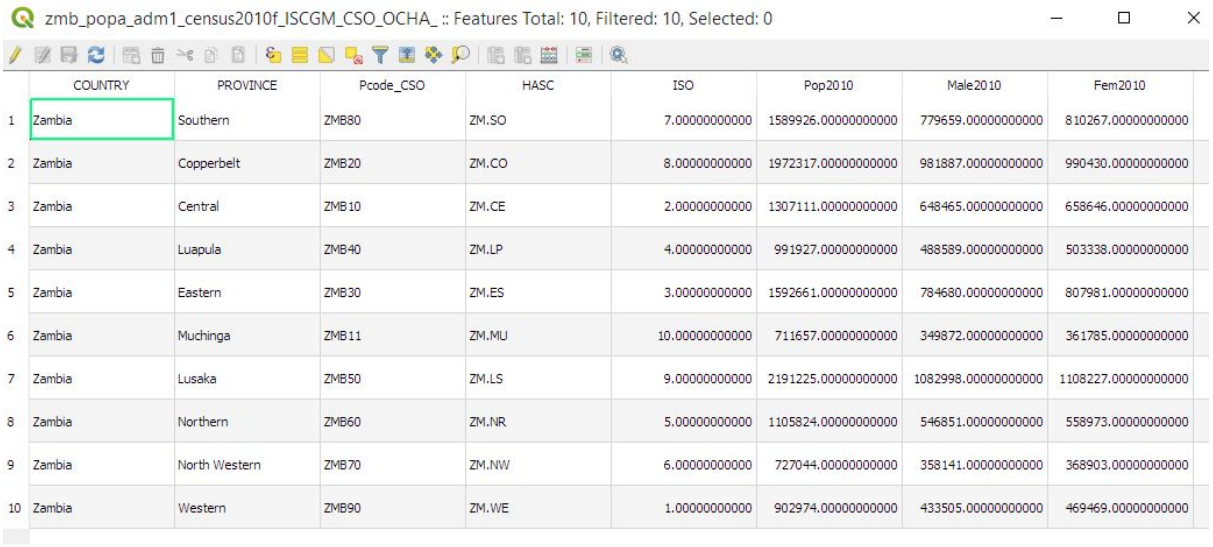
➔ Open QGIS, start a new Project with Project, New

➔ Layer, Add Layer, Add vector layer, browse to find this file from the provided files : **QGIS_training\data\spatial\gadm_ZMB_admin1_MIS.shp**

(Alternatively the administrative level 1 data for Zambia can be obtained from https://data.humdata.org or https://gadm.org/download_country_v3.html).

➔ Open up the attributes table to see what information is contained within the shapefile




zmb_popa_adm1_census2010f_ISCGM_CSO_OCHA_ :: Features Total: 10, Filtered: 10, Selected: 0

	COUNTRY	PROVINCE	Pcode_CSO	HASC	ISO	Pop2010	Male2010	Fem2010
1	Zambia	Southern	ZMB80	ZM.SO	7.00000000000	1589926.00000000000	779659.00000000000	810267.00000000000
2	Zambia	Copperbelt	ZMB20	ZM.CO	8.00000000000	1972317.00000000000	981887.00000000000	990430.00000000000
3	Zambia	Central	ZMB10	ZM.CE	2.00000000000	1307111.00000000000	648465.00000000000	658646.00000000000
4	Zambia	Luapula	ZMB40	ZM.LP	4.00000000000	991927.00000000000	488589.00000000000	503338.00000000000
5	Zambia	Eastern	ZMB30	ZM.ES	3.00000000000	1592661.00000000000	784680.00000000000	807981.00000000000
6	Zambia	Muchinga	ZMB11	ZM.MU	10.00000000000	711657.00000000000	349872.00000000000	361785.00000000000
7	Zambia	Lusaka	ZMB50	ZM.LS	9.00000000000	2191225.00000000000	1082998.00000000000	1108227.00000000000
8	Zambia	Northern	ZMB60	ZM.NR	5.00000000000	1105824.00000000000	546851.00000000000	558973.00000000000
9	Zambia	North Western	ZMB70	ZM.NW	6.00000000000	727044.00000000000	358141.00000000000	368903.00000000000
10	Zambia	Western	ZMB90	ZM.WE	1.00000000000	902974.00000000000	433505.00000000000	469469.00000000000

There are 10 provinces in the shapefile, and attached to each province is some population data taken from the 2010 census. The important column to pay attention to here is the PROVINCE column as this is what we will be using to join this geographical data to the malaria-related data.

You have been provided with two Excel files containing malaria-related data. We will first look at the ITN coverage data in the file DHSZambia_ITNownership.xlsx

	A	B	C
1	Province	ITNpercent	
2	Central	67.3	
3	Copperbel	74.6	
4	Eastern	77.1	
5	Luapula	63.3	
6	Lusaka	48.8	
7	Muchinga	72.9	
8	Northern	60.2	
9	North Wes	66.1	
10	Southern	79.5	
11	Western	76.5	
12			

This spreadsheet contains two columns. The first is the province name (*Province*), and the second is the percentage of surveyed households that owned at least one ITN (*ITNpercent*). In order to be able to join this to the shapefile we need to do the following two things:

- ➔ Check that the province names are written in an identical way to those in the shapefile *Attributes* table
- ➔ Convert the Excel file to a CSV

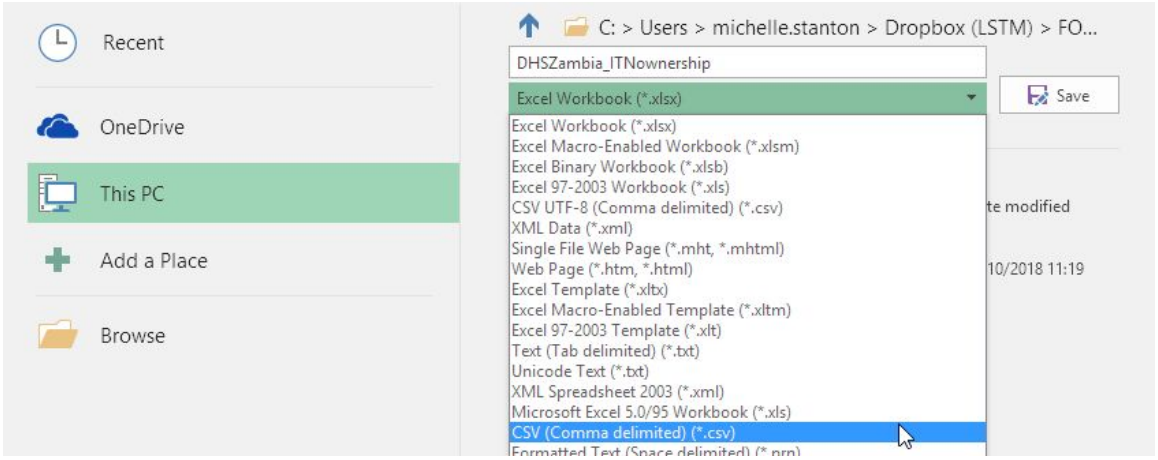
The joining process is dependent on the ID variable being written in **exactly the same way** in both of the files being joined. Text is case sensitive so words must be capitalised in the same way, the

same number of spaces need to be used etc. In this case, the ten provinces seem to be written identically in the two files being joined. If this was not the case, the *Province* column in the Excel file could be edited.

