NCG608 Data Science Assignment

Rahul Jadhav (17250785)

Contents

1. Introduction

In addition to food, cloth and shelter, internet has become one of the most important human needs these days. It is scary to imagine life without internet. There are various ways in which we can access this ocean of information. Domestic wired broadband, 3G/4G mobile broadband and satellite broadband are predominantly used options.

Surrounded by Atlantic Ocean, Republic of Ireland is a sovereign state in Northwestern Europe with about 5 million inhabitants spread little over 70000 square Kilometers.

31 counties in Republic of Ireland forms a basis of local government. These 31 counties are further divided into 3409 Electoral Divisions (EDs) which are the smallest legally defined administrative areas.

Purpose of this report is to describe in detail the study carried out to investigate the penetration of domestic broadband networking in Republic of Ireland. As part of this study, we will produce maps and provide relevant statistics (at County & ED level) highlighting key aspects with respect to domestic broadband penetration.

2. Data Used

2.1 Shapefiles

Source: data.gov.ie

- County Shapefile - Shapefile containing County Boundaries
- ED Shapefile - Shapefile containing ED Boundaries

2.2 CSV Files

Source: cso.ie/en/census

- County CSV File - CSV File with 2016 Census Statistics by County
- ED CSV File - CSV File with 2016 Census Statistics by ED

3. Tools Used

3.1 R

RStudio is open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics.

R is used for manipulating data in CSV files. It is also used for generating some informative plots based on the data in CSV files.

3.2 QGIS

QGIS is open-source cross-platform desktop GIS application that supports viewing, editing, and analysis of geospatial data.

It is used for converting Coordinate Referencing System (CRS) of both shapefiles from WGS84 to local Irish National Grid (EPSG:29902).

QGIS is also used for generation aesthetic maps telling the story of domestic broadband penetration in the country.

3.3 Postgres (with PostGIS extension)

Postgres is open-source RDBMS and PostGIS adds support for geographic objects to it.

Postgres is used to calculate area of Counties and EDs in square Kilometers, which is further used to calculate population density.

4. Data Manipulation

4.1 Using R

We have 2 CSV files containing 2016 Census data. One is at County level (with 31 records) and other at ED level (with 3409 records).

In CSV files, column with name 'T15_3_B' gives us number of households with Broadband Connection in respective County/ED.

Considering total number of households will not give precise picture in our analysis as all counties and EDs are not of same size in terms of total households. Thus, we have calculated proportion of houses having domestic broadband connection (dividing number of Broadband Households by Total Households). Thus, we have proportion of households with network broadband connection in last column of both dataframes.

Detailed steps are available in attached R markdown & PDF file.

R_Manipulations.Rmd

R Manipulations.pdf

Now, we have 2 CSV files (counties.csv and eds.csv) containing everything we need as part of our study.

4.2 Using QGIS

We have 2 shapefiles: one for Counties and other for EDs. Both of them are in CRS - World Geodetic System 1984 (i.e. WGS84). We have to convert them to local Irish National Grid (EPSG:29902).

We have opened these shapefiles in QGIS by adding Vector Layer. Once opened, by right clicking on the layer and selecting 'save as' option, we have saved these layers as geoJSON files by changing their CRS code to EPSG:29902.

This has been done to minimize the distortion caused by map projection and to get accurate analysis results.

4.3 Using Postgres

We need to know area of County and ED so as to calculate Population Density.

Using DB Manager of QGIS, we have imported geoJSON files (from above step) in Postgres as tables with names 'counties' and 'eds'.

In Postgres, following queries gave us area in square KMs.

```
select id, "COUNTY", ST_Area(geom)/1000000 as Area from counties;
```

```
select id, "ED_ENGLISH", ST_Area(geom)/1000000 as Area from eds;
```

Notice that here we have not used ST_Transform as our geoJSON files are already transformed to local CRS 29902.

Now, using following SQL commands, we have added ' area' as a column to tables counties and eds.

```
alter table counties add column area real default 0.0;
update counties set area = ST_Area(geom)/1000000;

alter table eds add column area real default 0.0;
update eds set area = ST_Area(geom)/1000000;
```

From DB Manager, adding these 2 tables as canvas in QGIS gives us 2 final layers to work with. We confirmed it by checking if calculated area appears as a column in 'Attribute Table'.

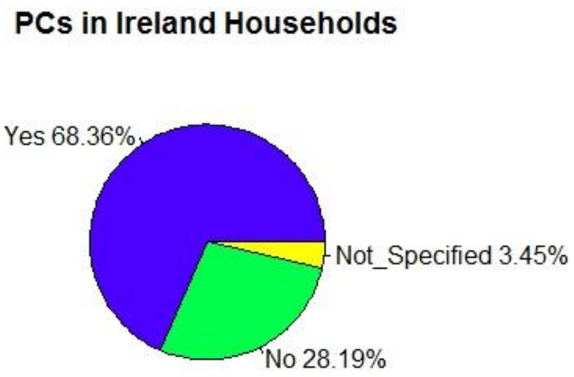
5. Plots in R & Choropleth Maps in QGIS

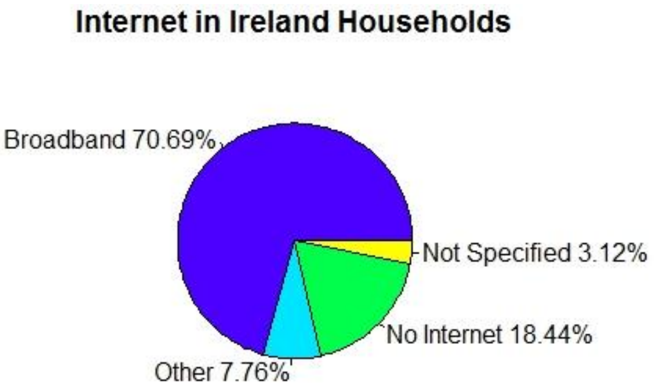
5.1 Plots in R

Let us draw Pie-Charts to visualize PC owner details and Internet Connection Type across Ireland:

Detailed R code in attached markdown file:

R_Plots.Rmd





5.2 Choropleth Maps in QGIS

a. Importing CSV files (generated after R Manipulations) in QGIS

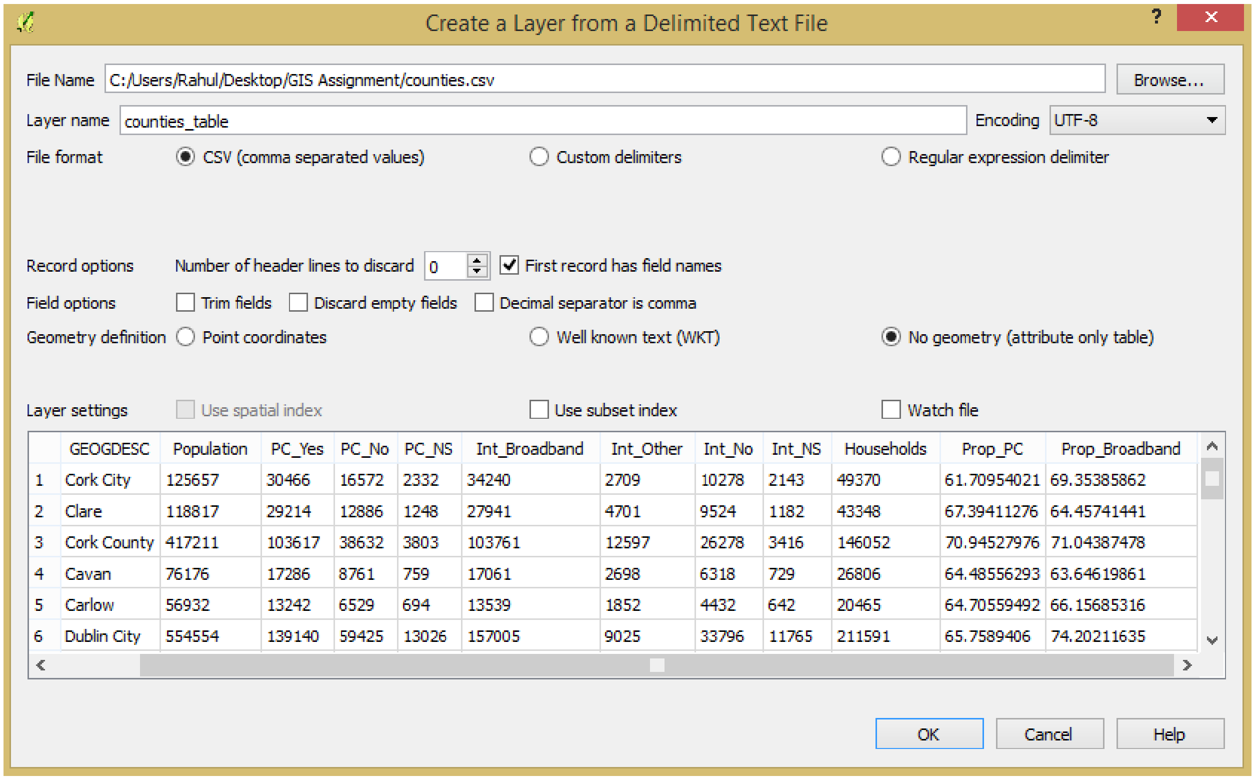
For Counties and EDs, we have to visualize data in these csv files on layers produced by adding counties and eds tables (as canvas) in QGIS.

We can directly import csv file into QGIS by 'Adding Delimited Text Layer'. However, doing so will generate a table in QGIS with all columns of type 'text'. Columns of our interest in csv files are numeric. Hence, to be able to perform mathematical operations on the data, we should have these column as numeric and not as a text.

To tackle this issue we need to create a sidecar file with .csvt extension in same folder. This file will have only 1 row specifying data type for each column in csv file.

Our csv files are 'counties.csv' and 'eds.csv'. So, let us create files 'counties.csvt' and 'eds.csvt' containing only 1 row with datatypes for all columns in csv files.

Now, as csv and csvt files are available in same directory, we will import csv files into QGIS as tables by selecting *'Layer -> Add Layer -> Add Delimited Text Layer'* menu options. We will ensure that file format chosen is 'csv'. Also, our csv files does not contain any geometry data, we specify it by checking 'No Geometry' option.



