Name: Paula McMahon

Student ID: 17185602

Module: NCG603 / NCG613

Title: Modelling Voter Turnout in Dublin

Date: 25th April 2018

# Introduction

The dub.voter dataset is one which explores spatial variation in voter turnout in the 2002 General Election. It is a built-in dataset to the GWModel R library and it contains 322 observations. There are eight predictors: *DiffAdd*, *LARent*, *SC1*, *Unempl*, *LowEduc*, *Age18_24*, *Age25_44*, *Age45_64* and the response variable for voter turnout is *GenEl2004*. The objective of this assignment is to see which variables influence the variation in voter turnout. The sections that follow will outline the approach taken, the results and a short conclusion. An R markdown document with code, comments and plots accompanies this document by way of reference.

# The collinearity issue

A problem in regression modelling is that of collinearity. Collinearity is a problem particularly in spatial datasets where, in heterogenous data, some locations may exhibit collinearity while others may not. In tackling this assignment, this issue must be borne in mind when choosing an optimum model to predict the voter turnout.

# Method

## Check the global correlation of the variables

Use the *pairs* function in R to produce a scatterplot matrix of all predictors and the response.

## Use geographically weighted summary statistics

Use the function, *gwss*, as a pre-cursor to running GW Principal Components Analysis later. The summary statistics comprise of:

(i) GW standard deviations that will highlight areas of high variability for a given variable.

(ii) GW correlations that provide an assessment of local collinearity between two independent variables of a GW regression.

## Run non-GW PCA

Scale the data first, run PCA using the *princomp* function, view the loadings and generate a scree plot which gives the proportion of variance per component.

## Run GW PCA

Convert the scaled data to a spatial data frame. Select a bandwidth using the *bw.gwpca* function. The parameter k, the number of components to retain, is set to 3. The bandwidth is the optimal cross validation (CV) score. GW PCA is run with the *gwpca* function using the chosen bandwidth.

## Run GW Regression

Again, select the bandwidth using approach=AICc and a bisquare kernel (bisquare means it uses a weighted moving window). Run GW regression with the *gwr.basic* function and using all predictors. Then, calculate some local correlations, local variance inflation factors (VIFs) and local condition numbers (CNs). (VIFs measure the degree to which the variance of an individual predictor is increased by the presence of collinearity with the other predictors and CNs measure the extent to which a cross product matrix (used to estimate coefficients) is ill-conditioned). The results section below will demonstrate that we have found evidence of local collinearity at this point. Let's now use Locally Compensated GW Ridge Regression.

### Run LCR GW Regression

With LCR-GWR, local ridge parameters are used, and they are only used at locations as set by the condition number threshold. Ridges of zero are specified at other locations. Therefore, not all the local regressions of a LCR-GWR model are necessarily biased. To run the regression, create a parameter for bandwidth using the *bw.gwr.lcr* function. Use a bisquare kernel and set a CN threshold of 30. Run the regression using the bandwidth with the *gwr.lcr* function. Re-check the VIFs and CNs. If they are not at the required values, we may need to reduce the model.

### Model reduction

The approach used was to firstly reduce the model by one term and then check the local VIFs and CNs. If the condition numbers have not decreased sufficiently, consider removing a second term.

## Results

Global Correlations: Strong correlations are seen between the following pairs of predictors: *DiffAdd* and *Age25_44* (0.7), *Age45_64* and *Age25_44* (-0.69), *LARent* and *Unempl* (0.67), *LARent* and *GenEl2004* (-0.68), *Age45_64* and *GenEl2004* (0.48)

Summary statistics: Fig 1 shows variability for the response variable *GenEl2004*. The areas of highest variability in turnout are North and West Dublin.

Local correlations: Fig 2 shows the degree of negative correlation between turnout and those living in local authority housing. The lightest blue areas in West Dublin and also some areas of Dublin City have the highest correlation. Fig 3 shows that variables *LARent* and *Unempl* are positively correlated so this map shows the area of highest correlation in South West, South East and North of Dublin city.

Basic PCA analysis: 73.6% of the variance is explained by the first three components (the scree plot displays this). The first component would appear to represent an older demographic (*Age45_64*). The second component, appears to represent affluent residents (*SC1*). The third component is mostly explained by the younger population (*Age18_24*).

GW PCA analysis: The output shows how data dimensionality varies spatially and how the original variables influence the components. Fig 4 show the percentage variance in the first three components. Fig 5 shows how the loadings from the first GW'd component correspond spatially to the electoral division codes. There is clear geographical variation in the influence of each variable on the first component. For GW PCA, *LowEduc* dominates in the northern and southwestern EDs, whilst *LARent* dominates in the EDs of central Dublin. *Unempl* dominates north of the city. *Age45_64* dominates in South suburban Dublin.

GW Regression: The model summary shows that the significant predictors are *LARent*, *Unempl* and *Age25_44*. Also, we have found evidence of local collinearity: Fig 6 shows local correlation is high between *DiffAdd* and *Age25_44* (more than 0.9), the VIFs are as high as 13 in some inner city EDs (see Fig 7) and CNs are as high as 110 in extreme North Dublin (see Fig 8). Furthermore, the correlation of *GenEl2004* with *Age45_64* is positive, but the sign of the GW regression coefficient is negative. Lastly, only four out of eight of the GW regression predictors are significant. Unexpected sign changes and relatively few significant variables are both indications of collinearity.

LCR GW Regression: Fig 9 shows the local ridge terms (λ) used in the LCR GW regression. The area with the largest ridge parameter is in North and South West Dublin EDs. After the regression has run,

we can see that most of the electoral areas have a VIF of less than 10 for the variable *DiffAdd* (Fig 10). Fig 11 shows that the CNs have reduced from a max of 110 to a max of 75 so there has been a marked improvement.

Model Reduction: Removing variables individually has little effect on the local condition number distributions. After removing *Age45_64* by itself, local CNs reduced from 75 to 55. This number suggests collinearity is still a problem. Removing the most collinear variables tends to provide lower condition number correlations. So, an LCR GW regression was run with *Age45_64* and *Age25_44* removed. Recall that the correlation coefficient of these predictors is -0.69. After removing *Age45_64* and *Age25_44*, VIFs are well below 10 (Fig 12) and CNs are a maximum of 35 (Fig 13).

## Conclusions

While *Age45_64* and *Age25_44* were strongly correlated and removing both from the model had the desired effect of reducing the collinearity effects, *Age25_44* is a significant predictor in the model according to GW basic regression with a p-value of $3.15e^{-06}$. So, we need to ask the question – what effect will its removal have on the goodness of fit of the model to the data?

- From the basic GW Regression with eight predictors, the Residual Sum of Squares value is 8805 and the AIC is 1999.
- From the basic GW Regression results (with *Age45_64* and *Age25_44* removed), the Residual Sum of Squares value is 9490 and the AIC is 2019.

In conclusion: the full eight predictor model is a better fit to the data but the reduced six predictor model (minus *Age45_64* and *Age25_44*) is better for reducing or eliminating the effects of collinearity.

## References

Isabella Gollini, Binbin Lu, Martin Charlton, Christopher Brunsdon, Paul Harris (2015). GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. Journal of Statistical Software, 63(17), 1-50. URL http://www.jstatsoft.org/v63/i17/.

Belsey D A, Kuh E, Welsch R E. Regression Diagnostics Identifying Influential Data and Sources of Collinearity: http://bayanbox.ir/view/4656937036272759497/Regression-Diagnostics-Identifying-Influential-Data-and-Sources-of-Collinearity.pdf

Charlton M: Living with collinearity, Lecture Notes, NCG, NUI Maynooth.