QGIS does not recognise the Excel Workbooks in .xlsx format. Like many other softwares, it uses a format called CSV. Therefore we are going to change our file from .XLSX to .CSV.

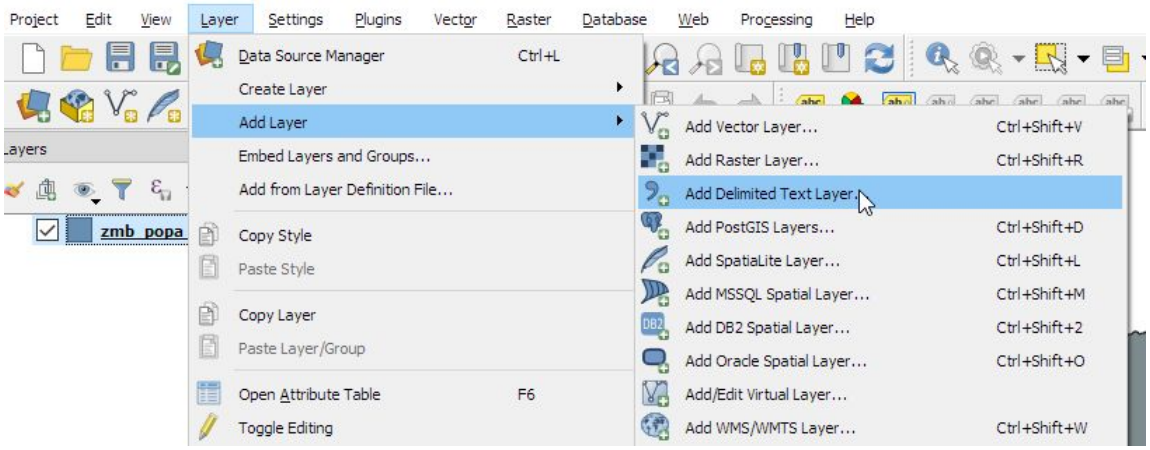
CSV stands for "comma separated values", also known as a comma delimited file. CSV files save the information as lists of the values (text or numbers), separated one from another by commas (or in languages where commas are used to denominate a decimal place, the values can instead be separated by a semi-colon to avoid confusion in interpretation of numbers). CSV files are commonly used for data transfer as they are simple and relatively small.

To convert our data from .xlsx to .csv we will use the "Save As" option in Excel (File/Save As) and simply change the file type in the drop down "save as type" option from "Excel Workbook (*.xlsx)", to "CSV (Comma delimited)(*.csv)".

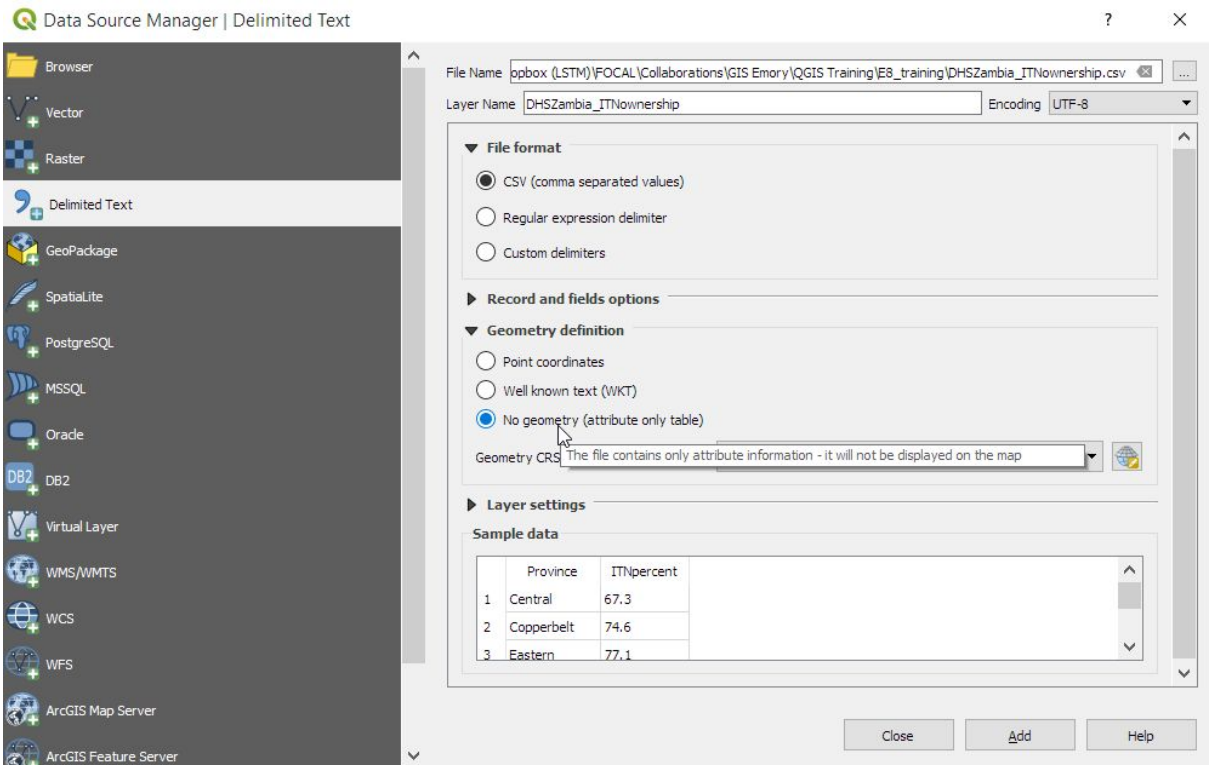


Before we can join the two datasets, we first need to add the CSV into QGIS.

➔ In QGIS select Layer, Add Layer, Add Delimited Text Layer



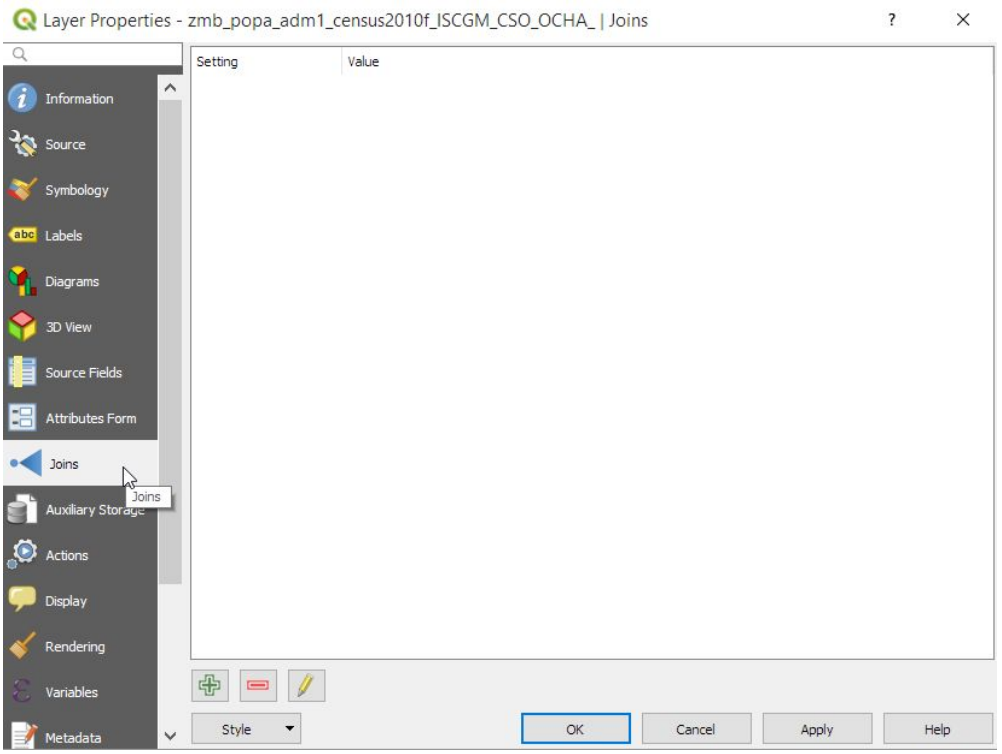
➔ Use the button next to File Name to browse to and select the ITN ownership CSV file.
➔ The default file format is CSV so we do not need to change this.
➔ Expand the Geometry definition box, and select *No geometry (attribute table only).*
➔ Click on Add
➔ Click Close



