# NCG 608 Project

# Broadband Network penetration in Ireland (Year 2016)

**BY**

**OMKAR SADEGAONKAR**

**STUDENT NUMBER - 17251802**



**Introduction:**

The world has started making way towards the digitization, for which internet connectivity is the most vital need. The era has unintentionally declared to have compulsory internet connection if you want to survive comfortably with experiencing the developments. By looking at the need, people started exploring on the technology and Broadband comes into the picture. So to see the development of specific area, we can consider one of the major parameter as the Penetration of Broadband network.

Here, as asked, the penetration of domestic broadband networking in Ireland is to be investigated. The most important requirement for the analysis, i.e. Data, is captured from the Central Statistics Office Website for year 2016. For better visualization, the spatial study in terms of decorating on maps is preferred using QGIS software. The data first processed using R then put into database PostGreSQL and later represented in terms of maps for the analysis.

The Flowchart about how the workflow occurred is as shown,

In MAPs using QGIS

Processing on data using R

Store in PostGreSQL Database

Data Analysis using MAPs

The implementation of each block is explained briefly in the following sections:

**Getting the data:**

The analysis is mainly based on the data, and hence the valid and trusted data is very important here. For the same, we can choose to have the census 2016 Small Area Population Statistics (SAPS) data from the Central Statistics Office Website. The boundary data in the form of Shapefile as well as the statistical measures in terms of CSV file are downloaded. Apart from Shapefile, there are various options for boundary data as CSV, GeoJSON and KML, etc., but the Shapefile looks more convenient to use here. The boundary data is available in ungeneralised and generalised options. The generalised has again 20m, 50m and 100m accuracies available. Here, the generalised 20m file is chosen, as the ungeneralised data has large size and is used for detailed analysis like point in polygon analysis, etc., which is not required here. Also the generalised has much of original data preserved and though has 10% of file size of ungeneralised file and hence, the 20m seems sufficient, which again gives the good combination for details but small size file.  As far as the statistical observations are concerned, the CSV file is considered which contains large amount of observations managed in the form of Themes and Tables. There are 15 Themes further divided into 804 tables. Theme 15 tables 2 and 3 have the statistics related to the availability of personal computers and broadband. We get the numbers of houses having broadband and total number of houses per county.

**Manipulating the data:**

The first two blocks in the Flowchart shown above describe this section. The Shapefile data taken can be used as it is in the QGIS for getting the boundaries of the counties and hence no separate processing is required as QGIS does the transformation of SRID task for us automatically. The statistical data in CSV file has 804 columns and hence need to be carefully selected the columns of our interest. The Glossary file has been provided on the same website which tells about how the distribution of data is done in themes and tables. As per the Glossary file, the column T15\_3\_B shows the count for number of houses with broadband connection, and the column T15\_3\_T gives the total number of houses count. As we can see the total population is different county wise, and hence, we need to consider the percentage of numbers i.e. percentage of houses with broadband. This can be achieved using some basic arithmetic operations. There are many options available to do the same, some of them we can use are, using MS Excel, R, Python, and Java or using PostGreSQL. Here, R is used for the operation, because it can be used for further statistical operations if needed. As the CSV data is imported into R, we get to know that the numbers are represented in the form of Digit Grouping using comma for better reading of numbers, which violates the rules of software programming by considering the data not as numbers. This can be solved using the powerful command of R and hence the data is now numeric and can easily processed using arithmetic operations. The percentage of household with broadband is calculated and ready to use for analysis. The best way to store the data is in database, and hence PostGreSQL database is used here for storing the data. There is no inbuilt support available for the connection to database using R and hence the package “RPostgreSQL” for database connection is installed and the library is used for it. The SQL queries for database are fired through R for which again, there are different commands for the type of Query to be fired. The SQL is case insensitive language but R is case sensitive, so doing SQL using R put us into the confusion for some cases, then with some exploration and trial and error, the answer got it as we need to go ahead by considering there is lower case in the SQL. The connection to the database is done and hence the data is imported into the PostGreSQL database and ready to use for analysis.

