Introduction

In this study, data collected on the 2002 General Election in Dublin is examined to determine the best pridictors for a model whose response variable is the voting turnout percentage of the population in Electorial Divisions. Due to the nature of the data's spatial distribution using models with global form predictors does not describe the data very well as their relative geographic locations to each other is not taken into account. As such, local statistical models called Geographically Weighted (GW) models are used instead when dealing with data with spatial heterogeneity. To that end, an R package called GWmodel and it's fuctions on Principle Component Analysis (GWPCA), Regression (GWR), and Collinearity are used to examine the data. These models work by using a moving window weighting technique where the window size is controlled by the bandwidth, an optimally adaptive bandwidth can be determined by cross validation with a specific kernal by using the bw.gwr() function. This kernal is what determines the weigthings at each location, starting at the window centre the kernal weights decay as the distance out increases until a set distance is reached or a specific number of nearest neighbours is attained. With these geographical weights taken into account, the data can be expressed more locally than globally, and provide a better discription of the voting habits for specific groups.

Task

To detremine the best predictor variables for a model with the response variable being the voter turnout percentage of the population in Electorial Divisions (EDs) of Dublin by addressing collinearity and creating geographically weighted models with Principle Component Analysis and Regression.

Approach

Th first step is to check for collinearity in the data, that is when independent variables are highly correlated. This collinearity causes problems by inflation of the variance and loss of procision, these problems are made even more prominant in GW due to the effect of being at smaller, more local samples, and due to spatial heterogenity resulting in differing collinearties at different locations. By plotting and checking the correlation matrix of the data using plot() and cor() fuctions, variables with potential collinearity problems can be determined. Once pairs of variables with high correlations are determined their Variance Inflation Factors (VIFs) can be determiined with the gwr.collin.diagno() function. Plotting the correlations and VIFs of the variables by the EDs of Dublin to examine the distribution of the variance inflations can help determine which, if any, variables need to be removed from the model to alleviate the collinearity problem.

Once the collinearty problems have been resolved, the remaining variable can used to contruct a model using PCA. PCA converts variables into linearly uncorrelated variables called principal components, these components then describe the variability of the data. For GW PCA local components are used to describe the data within a set distance or number of nearest neighbours. The PCA and GW PCA models are created using the gwpca() function and an optimally adaptive bandwidth is determined using the bw.gwr() function. Further GW Basic and Mixed Regression models were also made using the gwr.basic() and gwr.mixed() functions respectively. The difference between the basic and mixed models is that the basic GWR model treats all of the variables as local variables, while mixed only treats some of the variables as local and treats others as global. Both the basic and mixed GWR determine their bandwidths the same way as the previous models. To determine which variables are global and which are local, a Monte Carlo significance test is performed where the null hypothesis, H_0, is that the relationship between the response variable and a specific predictor variable is constant. This is performed using the montecarlo.gwr() function. If the p-value for a specific predictor variable is less than the 5% significance level, then the null hypothesis is rejected and the variable is considered local. However, if the p-value is greater than the significance level, then the null hypothesis is accepted, thus if the relationship between the response variable and the predictor variable is constant, the variable can be considered global.