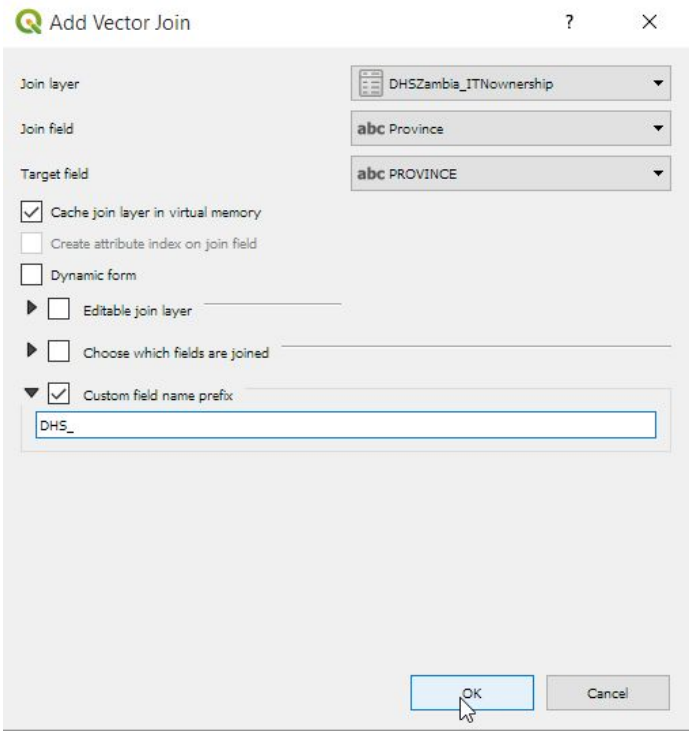
The layer will appear in the Layers panel as a table only, and no alterations will be made to the map.

To join the province-level ITN ownership data to the administrative level data, we now:

➔ Right-click on the administrative-level layer
➔ Select Properties.
➔ In the Properties window, click on Joins

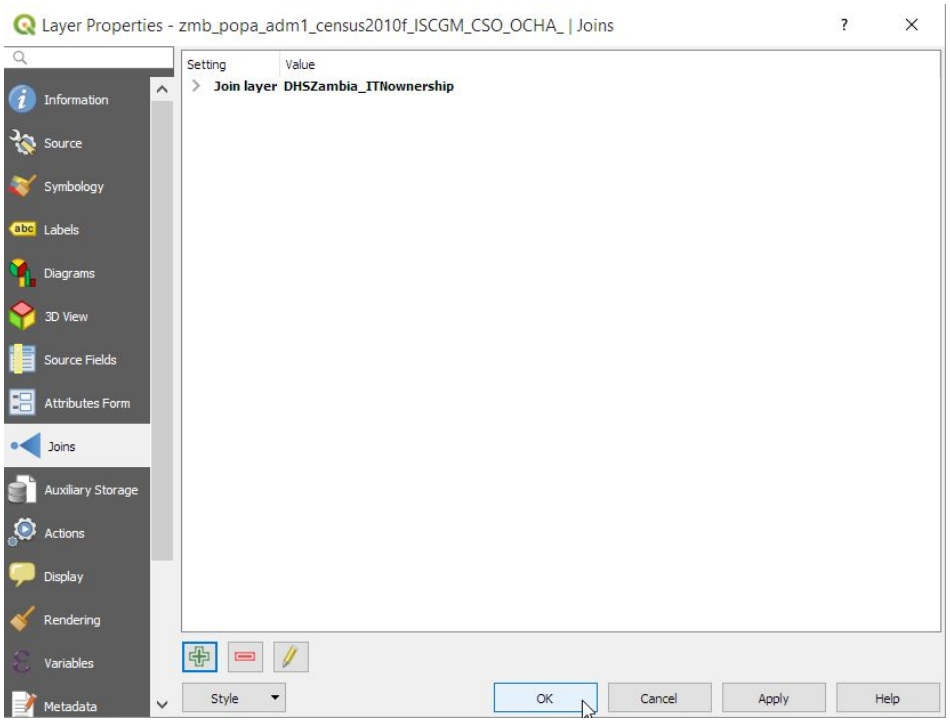


➔ Click the green + button to begin the join

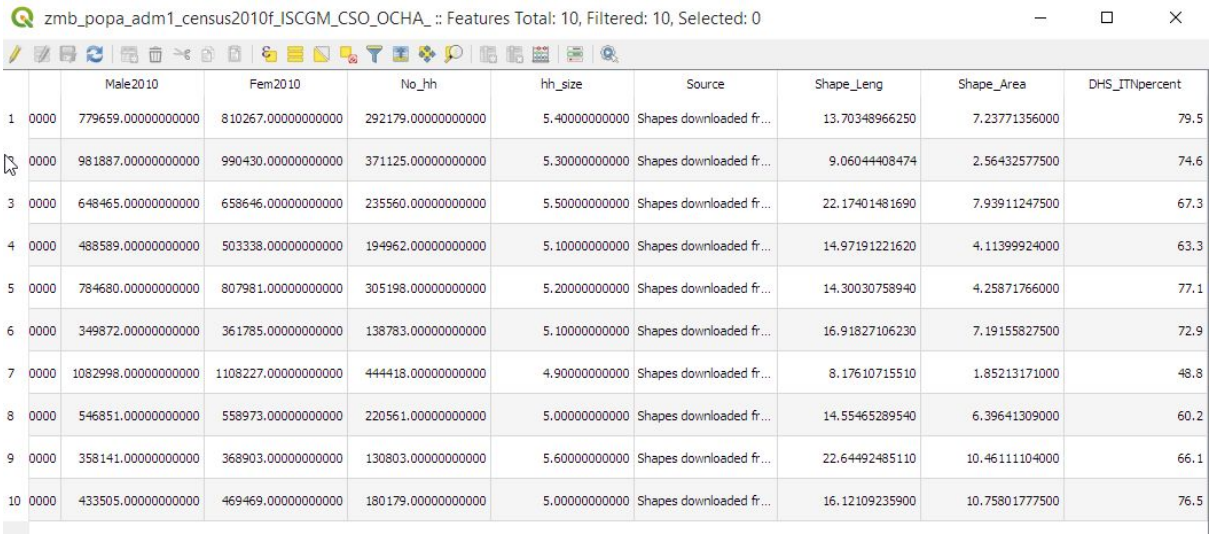


➔ The join layer is the csv-generated layer that you wish to join to the administrative level shapefile, which in this case is DHSZambia_ITNownership

