****

**by**

**Pratik Budhdeo**

**Introduction:**

Voting is a fundamental right of citizens and it enables them to choose the future leaders. Higher percentage of people voting depicts that majority people participate in deciding on measures like security, development and advancement of the country.

**Trends in Dublin over the two decades (1970-90) showed an increase in level of voter abstention**. Over a span of general elections from 1969 to 1981, turnout was at 76.5% in average. However, since 1981, all the movement has been declining, it dipped to 65.9 per cent in the 1997 general election.

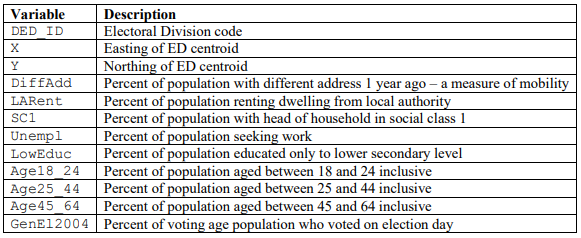
It is ambitious to precisely identify trends in turnout, but there is a developing demand for knowledge which would explain the cause of this phenomena. The aim is to find the reasons that influence the variation in voter turnout most strongly using regression.

**Specification:**

For analysis, different regression techniques are **implemented using R** to find the best set of predictors to predict the Turnout (Response) and **geographic variability** within them.

**DubVoter Dataset:**

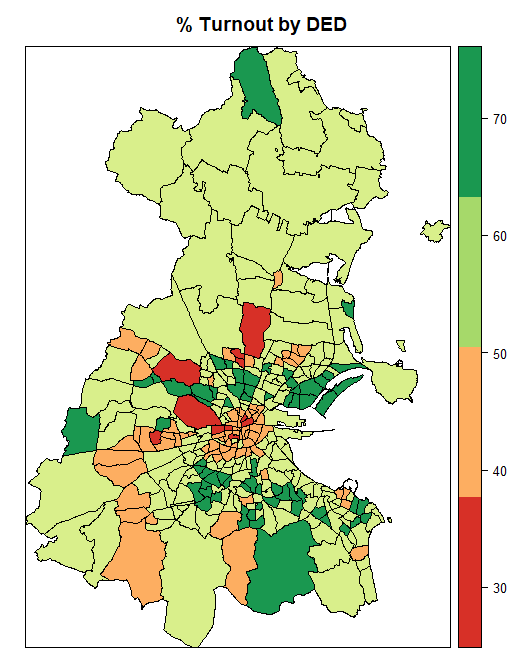
The DubVoter (within ***GWmodel***) dataset is used having information about t**urnout in 2002 general election along with socio-economic characteristics** for the ***322 Electoral Division in Dublin*** city. A unique ID is assigned for each ED details of predictors are given below: -





The above 11 independent variables reflect the predictors that might influence the response as dependent variable GenEl2004.