Create a Layer from a Delimited Text File

File Name C:/Users/Rahul/Desktop/GIS Assignment/counties.csv

Layer name counties_table Encoding UTF-8

File format ● CSV (comma separated values) ○ Custom delimiters ○ Regular expression delimiter

Record options Number of header lines to discard 0 ☑ First record has field names

Field options □ Trim fields □ Discard empty fields □ Decimal separator is comma

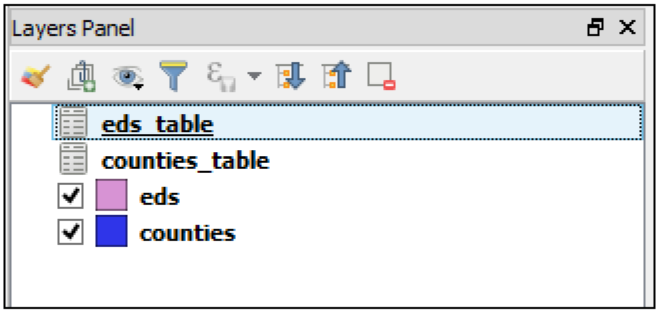
Geometry definition ○ Point coordinates ○ Well known text (WKT) ● No geometry (attribute only table)

Layer settings □ Use spatial index □ Use subset index □ Watch file

	GEOGDESC	Population	PC_Yes	PC_No	PC_NS	Int_Broadband	Int_Other	Int_No	Int_NS	Households	Prop_PC	Prop_Broadband
1	Cork City	125657	30466	16572	2332	34240	2709	10278	2143	49370	61.70954021	69.35385862
2	Clare	118817	29214	12886	1248	27941	4701	9524	1182	43348	67.39411276	64.45741441
3	Cork County	417211	103617	38632	3803	103761	12597	26278	3416	146052	70.94527976	71.04387478
4	Cavan	76176	17286	8761	759	17061	2698	6318	729	26806	64.48556293	63.64619861
5	Carlow	56932	13242	6529	694	13539	1852	4432	642	20465	64.70559492	66.15685316
6	Dublin City	554554	139140	59425	13026	157005	9025	33796	11765	211591	65.7589406	74.20211635

By clicking 'OK' we will have counties.csv file imported in QGIS as table. We will do the same for eds.csv file.

In QGIS Layers Panel we have layers 'eds' and 'counties' added from postgres tables. Also, we have 'eds_table' and 'counties_table' tables imported from csv files generated after manipulations in R.



Layers Panel

eds_table
counties_table
☑ eds
☑ counties

Subsequently, we are going to rename these vector layers as per the information they depict.

b. Joining Data from table files with layers

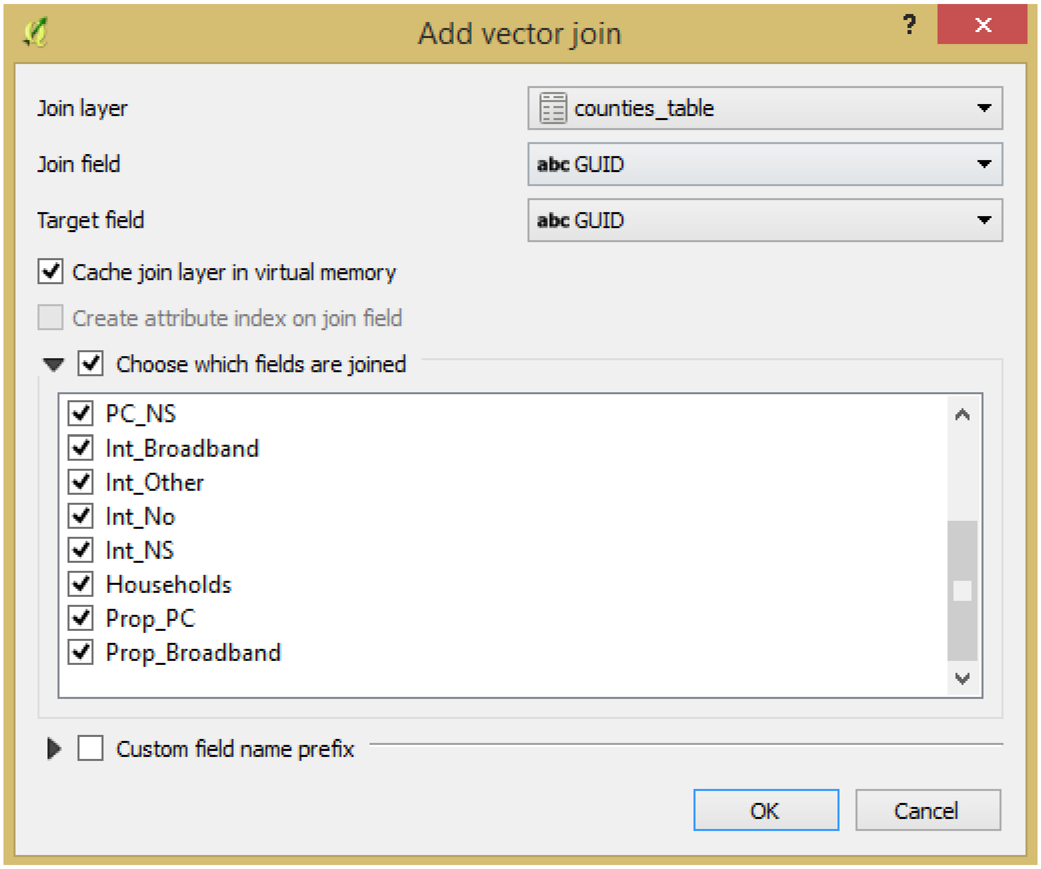
Now, all the data we require for analysis is in QGIS and it is in the format we need.

Next task is of joining tables with layers. This is due to the fact that our Boundaries and Area details are in layers and 'Proportion of PC/Broadband Households' is in tables.

For this purpose we will use ***'Table Join'*** feature of QGIS. Yes, QGIS is as good as any RDBMS in making joins just in few clicks! We will make a join on GUID field which is available in both tables and layers.

Right Click on layer to navigate to properties and then we have to select 'Joins' option. In Layer properties dialog, we have to select + button at the bottom to make a join with respective table.