**Creating the maps:**

The most important section of data analysis is data visualization. There are various ways of doing the same. For our case, the spatial database is considered, for which the convenient tool is QGIS. Hence the data from database is imported into QGIS. There are various ways to import data in QGIS, like, importing CSV, import data from MS Excel, data with custom delimiters, etc. Here the processed data from R is stored in PostGreSQL database which is further extended to QGIS. Some additional things required to be done include, adding PostGIS layer into the QGIS to access the database through DB Manager. Now we get two different datasets available into the QGIS as Shapefile and the processed statistical data from Database. The most important thing to do is to make the data synchronized, which can be done using the Table join. We can join the tables using Join facility provided by QGIS and hence we get the data available area wise, using the same value of GUID in both tables.

After joining the tables, the actual drawing of maps comes into picture, which can be done using the style option in the property of joined table on canvas. As the data is continuous and not categorical, we can use the graduated option and then desired color scheme for the classification of data depending on various types of intervals available in the software. There are 5 different types of intervals available in the software. We can refer the histogram for the classified values and depending on the distribution of data on histogram; we can set the intervals among Equal interval, Quintile, Natural breaks, Standard deviation or pretty breaks. By looking at various types of intervals available, and analyzing the histogram, the Pretty Breaks type of interval classification shows good distribution of values in terms of histogram. The pretty breaks classify the data into the discrete levels which are easy and convenient to do the analysis. The levels of 10s make the analysis easier as the statistics is in terms of the percentages.

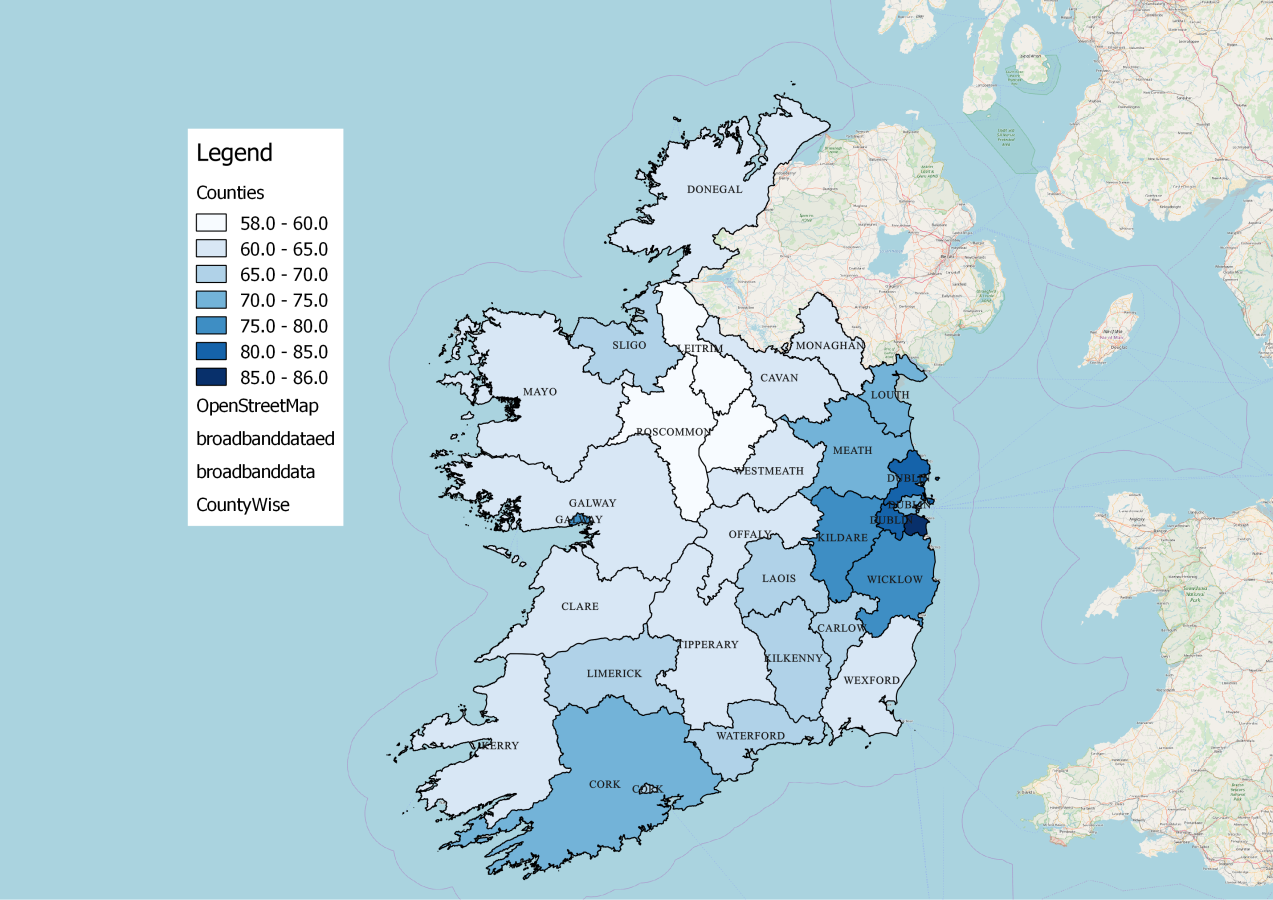
Another most important aspect for creating the map is the color of the map, which color scheme to use is decided by the type of map. Here the thematic map is shown and hence the gradient color scheme is used which best describes the concentration of networking. By considering the color blindness problem and all other problems, the blue color gradient scheme is been chosen here.