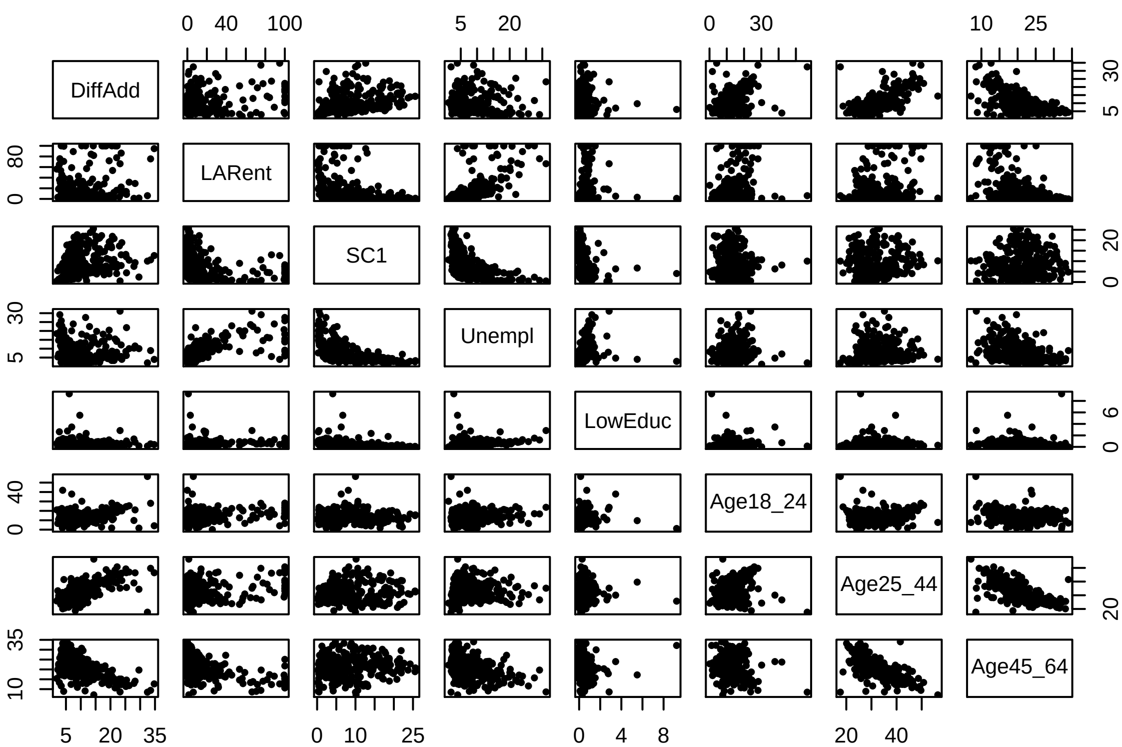
Data

The data itself is a spatial polygons data frame with a set of polygons for the EDs, a set of variables describing the proportion of each ED populations in respect to some social groupings, and the total proportion of voter turnout. There is also a unique ID vector and X and Y coordinates. All data is in relation to the 322 EDs of Dublin during the 2002 Dail elections. The variables are as follows:

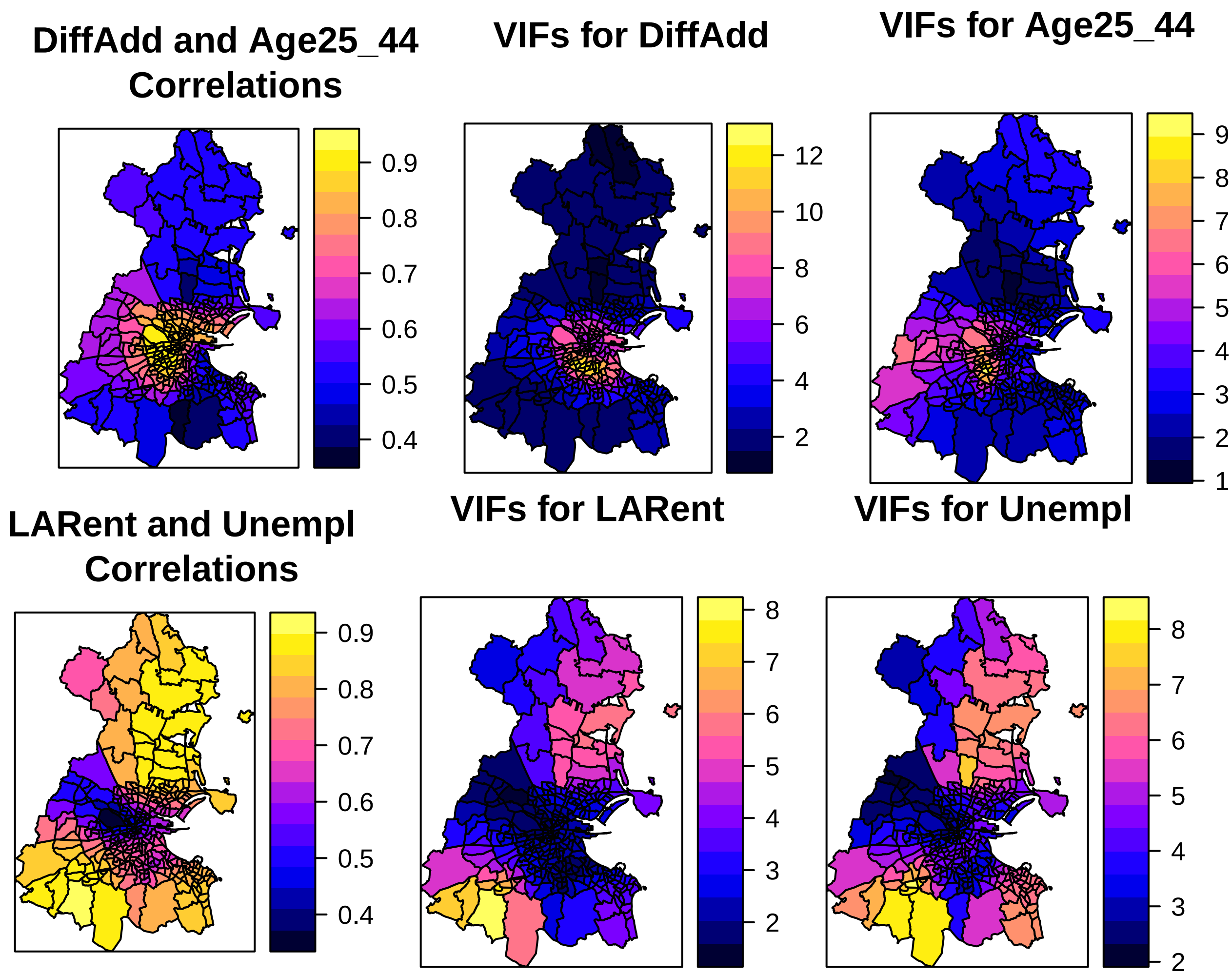
- DED_ID - Vector of unique IDs
- X - X coordinates
- Y - Y coordinates
- DiffAdd - Percentage of population who are one year migrants
- LARent - Percentage of population who are local authority renters
- SC1 - Percentage of population who are social class one
- Unempl - Percentage of population who are unemployed
- LowEduc - Percentage of population who have little formal education
- Age18_24 - Percentage of population who are within the age group 18-24
- Age25_44 - Percentage of population who are within the age group 25-44
- Age45_64 - Percentage of population who are within the age group 45-64
- GenEl2004 - Percentage of population who voted in 2004 election