**Voter Turnout by DED**



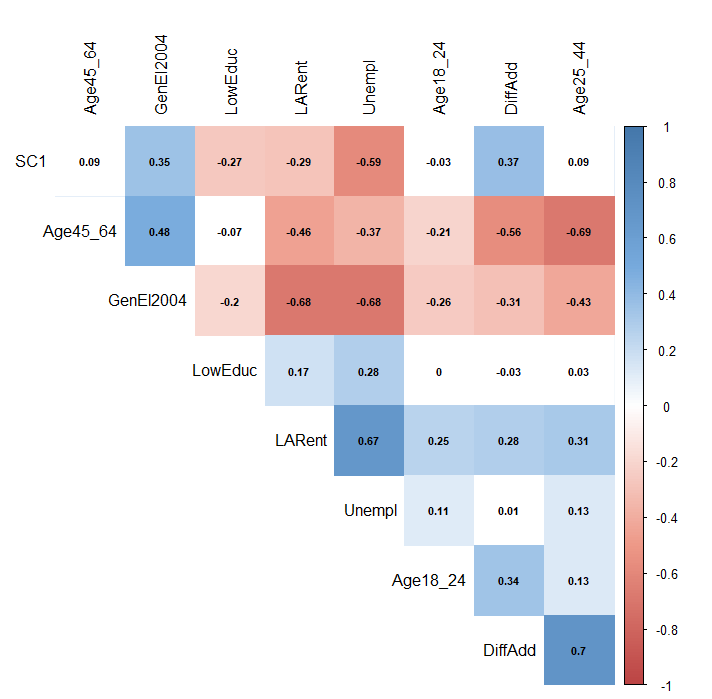


Above map depicts percentage of voter turnout in each Dublin ED. There is significant **variation in turnout and variation appears to be spatially autocorrelated**. In **most of the region only 60 % of the people** have voted and 70% in few ED’s. There are ***11 places with least percentage (30%)*** turnout.

**Collinearity**

As a first step it is important to identify the linear relationship between the predictors so that we can prevent if there is any high influence from those variables on the result. Best way is to **minimize the collinearity** between the variables to reduce the influence.

* **Correlation:**

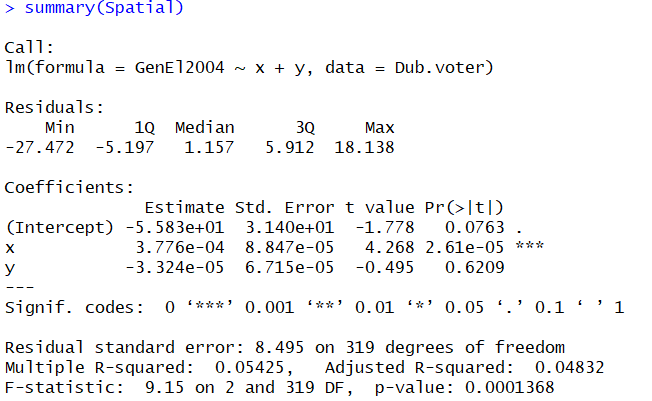


Response GenEl2004 has **high correlation** with all the predictors and some of the predictors themselves have high correlation. The pairwise predictors having correlation greater than 0.5 are: -

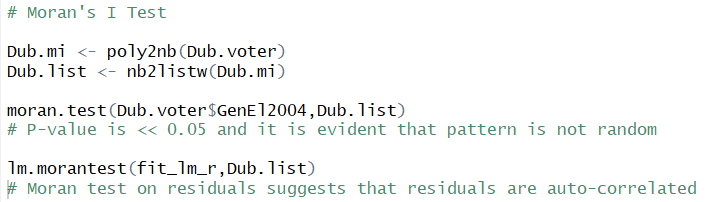
* Age25-44, Age45\_64 with DiffAdd
* Unempl and LARent
* Unempl and SC1
* Age45\_64 and Age25\_44

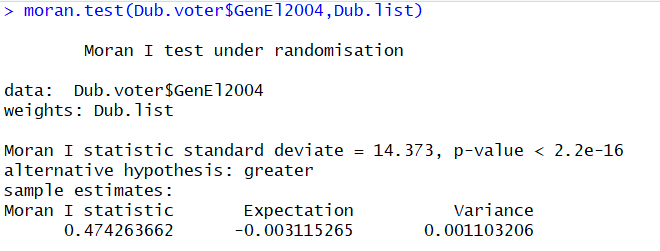


**Spatial Pattern**



It is evident that **X (Easting) is significant in prediction** as a spatial component, and hence we can compute **Moran’s I test** to test autocorrelation and spatial pattern in dataset.