- ➔ The join field is the column in DHSZambia_ITNownership which contains the unique joining ID which is *Province*
- ➔ The Target field is the column in the administrative level shapefile which contains the unique joining ID which is *PROVINCE*
- ➔ Add a custom prefix of DHS_ which will simply add 'DHS_' to the variable name of those attributes joined from the DHSZambia_ITNownership layer
- ➔ Click OK, then click OK in the Layer Properties window



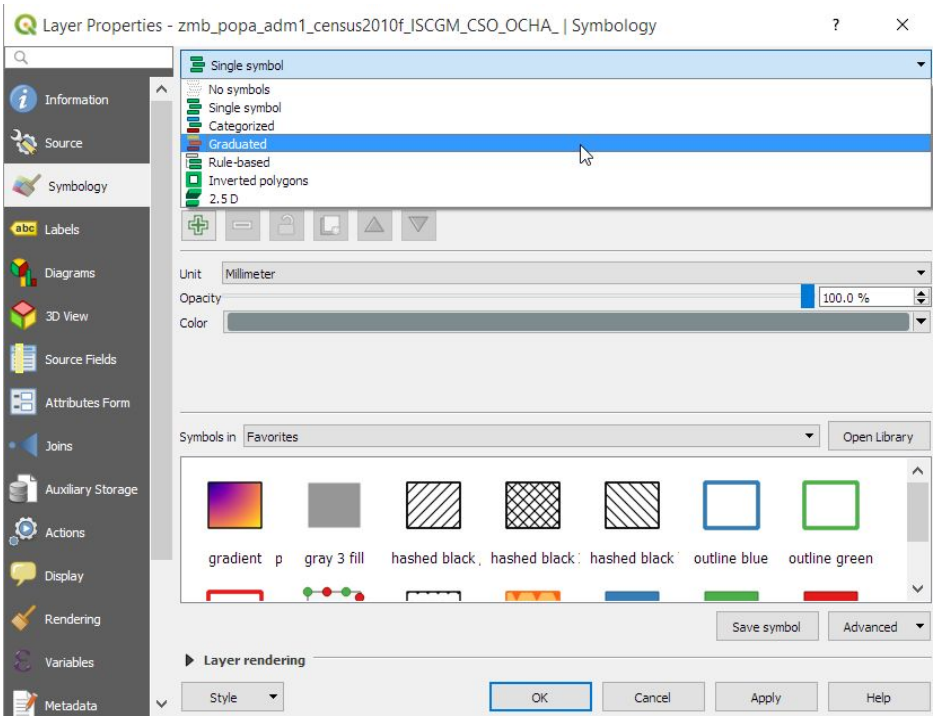
If you now look at the attribute table for the administrative level shapefile, there is now an additional column called DHS_ITNpercent