Analysis

Collinearity Diagnostics



```
##                DiffAdd     LARent         SC1       Unempl       LowEduc
## DiffAdd   1.000000000  0.2757630  0.37229879  0.005779965 -0.0318577039
## LARent    0.275763007  1.0000000 -0.29226328  0.668776169  0.1675897500
## SC1       0.372298795 -0.2922633  1.00000000 -0.591567897 -0.2727821424
## Unempl    0.005779965  0.6687762 -0.59156790  1.000000000  0.2828183109
## LowEduc  -0.031857704  0.1675897 -0.27278214  0.282818311  1.0000000000
## Age18_24  0.335300415  0.2524328 -0.03295145  0.113380829 -0.0001270398
## Age25_44  0.703062428  0.3124497  0.09075550  0.131741285  0.0283071619
## Age45_64 -0.561283893 -0.4626929  0.08930325 -0.374268612 -0.0723867788
##                Age18_24    Age25_44    Age45_64
## DiffAdd   0.3353004149  0.70306243 -0.56128389
## LARent    0.2524327691  0.31244970 -0.46269291
## SC1      -0.0329514460  0.09075550  0.08930325
## Unempl    0.1133808290  0.13174128 -0.37426861
## LowEduc  -0.0001270398  0.02830716 -0.07238678
## Age18_24  1.0000000000  0.12619751 -0.21154554
## Age25_44  0.1261975141  1.00000000 -0.69322998
## Age45_64 -0.2115455410 -0.69322998  1.00000000
```