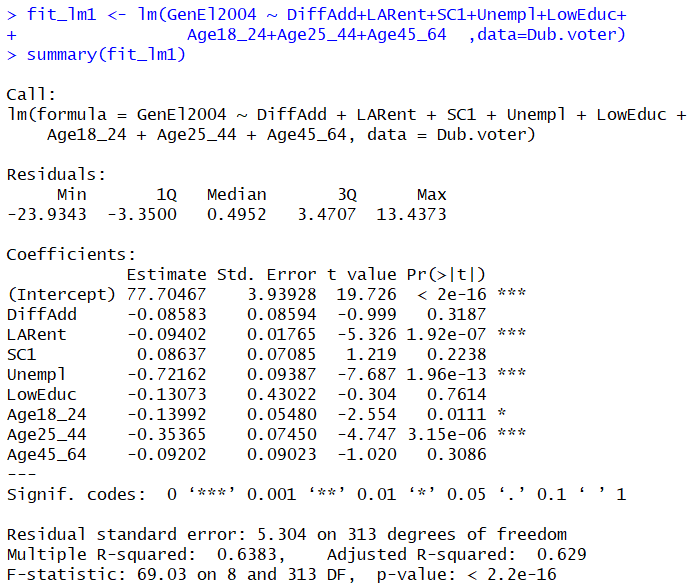




We interpret that there is **significant autocorrelation** and spatial pattern in voter turnout.

**Linear Regression**

Linear regression helps to model the relationship between a dependent variable and explanatory variables (or independent variables). Two variables DED\_ID and X are ID and coordinates respectively, so these variables are not considered for analysis.

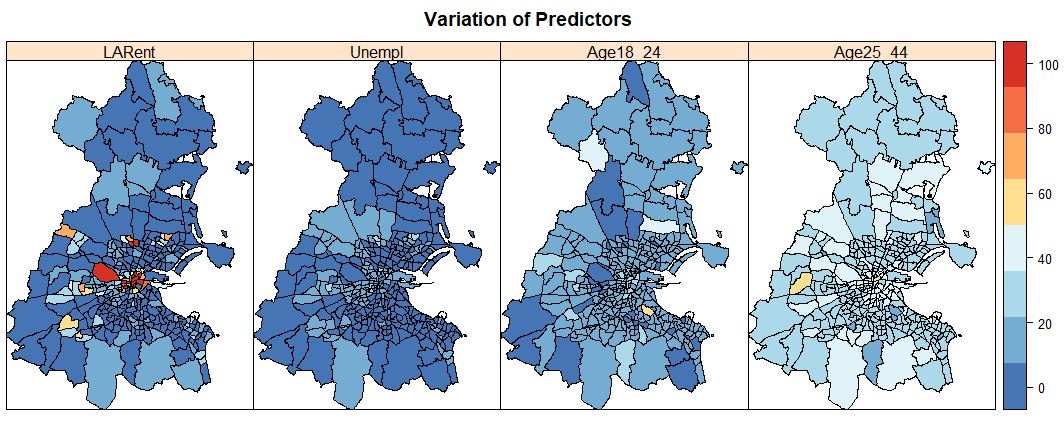


Analysing the results based on P-Value (P < 0.05), ***significant predictors are LARent, Unempl, Age18\_24, Age25\_44.***

For all the significant predictors, **coefficients are negative**, i.e. percentage of population who voted decrease with increase in the percentage of the predictors.

**Visualising Coefficients**

Significant Predictors from Linear Regression (Reduced model)





From this Map, it is clear that there is significant spatial variation in predictors, especially in LARent.

**Principle Component Analysis**

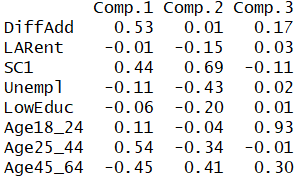
**PCA** is used for analysis of multivariate data. It commonly explains the covariance structure of a multi-dimensional data space in a low-dimensional alternative. The components (Principal) are linear aggregations of the original predictors that provide a better understanding of differing sources of variation and structure in the interpretation of information.

All eight predictors are not of a similar magnitude although measured on a similar scale. Thus, we have standardized the data to make each variable have same relevance in the data analysis.