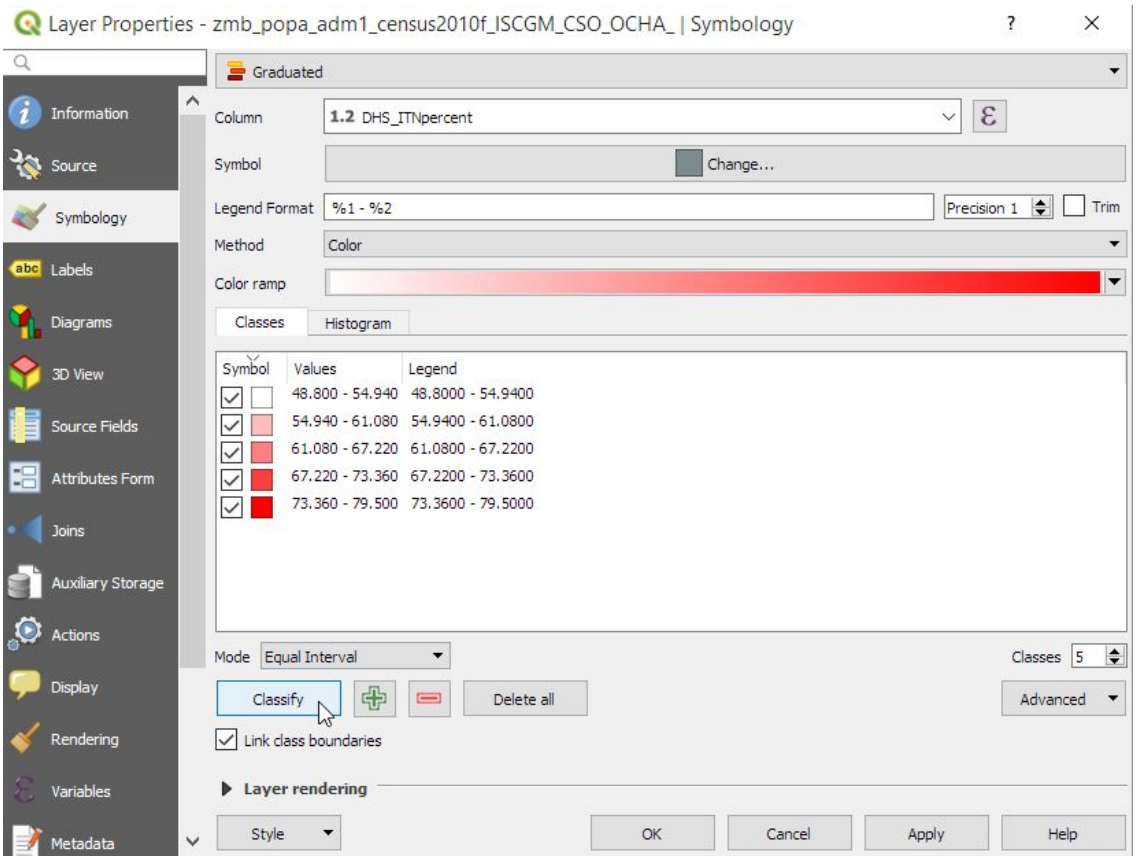


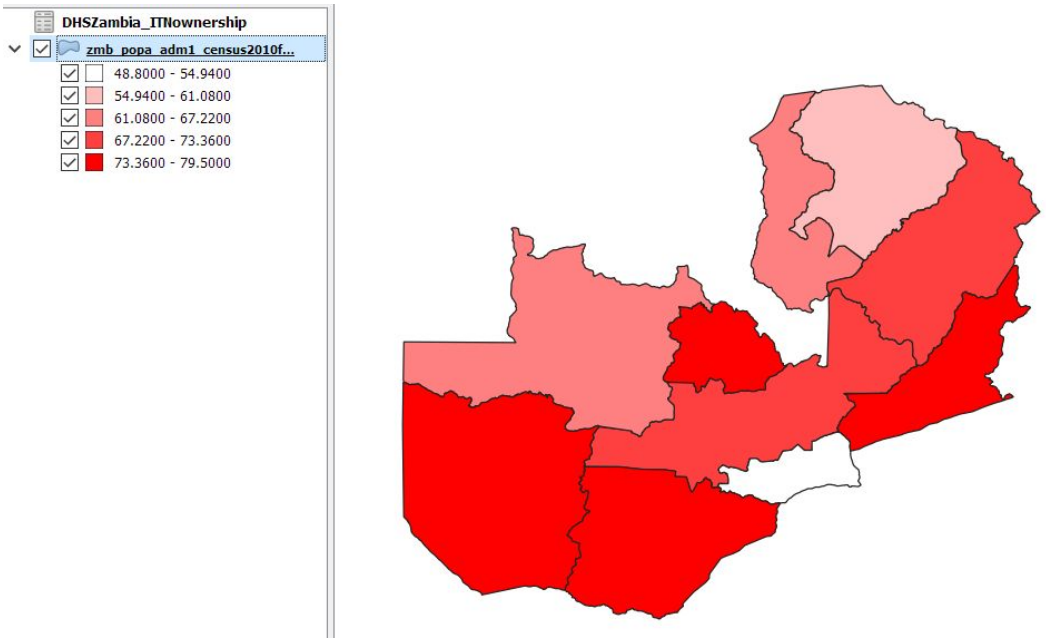
		Male2010	Fem2010	No_hh	hh_size	Source	Shape_Leng	Shape_Area	DHS_ITNpercent
1	0000	779659.00000000000	810267.00000000000	292179.00000000000	5.40000000000	Shapes downloaded fr...	13.70348966250	7.23771356000	79.5
2	0000	981887.00000000000	990430.00000000000	371125.00000000000	5.30000000000	Shapes downloaded fr...	9.06044408474	2.56432577500	74.6
3	0000	648465.00000000000	658646.00000000000	235560.00000000000	5.50000000000	Shapes downloaded fr...	22.17401481690	7.93911247500	67.3
4	0000	488589.00000000000	503338.00000000000	194962.00000000000	5.10000000000	Shapes downloaded fr...	14.97191221620	4.11399924000	63.3
5	0000	784680.00000000000	807981.00000000000	305198.00000000000	5.20000000000	Shapes downloaded fr...	14.30030758940	4.25871766000	77.1
6	0000	349872.00000000000	361785.00000000000	138783.00000000000	5.10000000000	Shapes downloaded fr...	16.91827106230	7.19155827500	72.9
7	0000	1082998.00000000000	1108227.00000000000	444418.00000000000	4.90000000000	Shapes downloaded fr...	8.17610715510	1.85213171000	48.8
8	0000	546851.00000000000	558973.00000000000	220561.00000000000	5.00000000000	Shapes downloaded fr...	14.55465289540	6.39641309000	60.2
9	0000	358141.00000000000	368903.00000000000	130803.00000000000	5.60000000000	Shapes downloaded fr...	22.64492485110	10.46111104000	66.1
10	0000	433505.00000000000	469469.00000000000	180179.00000000000	5.00000000000	Shapes downloaded fr...	16.12109235900	10.75801777500	76.5

We will now create a map of ITN ownership, using graduated colours to represent the different levels of ownership.

- ➔ Right-click the administrative levels layer, and select Properties
- ➔ In the properties window, select Symbology
- ➔ In the dropdown menu of symbology styles, select Graduated
- ➔ Select the column DHS_ITNpercent
- ➔ Click the Classify button to ground the percentages into categories
- ➔ Click OK

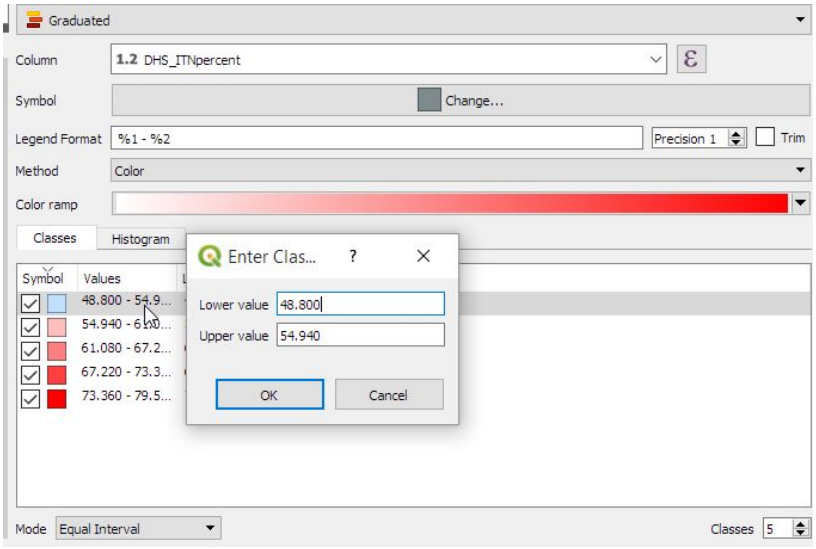






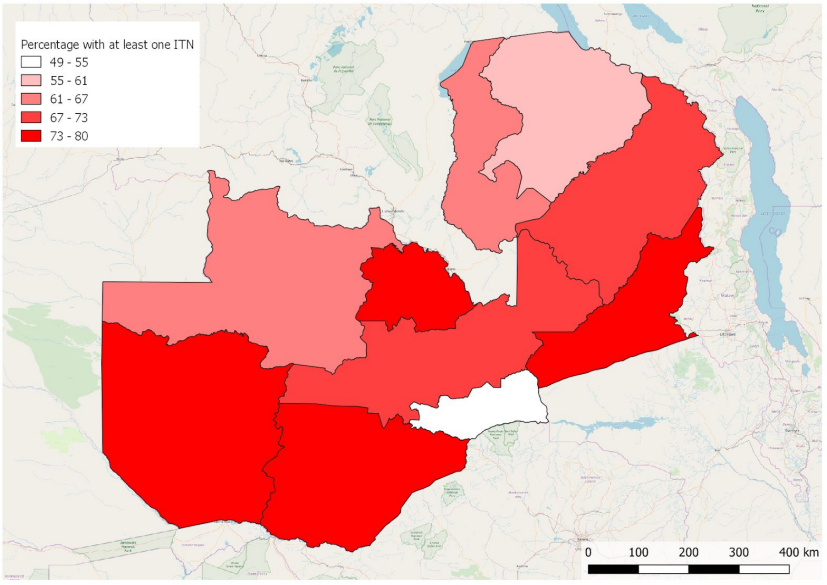
You can change the colour, and classification boundaries within the Symbology window to change the appearance of the map.

Also, remember to save your new project!



Exercise

Produce a map of the ITN ownership using the Print Layout



Exercise

Repeat the process using the DHSSambia_FeverUnder5 Excel file to produce a map of the percentage of under fives who had experienced a fever in the last two weeks.

Adding point-level data from an Excel spreadsheet

So far we have been working with point-level data that has already been saved in a spatial data format i.e. a shapefile. Commonly, point data may be given to you in the format of an Excel file with

columns representing the coordinates of the points, plus additional columns containing information relating to those points. In this section we will show you how to convert this type of data to a shapefile that you can display in QGIS.

As an example, we have randomly generated some points within one of the districts in Zambia to represent locations where some resistance testing has been undertaken. At each site we have simulated the mosquito mortality percentage resulting from these hypothetical tests. We suppose that the coordinates for each site were recorded using a GPS receiver, and typed up in an Excel file. The coordinates are in the format of longitude and latitude (WGS84 geographical coordinate system).