Following is the screenshot of 'Add Vector Join' for Counties.



Add vector join

Join layer: counties_table

Join field: abc GUID

Target field: abc GUID

☑ Cache join layer in virtual memory

□ Create attribute index on join field

☑ Choose which fields are joined

- ☑ PC_NS
- ☑ Int_Broadband
- ☑ Int_Other
- ☑ Int_No
- ☑ Int_NS
- ☑ Households
- ☑ Prop_PC
- ☑ Prop_Broadband

□ Custom field name prefix

Here, we are joining table and layer on GUID column. Once join is established, we can see columns from table appears under 'Open Attribute Table' option for layers 'counties' and 'eds'.

We will add one column as 'Pop_Density' (Population/Area) through attribute table.

This ultimately means that we can now happily visualize area, population density, proportion of households with Broadband Internet on maps.

c. Cartography in action!

Map is the model of reality. Any map generally provides a snapshot of some process, it is a frozen picture showing status at one point in time. However, it is harder to compare changes over time.

We are creating Thematic Maps in this study. More precisely 'Choropleth Maps' where boundaries around zones (Counties and EDs) are established independently of the data.

We have chosen to use OpenStreetMap as a background layer.

For both Counties and EDs, we have to navigate to properties of layer and apply 'Graduated' style on 'prop' column which contains proportion of households with broadband internet. We can choose color ramp in such a way that it looks aesthetic with respect to the background map we have chosen and context of the problem statement we are trying to solve/study. One of the very important aspect is the choice of class interval. By default QGIS chooses equal intervals with 5 classes. In this case, range of proportion gets equally divided in 5 classes.

Instead of 'Equal Intervals', if we choose mode as 'Quantile (Equal Count)' then sorted column values get arranged in number of classes we have specified.

We have another mode as 'Natural Breaks'. Using this mode, QGIS creates classes by finding natural groupings of data. The resulting classes will be such that there will be maximum variance between individual classes and minimum variance within each class. Variance is one of the important measure to study the scatter within dataset with respect to mean.

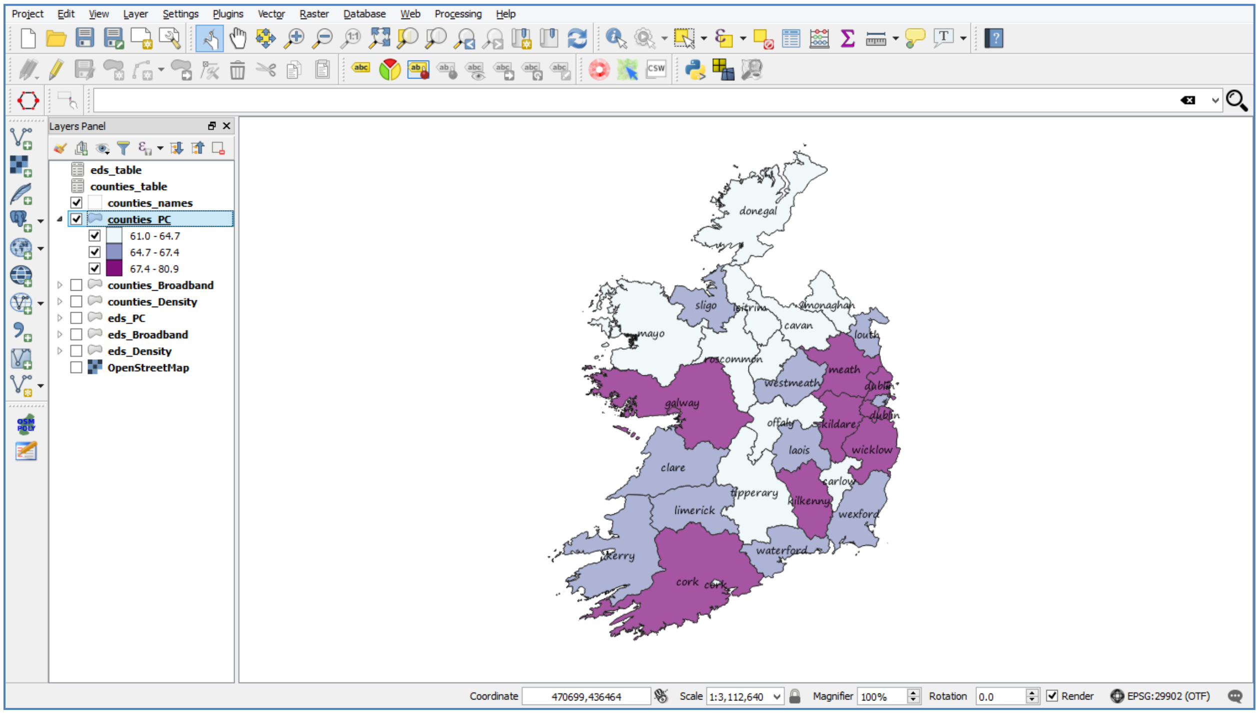
County level Maps:

In counties data, we have 31 counties so we have 31 values in proportion column(s). On choosing **'Quantile (Equal Count)'** mode for graduation, QGIS sorts data in proportion column in ascending order and as we have chosen classes as 3, it segregates those sorted 31 values with 10 in each class (with 11 in last as we have 31 counties). This has nothing to do with concept of quantiles in statistics which actually calculates percentiles to form 5 number summary (min, 25th percentile, 50th percentile (median), 75th percentile & max). For counties map, we have chosen **3 classes** considering the small number of counties (i.e. 31). It becomes easy to visualize less data in less number of classes.