**PCA**

 **Robust PCA**

Using the PTV (percentage of total variance) values, the 3 components (Comp.1, Comp.2 and Comp.3) collectively **represent** **74% (PCA) and 87% (Robust PCA) of variation** in the data. As Robust PCA performs better, we will take a look at the tables of loadings for the same,



Comp.1 is an overall measure of the variance - most heavily influenced by SC1 and DiffAdd and Age25/45.

Comp.2 gives more weight to SC1, Unempl and Age45\_64.

Comp.3 is analysing the percentage vote by younger people Age18\_24.

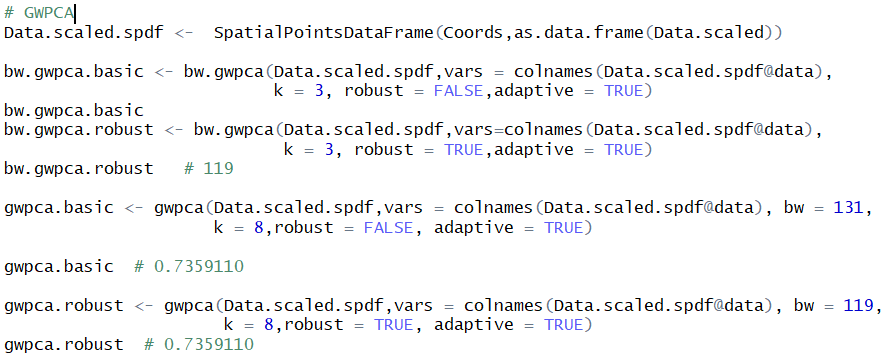
These whole-map statistics give overall average over Dublin region, but do not truly represent local social structure.

**Geographically Weighted** **PCA** analyses spatial heterogeneity within the multivariate data. GW PCA accounts:

* How the original predictors influence individual spatially-varying component,
* How data dimensionality varies spatially

Under GWPCA, a `leave-one-out' cross-validation score is calculated for complete possible list of Kernel bandwidths and an optimal bandwidth selected for smallest CV score. We have to decide ‘k’ number of components to retain.

**Robust Geographically Weighted** **PCA** This method helps to trim the effect of abnormal inputs on the output. Each individual local covariance matrix is approximated using the robust Minimum Covariance Determinant (MCD) measure, that examine h data points that have the smallest determinant for their basic covariance matrix. Default value of h = 0:75n.



On comparing results from both the methods, **optimal bandwidths of N = 131 and N = 119** selected to calibrate the basic and robust GW PCA fits (Using k=3) respectively. Later, we specify all components (k = 8), but will target first 3 components. This guarantees that we correctly estimate the variation locally represented for by each component.