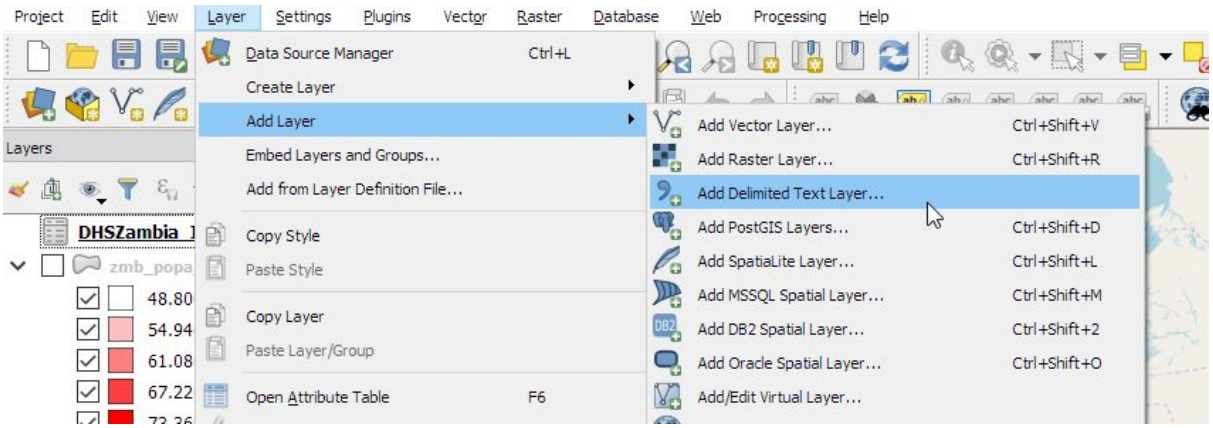
Note that when converting data from Excel into a shapefile in QGIS, the column names become truncated so that only the first 10 characters are used. Therefore you need to make sure that the names you use as your column headings are still interpretable after they have been truncated.

	A	B	C	D	E
1	Longitude	Latitude	SiteID	perc_mortality	
2	32.43885	-12.4838	1	99	
3	33.10082	-12.4773	2	100	
4	32.26809	-13.6394	3	91	
5	30.7441	-14.5338	4	100	
6	32.33721	-14.052	5	97	
7	32.69667	-12.1916	6	89	
8	30.73129	-14.3498	7	92	
9	31.98902	-12.9122	8	98	
10	30.08276	-14.4896	9	100	
11	32.54188	-13.5443	10	94	
12	33.01506	-12.462	11	89	

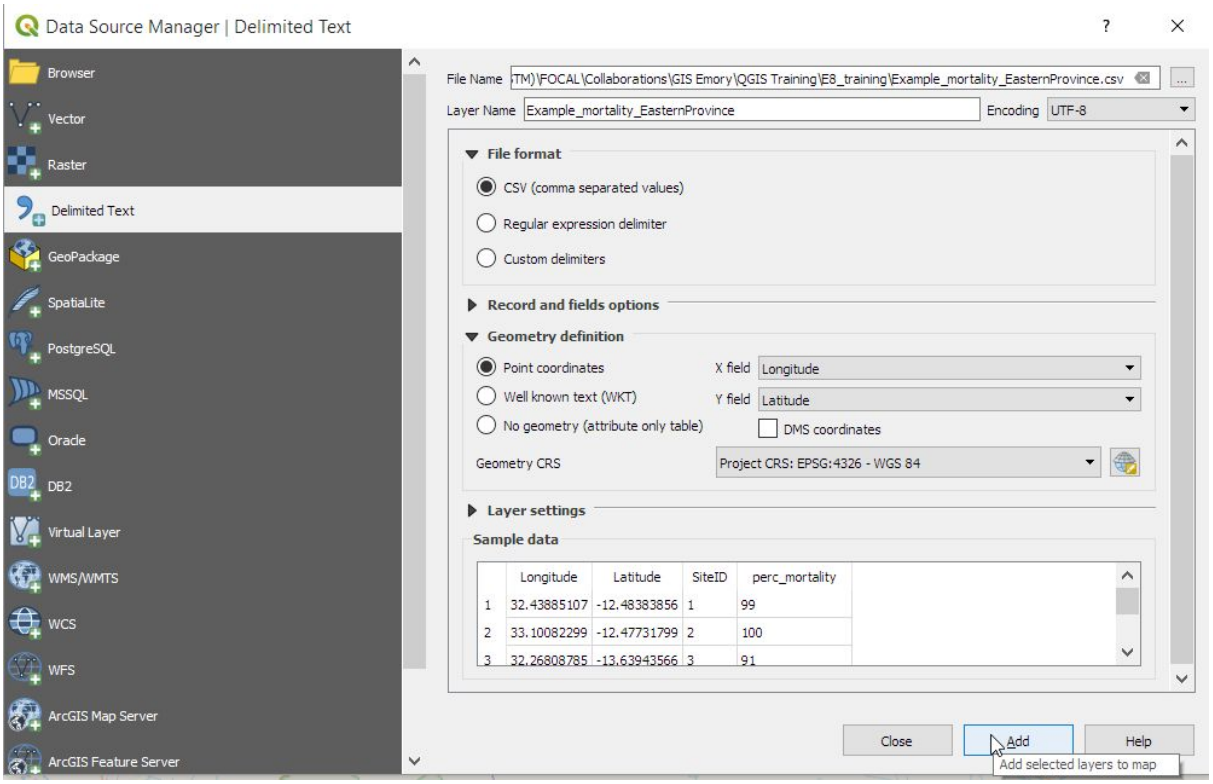
Data from 20 sites were generated. We will now add these to QGIS, convert them to a spatial object and create a map of the data.

As before, we first need to convert the Excel file to the CSV format (File, Save As, CSV (Comma delimited)). Once it has been converted we can add it to QGIS as follows:

➔ Go to Layer, Add Layer, Add delimited text layer

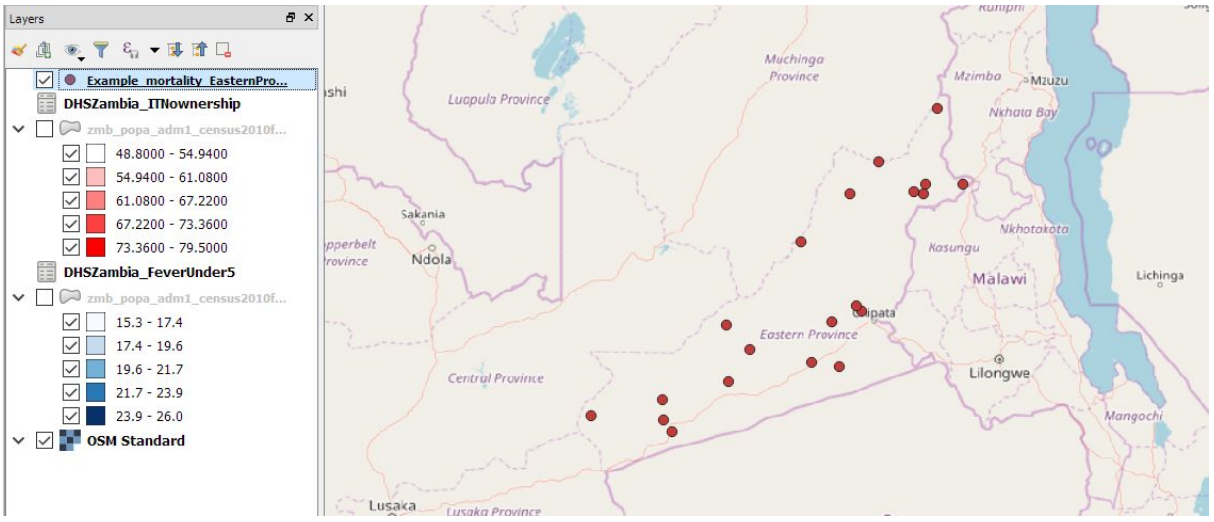


- ➔ In the resulting window, select the CSV you have just saved (in this case Example_mortality_EasternProvince.csv)
- ➔ Keep the default file format of CSV
- ➔ Expand the Geometry definitions options and select Point coordinates. The X field is the column representing longitude, and the Y field is that representing latitude (in this example)
- ➔ Select the coordinate system in which the coordinates are recorded which in this case is WGS84
- ➔ Click Add