The final time comes now to display the map with respect to some background mapping. The Shapefile on the canvas shows the distribution of the broadband network across Ireland, County wise as well as Electoral Division wise. But for some better visualization, we can put some background map with increasing some transparency of our map. For background mapping, various options are available, like openstreetmap, google map, bing map, etc. The openstreet map chosen here. As the Shapefile representation of map found to be not fitted properly with the background map, we need to transform the boundary data to Irish National Grid with ESPG 2157 standard. QGIS provides this facility and transforms the data automatically into the suitable format. Therefore the map is ready for the analysis now, which can be saves as Image and use for analysis.

**Discussions:**

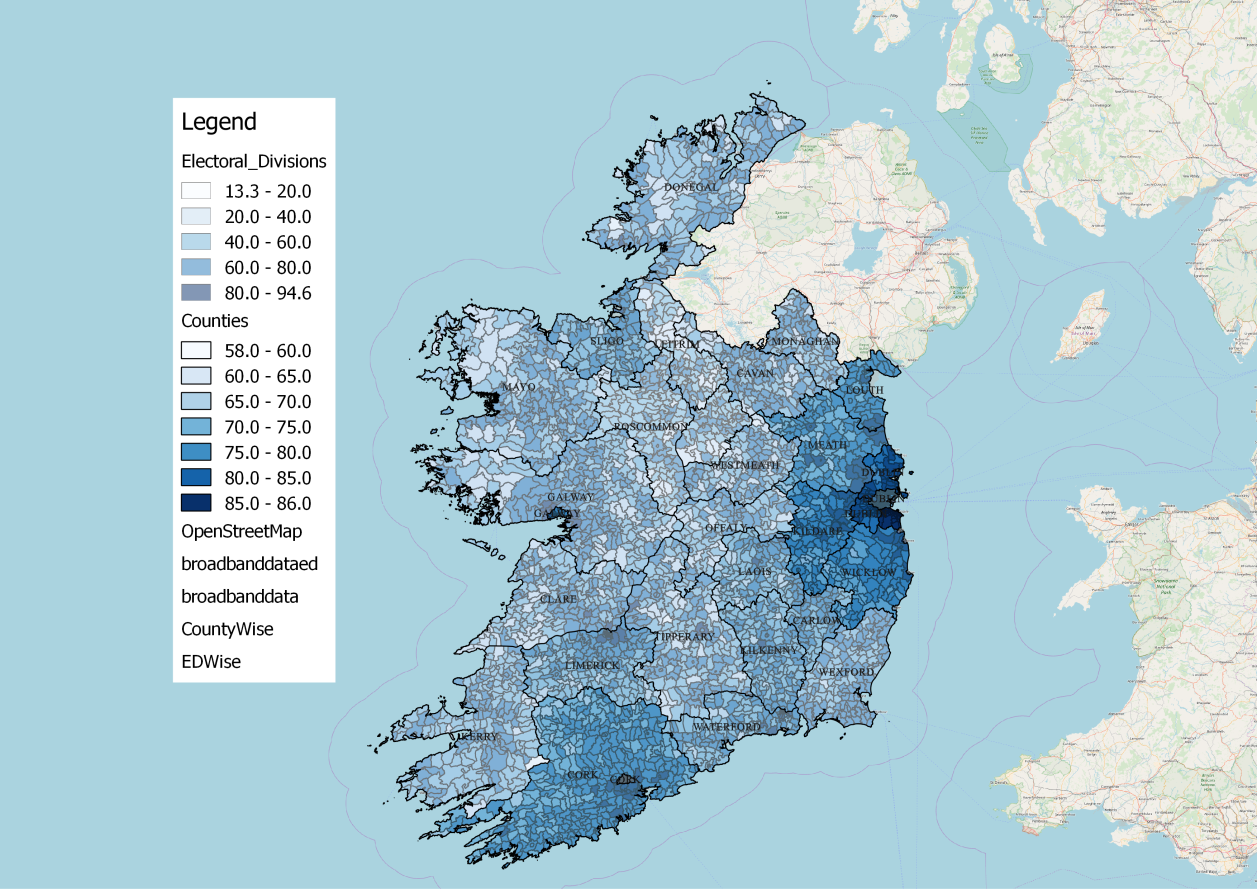
The discussion can be done in mainly two parts. One at County level and by zooming in, another can be at Electoral Division level.

1. **County Level**: As per the given dataset, Ireland is divided into 31 counties. The analysis has been done for the percentage of broadband usage in each county. The pattern showing distribution is as per below image.



As per the map shown above, the counties of Ireland are divided into 7 categories, each category showing some intensity of the blue color. Higher the intensity of the color more is the usage of broadband in the county. As per the data, the usage of broadband in each county is above 58%, which shows the digital development of the country. Around 30% of counties having broadband usage more than average. We can clearly say that the eastern part of counties have very good broadband usage than rest of the