To graduate the household broadband proportion data for Counties, we are using 3 classes formed by applying 'Quantile (Equal Count)' mode. Here we have less number of observations (31) and Quantile mode will give us 3 classes with 10 counties each. It becomes easy for naked eye to distinguish difference in scatter if we limit classes to 3 (say LOW, MEDIUM and HIGH).

Colors chosen are self explanatory in terms of displaying the underlying information.

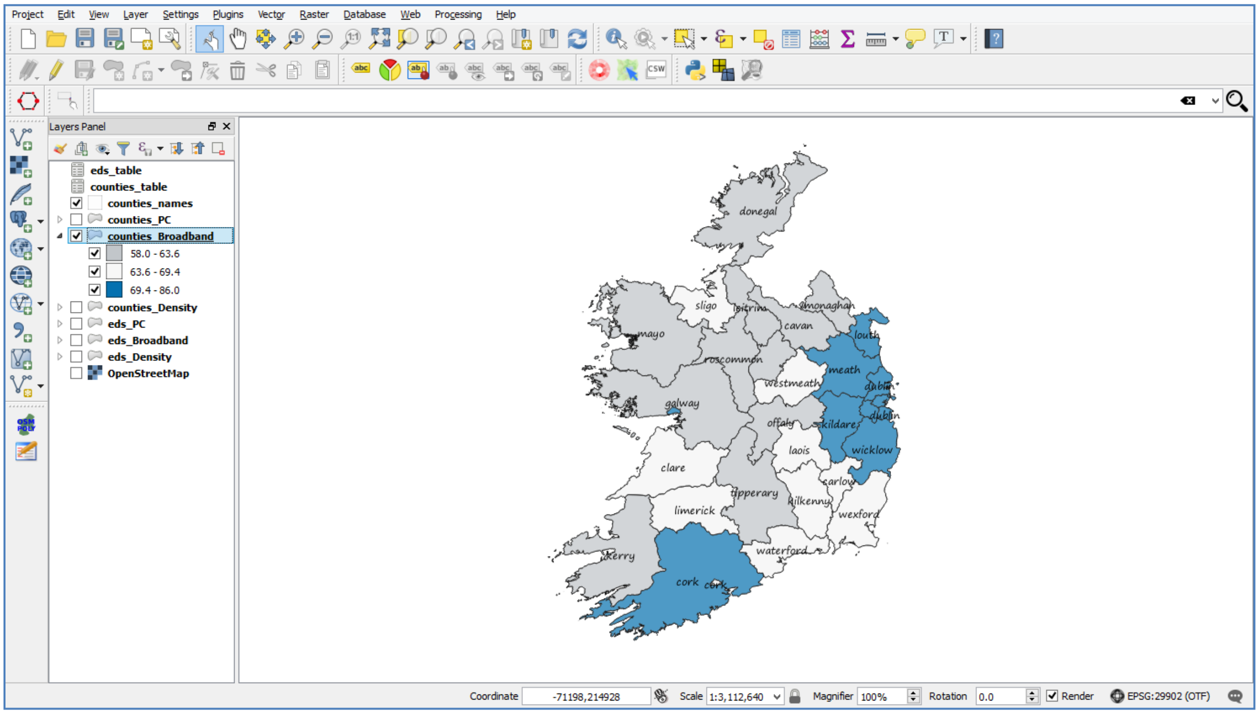
PC proportion map



- ▦ eds_table
- ▦ counties_table
- ☑ counties_names
- ☑ counties_PC
  - ☑ □ 61.0 - 64.7
  - ☑ ▨ 64.7 - 67.4
  - ☑ ■ 67.4 - 80.9
- ☐ counties_Broadband
- ☐ counties_Density
- ☐ eds_PC
- ☐ eds_Broadband
- ☐ eds_Density
- ☐ OpenStreetMap

Coordinate 470699,436464 Scale 1:3,112,640 Magnifier 100% Rotation 0.0 ☑ Render EPSG:29902 (OTF)