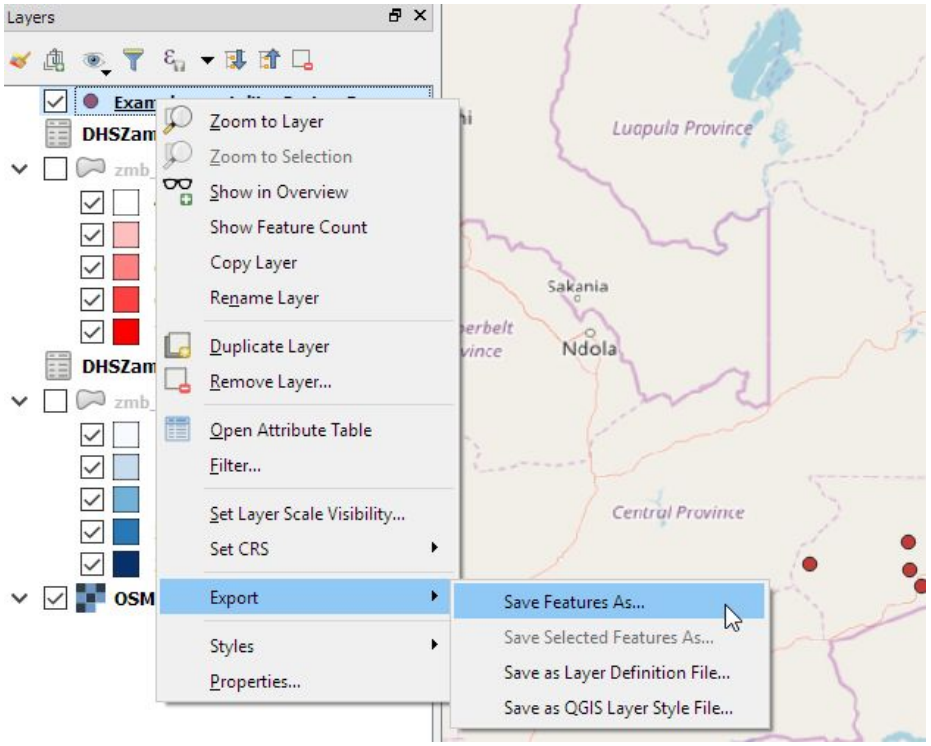
The points should now appear in the Eastern Province of Zambia.

- ➔ Add a basemap to check the points are in the area where you expect them to be! (Web, Quickmapservices)



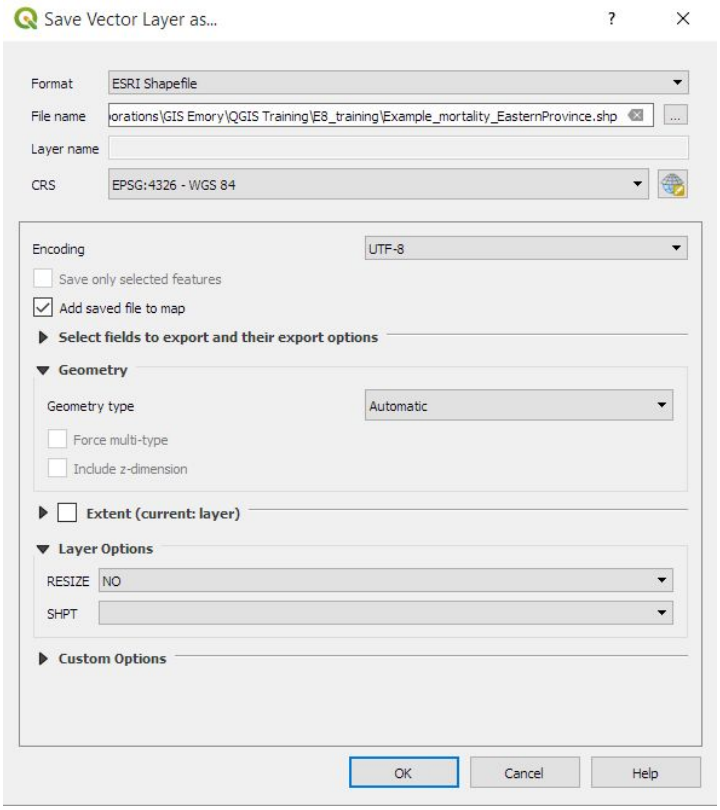
You now want to save this layer as a shapefile. By doing this, you don't have to repeat the conversion process every time you want to create a map using these data.

➔ Right-click the layer, select Export, then Save Features As...



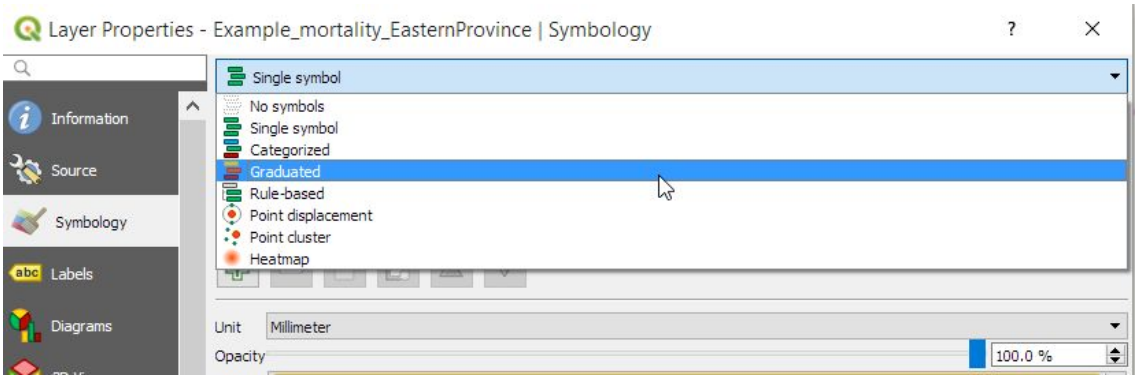
➔ Change the format to ESRI Shapefile, and select a file name. Then click OK.

➔ The shapefile will be added to the Layers panel and you can now remove the layer created from the CSV file.