area. Higher broadband usage can be seen in Dublin, Kildare, Meath , Louth, Wicklow, Cork and Galway city. The more usage of Broadband in those counties shows the digital improvements and technology usage in the area. Least usage of Ireland can be observed in Leitrim and Roscommon.

1. **Electoral Division level**: By going into details of each county, we can have look at the statistics at electoral division level.



The discussion has been moved into more details i.e. Electoral division wise. The counties are further divided into 3409 EDs across the country. On top of county distribution, the EDs are distributed and classified according to the blue color intensity. More the intensity of the blue color more is the broadband usage. The usage of broadband in EDs is as low as 13.3% and as high as 94.6% which shows that it has the detailed county wise distribution. With reference to county wise map, we can see that, as higher broadband usage EDs make the counties to have more broadband usage and vice versa. So the pattern “darker becomes darker and lighter becomes lighter” can be observed.

Finally, to conclude, we can say that as the percentage of broadband usage in Electoral Divisions decide the percentage of broadband usage in counties. And hence the counties having more broadband usage EDs are seen to be having more broadband usage across the country. This can be supported with the help of maps as shown in above section. The more number of darker color of electoral divisions results into darker county and vice versa for opposite.

In this way, using the mathematical part of data manipulation using R, storing data in PostGreSQL database and visualization using QGIS Maps, the penetration of Broadband Network across Ireland has been studied.