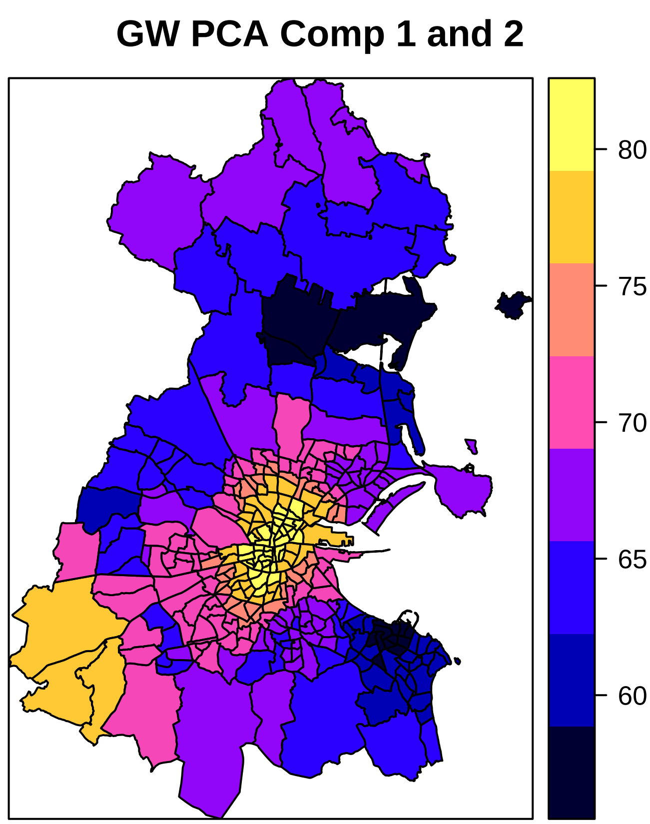


Plotting the data and examining the correlation matrix shows that DiffAdd and the Age25_44 is strongly correlated with a value of ≈ 0.7, and that the variable LARent and Unempl also strongly correlated with a value of ≈ 0.67. Now that the variable pairs with stong correlations are known there collinearity can be examined. As seen in the graphs above significant collinearity can be seen in central Dublin between DiffAdd and Age24_44, with the VIFs for DiffAdd being more densly packed than the VIFs for the Age25_44 group. The graphs for the corelation between LARent and Unempl are consistently strong across Dublin, with the VIFs for both groups also being high and spread out. Since collinearity will cause problems once the geographical weighting becomes a factor, to alleviate the problem one the variables should be removed, examining the AIC values shows that only removing DiffAdd makes any real change to the data, as removing any of the other variables has little effect.

Geographically Weighted Principle Component Analysis (GWPCA)

```
##    ***********************************************************************
##    *                       Package   GWmodel                             *
##    ***********************************************************************
##    Program starts at: 2019-05-08 00:36:21
##    Call:
##
##    Variables concerned:  LARent SC1 Unempl LowEduc Age18_24 Age25_44 Age45_64
##    The number of retained components:  7
##    Number of data points: 322
##    ***********************************************************************
##    *                Results of Principal Components Analysis               *
##    ***********************************************************************
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.6240067 1.2442647 0.9765011 0.9035621 0.73630994
## Proportion of Variance 0.3767711 0.2211707 0.1362221 0.1166321 0.07745033
## Cumulative Proportion  0.3767711 0.5979418 0.7341638 0.8507959 0.92824625
##                            Comp.6     Comp.7
## Standard deviation     0.53304287 0.46705627
## Proportion of Variance 0.04059067 0.03116308
## Cumulative Proportion  0.96883692 1.00000000
##
## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## LARent    0.503  0.015  0.143  0.178  0.627  0.334  0.436
## SC1      -0.317 -0.526 -0.032 -0.199  0.662 -0.259 -0.277
## Unempl    0.500  0.310  0.022  0.242  0.155 -0.285 -0.699
## LowEduc   0.216  0.343 -0.356 -0.832  0.118 -0.021  0.042
## Age18_24  0.202 -0.184  0.846 -0.415 -0.171  0.007 -0.095
## Age25_44  0.328 -0.552 -0.320 -0.069 -0.232  0.566 -0.326
## Age45_64 -0.448  0.414  0.183 -0.028  0.219  0.647 -0.356
##
##    ***********************************************************************
##    *   Results of Geographically Weighted Principal Components Analysis  *
##    ***********************************************************************
##
##    *********************Model calibration information*********************
##    Kernel function for geographically weighting: bisquare
##    Adaptive bandwidth for geographically and temporally  weighting: 116 (number of nearest neighbours
##    Distance metric for geographically weighting: A distance matrix is specified for this model calib
```

```
##
##    ****************     Summary of GWPCA information:    *****************
##    Local variance:
##               Min.  1st Qu.   Median  3rd Qu.   Max.
##    Comp.1 1.536711 3.068041 4.332136 5.940152 7.9615
##    Comp.2 0.931533 2.071545 2.356602 2.881007 4.6395
##    Comp.3 0.529633 1.105253 1.325191 1.683833 2.4005
##    Comp.4 0.270787 0.587662 0.670022 0.783144 1.3092
##    Comp.5 0.115799 0.343259 0.401930 0.463923 0.6075
##    Comp.6 0.075496 0.205034 0.237839 0.282046 0.4349
##    Comp.7 0.021830 0.073764 0.131214 0.164556 0.2641
##    Local Proportion of Variance:
##                    Min.   1st Qu.    Median   3rd Qu.     Max.
##    Comp.1      30.07411  39.15401  45.18099  50.42895  58.8209
##    Comp.2      17.64900  23.19946  25.76007  28.20702  33.0639
##    Comp.3       7.89154  10.07538  13.68997  18.67213  25.9903
##    Comp.4       3.55990   5.95048   6.97597   8.14268  12.9769
##    Comp.5       1.67337   3.30362   4.25634   5.01018   6.6513
##    Comp.6       1.05515   1.96108   2.50314   2.89324   4.9192
##    Comp.7       0.41927   0.80491   1.20446   1.63474   2.6682
##    Cumulative 100.00000 100.00000 100.00000 100.00000 100.0000
##
##    ***********************************************************************
##    Program stops at: 2019-05-08 00:36:22
```