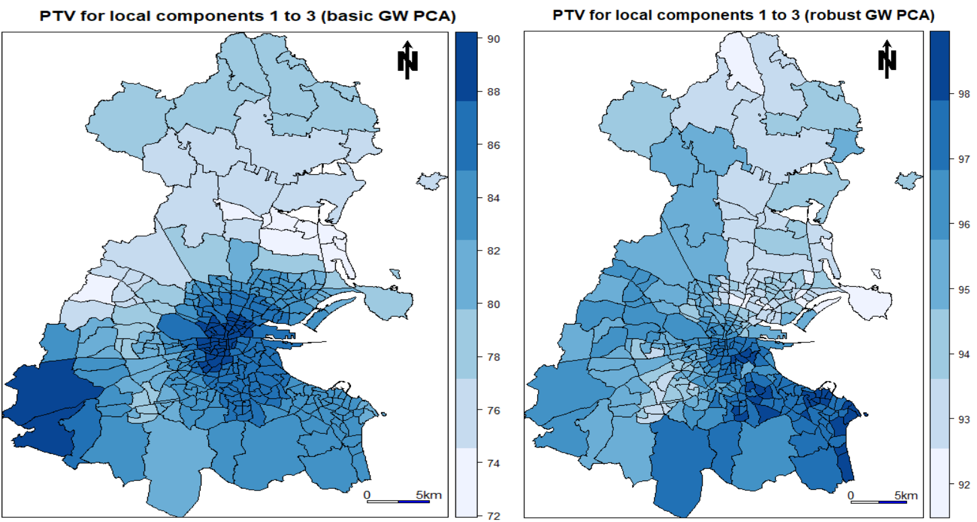
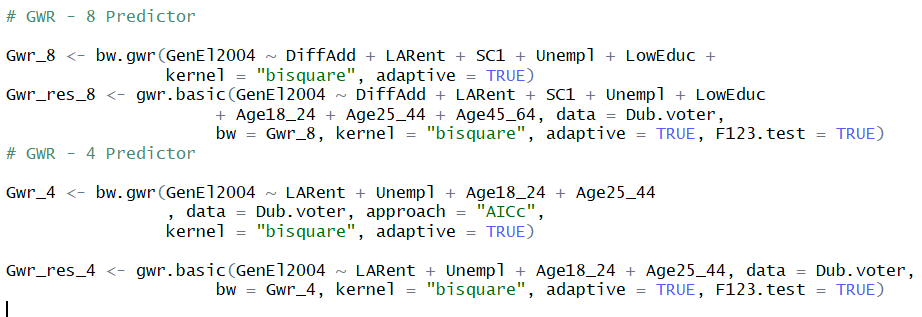


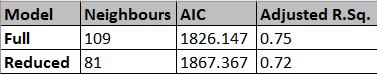
  

Figure 5 represents the local maps (PTV) for the 2 GWPCA for the first three local components. There is clear land variety in the PTV information and a higher PTV is for the most part represented in the neighborhood case, than in the global case. The spatial pattern in the two maps are comprehensively comparative, with **higher rates situated in the south, while lower rates are situated in the north**. As would be normal, the robust PTV data is reliably higher than the basic PTV data. Substantial contrasts between the basic and robust PTV (e.g., in south-west Dublin) can be taken to demonstrate the presence of neighborhood multivariate exceptions.

**Geographically Weighted Regression**

In GWR spatially changing relationships are investigated between the independent and dependent factors. The shape and size of the BW is subject to input for the Kernel type, technique, distance, and Number of neighbor's parameters.

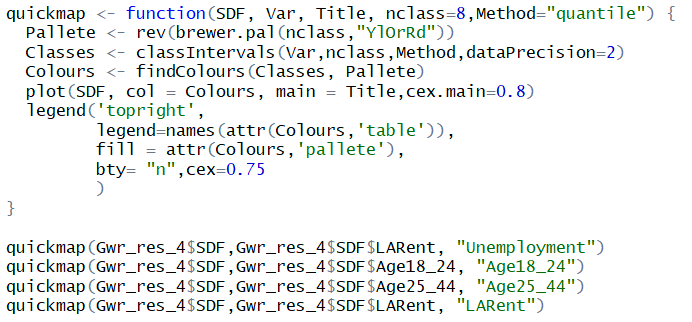


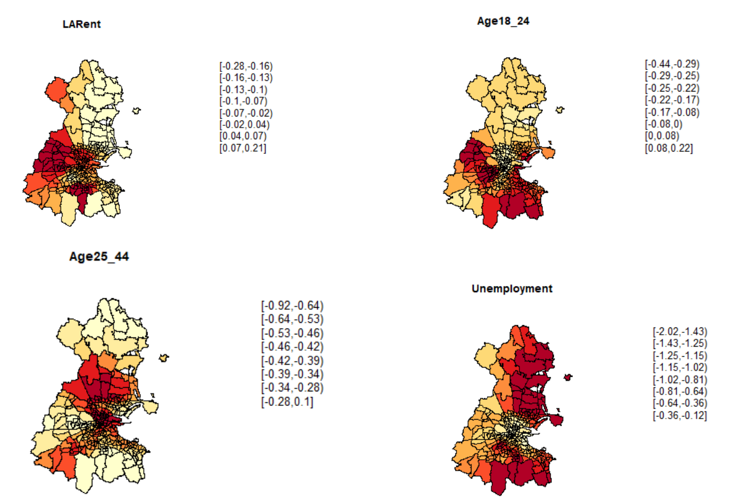


Based on AIC & Adjusted R-Sq. value, we can say that **GWR reduced model provides close fit to data**.