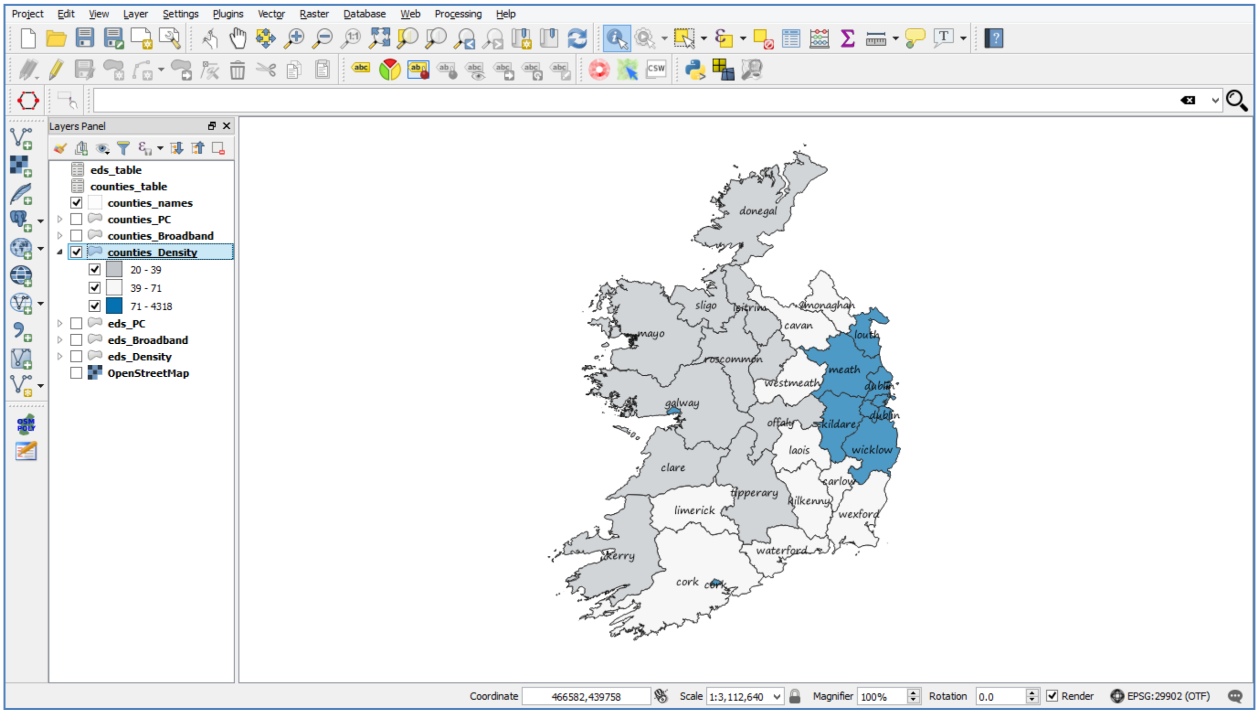
Broadband proportion map



- ▦ eds_table
- ▦ counties_table
- ☑ counties_names
- ☐ counties_PC
- ☑ counties_Broadband
  - ☑ ▨ 58.0 - 63.6
  - ☑ □ 63.6 - 69.4
  - ☑ ■ 69.4 - 86.0
- ☐ counties_Density
- ☐ eds_PC
- ☐ eds_Broadband
- ☐ eds_Density
- ☐ OpenStreetMap

Coordinate -71198,214928 Scale 1:3,112,640 Magnifier 100% Rotation 0.0 ☑ Render EPSG:29902 (OTF)

Population Density map



Layers Panel

- eds_table
- counties_table
- ☑ counties_names
- ☐ counties_PC
- ☐ counties_Broadband
- ☑ counties_Density
  - ☑ 20 - 39
  - ☑ 39 - 71
  - ☑ 71 - 4318
- ☐ eds_PC
- ☐ eds_Broadband
- ☐ eds_Density
- ☐ OpenStreetMap

Coordinate 466582,439758 Scale 1:3,112,640 Magnifier 100% Rotation 0.0 ☑ Render EPSG:29902 (OTF)

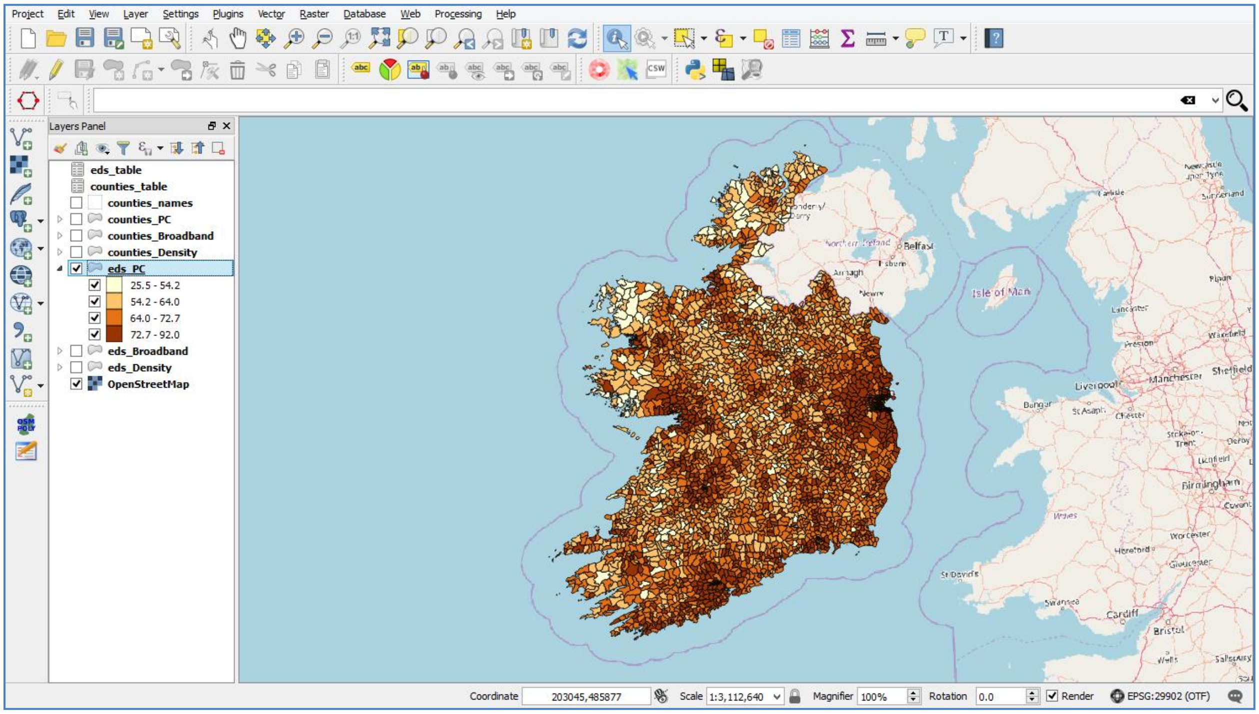
ED level Map:

For eds data (), we use 'Natural Breaks' to classify the proportions. Using this mode, QGIS creates classes by finding natural groupings of data. The resulting classes will be such that there will be maximum variance between individual classes and minimum variance within each class. Variance is one of the important measure to study the scatter within dataset with respect to mean.