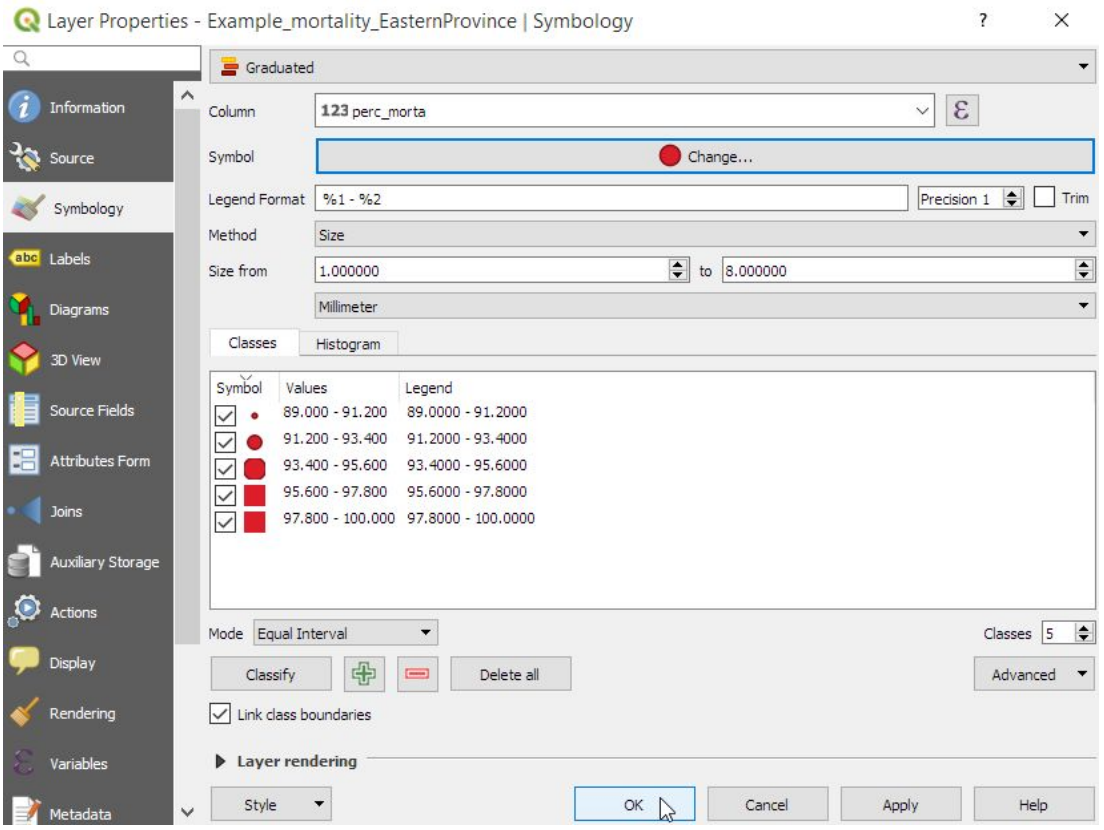


We now want to create a map where the size of the circles represents the mosquito mortality rate.

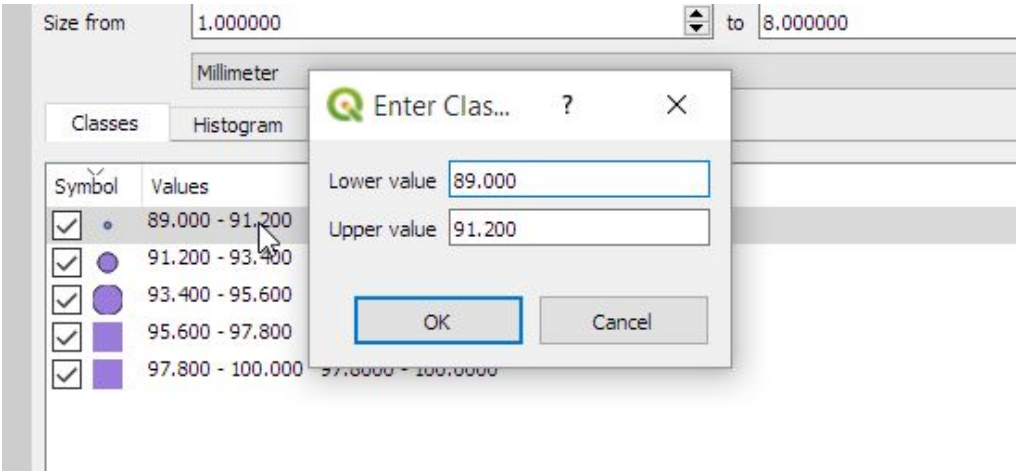
- ➔ Right-click the layer in the layers panel and select Properties
- ➔ Change the format type from Single symbol to Graduated



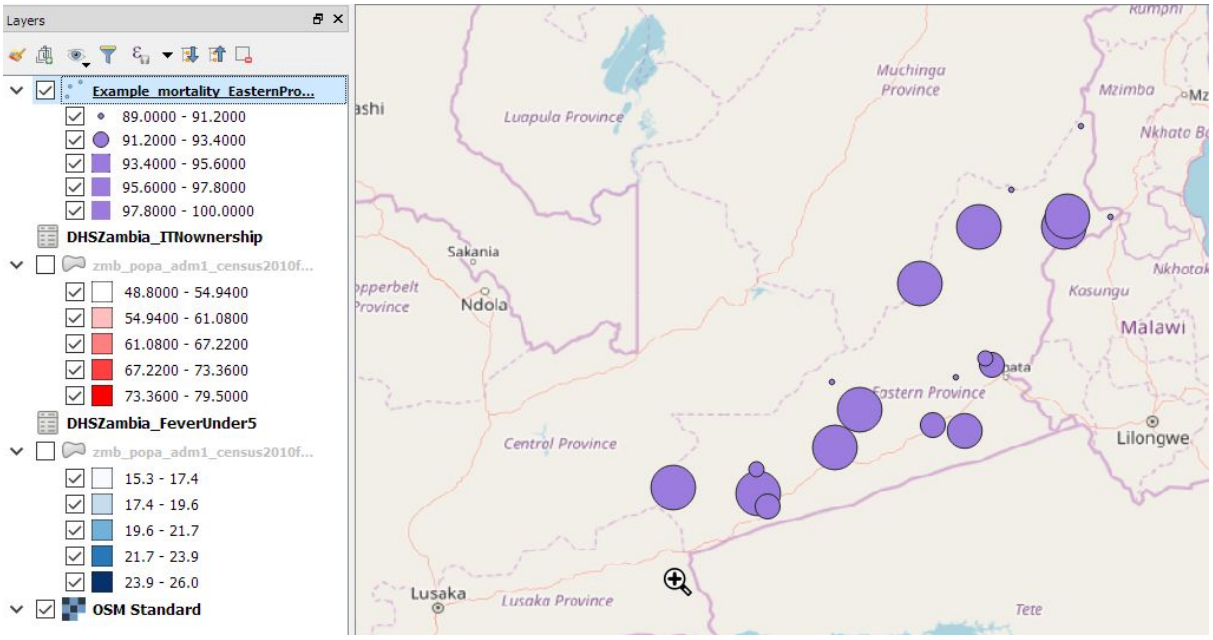
- ➔ Select the column on which to base the symbol size on, which has been truncated to perc_morta, and change the method from Colour to Size
- ➔ Click Classify. If you're happy with the automated category intervals, click OK



If you are not happy with the automated intervals you can change these either by changing the mode in which the intervals are generated from Equal Interval to another method, or you can edit the Values column manually.



Your map will now look something like this



Exercise

Produce a map of the mosquito mortality rate using the Print Layout

Creating maps from your own data

We are dedicating the afternoon to producing maps from data that you have brought with you. This could be :

- an excel file with coordinates : you will need to save as a csv then use Layer, add layer, add delimited text layer
- an excel file with administrative names : you will need to Join to an admin polygons layer after saving as a csv and use Layer, add layer, add delimited text layer.
- a gpx file from a gps device : you will need to use Layer, add layer, add vector layer
- you could search on https://data.humdata.org/dataset to find other data that can be useful for your work.

We will be available to answer any questions and difficulties.

By the end of the afternoon we would like everyone to have made a map using QGIS and Print Layout so that they can show this to the group and say a few words about it.

Good luck !