**Visualising Coefficients**

Significant Predictors from GW Regression (Reduced model)





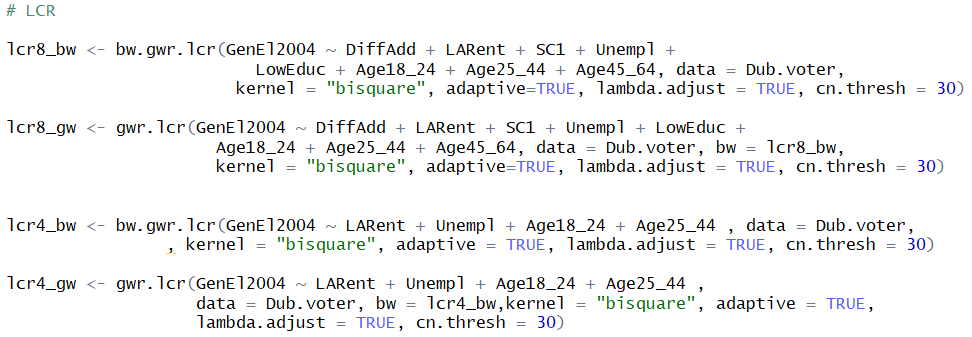


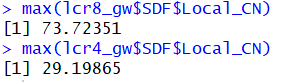
If these maps are compared with the maps of reduced model before performing GWR model (Fig 4), we can notice a significant change in autocorrelation and spatial pattern as explained below: -

* Local authority renters located in NE and SE parts of Dublin are majority of the population of voters. This decreases in western and central parts of Dublin.
* Young voters (18-24) in have contrast trend for north and south Dublin ED’s.
* Middle-Aged voters (25-44) are predominant in the ED’s close to the south and north of central Dublin, negative influence at all the remaining places.
* Overall Unemployment rate is less, but with increase in Unemployment the voter turnout decreases in general specially in the central region.

**Collinearity**

Locally Compensated GW Ridge Regression is used to deal with the effect of collinearity (leads to loss in precision and power in the parameter estimates). This can be accomplished by altering the lambda value (lambda.adjust = TRUE) in Ridge regression and by setting the threshold value of condition number to 30 (cn.thresh = 30).





The Condition Number is a measure of the extent to which a cross-product matrix is ill-conditioned. A matrix exhibits worrying levels of collinearity if its k is greater than 30. As shown above, the condition number of 73.72 in 8 predictor model has been brought down to 29.2 in 4 predictor model.

**Conclusion:**

In 2002 Dublin general elections, we found that **LARent, Unempl, Age18\_24 and Age25\_44 are the most important predictors** in understanding the variability in the voter turnout assessment. **SC1, DiffAdd, LowEduc and Age45\_64 are not significant** in the model predictions. It was evident as the variables were suggested by a geographer who was not an expert in statistical modelling that there can be strong collinearity present among predictors under analysis. The main of the study was to decrease the collinearity which hen helped to better understand the actual geographical local variation among the ED’s in Dublin. We created various visual maps to understand the pattern and localisation effect.

We found that **GWR and GWPCA are better than Classical Regression and PCA** as they help to reduce the residual by taking into consideration the spatial relation between the different predictors. We were successful in **reducing the collinearity among predictors and from 73.73 to 29.2 by reducing the model with 8 predictors to 4 predictors**, which is in line with the results from classical regression.