To graduate the household broadband proportion data for EDs, we are using 4 classes formed by applying 'Natural Breaks'. This is because as we have more data values and it is logical (and visually gratifying) to divide proportions based on variance (or scatter) within and among the classes.

As there are 3409 EDs with small area to display on the Ireland map. For population density map, we have chosen quantile (Equal Counts) classification so that it groups EDs by equal number of observations. This is useful from visualization point of view.

PC proportion map

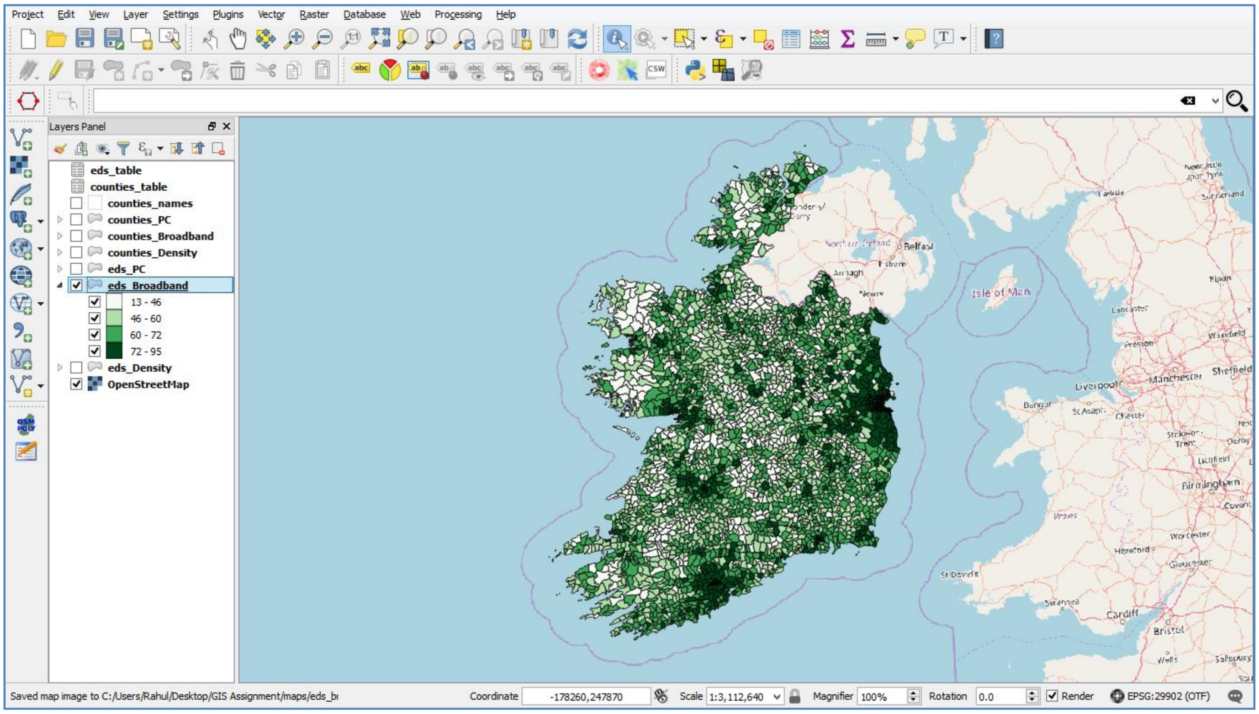


Layers Panel

- eds_table
- counties_table
- □ counties_names
- □ counties_PC
- □ counties_Broadband
- □ counties_Density
- ☑ eds_PC
  - ☑ 25.5 - 54.2
  - ☑ 54.2 - 64.0
  - ☑ 64.0 - 72.7
  - ☑ 72.7 - 92.0
- □ eds_Broadband
- □ eds_Density
- ☑ OpenStreetMap

Coordinate 203045,485877 Scale 1:3,112,640 Magnifier 100% Rotation 0.0 ☑ Render EPSG:29902 (OTF)

Broadband proportion map

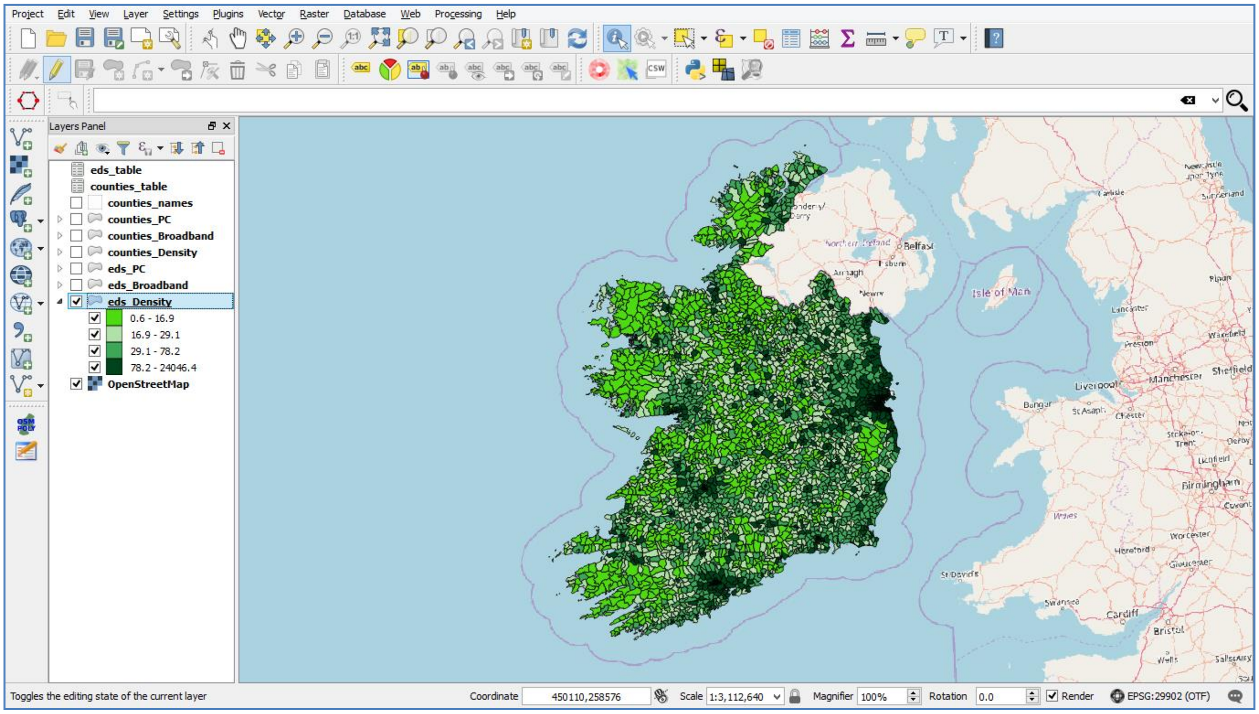


Layers Panel

- eds_table
- counties_table
- □ counties_names
- □ counties_PC
- □ counties_Broadband
- □ counties_Density
- □ eds_PC
- ☑ eds_Broadband
  - ☑ 13 - 46
  - ☑ 46 - 60
  - ☑ 60 - 72
  - ☑ 72 - 95
- □ eds_Density
- ☑ OpenStreetMap

Saved map image to C:/Users/Rahul/Desktop/GIS Assignment/maps/eds_bı Coordinate -178260,247870 Scale 1:3,112,640 Magnifier 100% Rotation 0.0 ☑ Render EPSG:29902 (OTF)

Population Density map


Project Edit View Layer Settings Plugins Vector Raster Database Web Processing Help
Layers Panel
Belfast
Isle of Man
Liverpool
Manchester
Sheffield
Birmingham
Cardiff
Bristol

- eds_table
- counties_table
- counties_names
- counties_PC
- counties_Broadband
- counties_Density
- eds_PC
- eds_Broadband
- ☑ eds_Density
  - ☑ 🟩 0.6 - 16.9
  - ☑ 🟩 16.9 - 29.1
  - ☑ 🟩 29.1 - 78.2
  - ☑ 🟩 78.2 - 24046.4
- ☑ OpenStreetMap

Toggles the editing state of the current layer Coordinate 450110,258576 Scale 1:3,112,640 Magnifier 100% Rotation 0.0 ☑ Render EPSG:29902 (OTF)

6. Patterns Observed

County Level:

It is evident from Counties map that Dublin & Cork counties have **highest** (69% to 86%) proportion of households with broadband internet connection. This seems logical as Dublin and Cork being largest cities in Ireland having thriving business eco structure.

Waldo Tobler's first law of geography say that everything is related to everything else, but near things are related more than distant things. We also know that 'Spatial Autocorrelation' is the measurement of tendency of neighbors to exhibit similarity.

Counties adjacent to Dublin and Cork come under **medium** category with 63% to 70% of households connected to internet broadband network. Remaining counties which are rural and less populated stands last with 58% to 63% of households connected to broadband network.

All this spatial variability can be attributed to costs associated with setting up infrastructure required for broadband network. For telecom authorities and government bodies, it makes sense to invest in infrastructure in larger cities like Dublin & Cork given population and existence of businesses in the area. However, in rural counties with less population, it is not wise to invest large sums on infra setup. However, there are other technologies at our disposal like 3G/4G mobile broadband and satellite broadband. High speed internet can be availed in rural counties using Wi-Fi routers and signal extenders using mentioned technologies.

ED Level:

'eds_broadband map' is categorized in 4 classes with equal number of counts (3409/4). It shows that half of EDs in Ireland have less than 60% broadband penetration.

EDs displayed in dark green (class with broadband penetration > 72%) are mostly within urban area i.e. area with higher population density. Higher concentration of these EDs is in Dublin, Cork, Limerick and Galway counties.

However, there are few EDs in rural area (considering population density) with higher broadband penetration.

7. Challenges Faced

- Performing analysis at higher level (County) and at subset level (ED) based on same parameter (proportion of households with broadband) is tricky. Example - In County map, Kildare is categorized with HIGH proportion of broadband penetration. However, at ED level map there are many EDs within Kildare county where broadband penetration is considerably lower. This is nothing but the 'Specification Problem' - Shall we use points or area in Choropleth map?

- Results obtained from aggregated data (e.g. counties) cannot be assumed to apply to individual EDs within the county, this is nothing but Ecological fallacy while interpreting statistical data.

- Choosing colors for map which will tell the story. It is more art than science.

8. Conclusion

Research is more about people and less about results. Here, from studying population density & broadband penetration across the country, we can conclude that Ireland is a rural country with several urban centers. This analysis easily boils down to the fact that Agriculture is an important part of the country's economy.

If we had the data about farm area under cultivation per county/ed, we would have easily noticed less broadband penetration in counties/eds with more farm area. Also, we could have visualized the data more precisely using heatmaps if we had point data for households.