Applying GW PCA the variables LARent, SC1, Unempl, LowEduc, Age18_24, Age25_44 and Age45_64 are used. It can be seen from the standard PCA that the first two components with global variables account for ≈ 60% of the data variation. The loadings show that the first components represents the percentage of renters and the percentage of unemployed, with the second component representing percentage of 25 to 44 years olds and the percentage of social class one, with the percentage of age group 45 to 64 year olds being next higest in both components. A bandwidth is then determined via cross validation to be 116, and the GW PAC is computed. The first two cumulative proportion is then determined for each ED and plotted. This plot shows that the local determination of proportion is generally higher in the GW PCA than the global values of 60% for the standard PCA.

Geographically Weighted Regression (GWR)

```
##    ***********************************************************************
##    *                       Package   GWmodel                             *
##    ***********************************************************************
##    Program starts at: 2019-05-08 00:36:26
##    Call:
##    gwr.basic(formula = GenEl2004 ~ LARent + SC1 + Unempl + LowEduc +
##     Age18_24 + Age25_44 + Age45_64, data = Dub.voter, bw = bw.gwr.2,
##     kernel = "bisquare", adaptive = TRUE)
##
##    Dependent (y) variable:  GenEl2004
##    Independent variables:  LARent SC1 Unempl LowEduc Age18_24 Age25_44 Age45_64
##    Number of data points: 322
##    ***********************************************************************
##    *                    Results of Global Regression                     *
##    ***********************************************************************
##
##    Call:
##     lm(formula = formula, data = data)
##
##    Residuals:
##     Min      1Q  Median      3Q     Max
## -23.537  -3.131   0.635   3.452  12.958
##
##    Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
##    (Intercept) 78.39806    3.87761  20.218  < 2e-16 ***
##    LARent      -0.09581    0.01756  -5.456 9.87e-08 ***
##    SC1          0.05240    0.06215   0.843  0.39981
##    Unempl      -0.72446    0.09383  -7.721 1.56e-13 ***
##    LowEduc     -0.15193    0.42969  -0.354  0.72389
##    Age18_24    -0.15982    0.05105  -3.131  0.00191 **
##    Age25_44    -0.39320    0.06311  -6.231 1.49e-09 ***
##    Age45_64    -0.07709    0.08898  -0.866  0.38699
##
##    ---Significance stars
##    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##    Residual standard error: 5.304 on 314 degrees of freedom
##    Multiple R-squared: 0.6371
##    Adjusted R-squared: 0.629
##    F-statistic: 78.75 on 7 and 314 DF,  p-value: < 2.2e-16
##    ***Extra Diagnostic information
```

```
##    Residual sum of squares: 8833.313
##    Sigma(hat): 5.253961
##    AIC:  1998.175
##    AICc:  1998.752
##    ***********************************************************************
##    *          Results of Geographically Weighted Regression              *
##    ***********************************************************************
##
##    *********************Model calibration information*********************
##    Kernel function: bisquare
##    Adaptive bandwidth: 109 (number of nearest neighbours)
##    Regression points: the same locations as observations are used.
##    Distance metric: Euclidean distance metric is used.
##
##    ****************Summary of GWR coefficient estimates:******************
##                    Min.    1st Qu.     Median    3rd Qu.     Max.
##    Intercept 54.9441744 74.2375737 81.9341207 94.2553001 116.5727
##    LARent    -0.1884996 -0.1187566 -0.0799932 -0.0409452   0.1089
##    SC1       -0.2738959 -0.0016703  0.2668400  0.4560556   0.7460
##    Unempl    -2.4545373 -1.1311798 -0.7508024 -0.4662021  -0.1106
##    LowEduc   -7.6594089 -0.9218641  0.3375128  1.6582372   3.0561
##    Age18_24  -0.4291152 -0.2797998 -0.1378394 -0.0280735   0.1760
##    Age25_44  -1.0482955 -0.7184957 -0.5066720 -0.3891038   0.1813
##    Age45_64  -0.9608610 -0.3681452 -0.0184767  0.1189194   0.5377
##    ************************Diagnostic information*************************
##    Number of data points: 322
##    Effective number of parameters (2trace(S) - trace(S'S)): 71.55592
##    Effective degrees of freedom (n-2trace(S) + trace(S'S)): 250.4441
##    AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 1923.867
##    AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 1842.416
##    Residual sum of squares: 4851.669
##    R-square value:  0.8006834
##    Adjusted R-square value:  0.7435072
##
##    ***********************************************************************
##    Program stops at: 2019-05-08 00:36:26
```

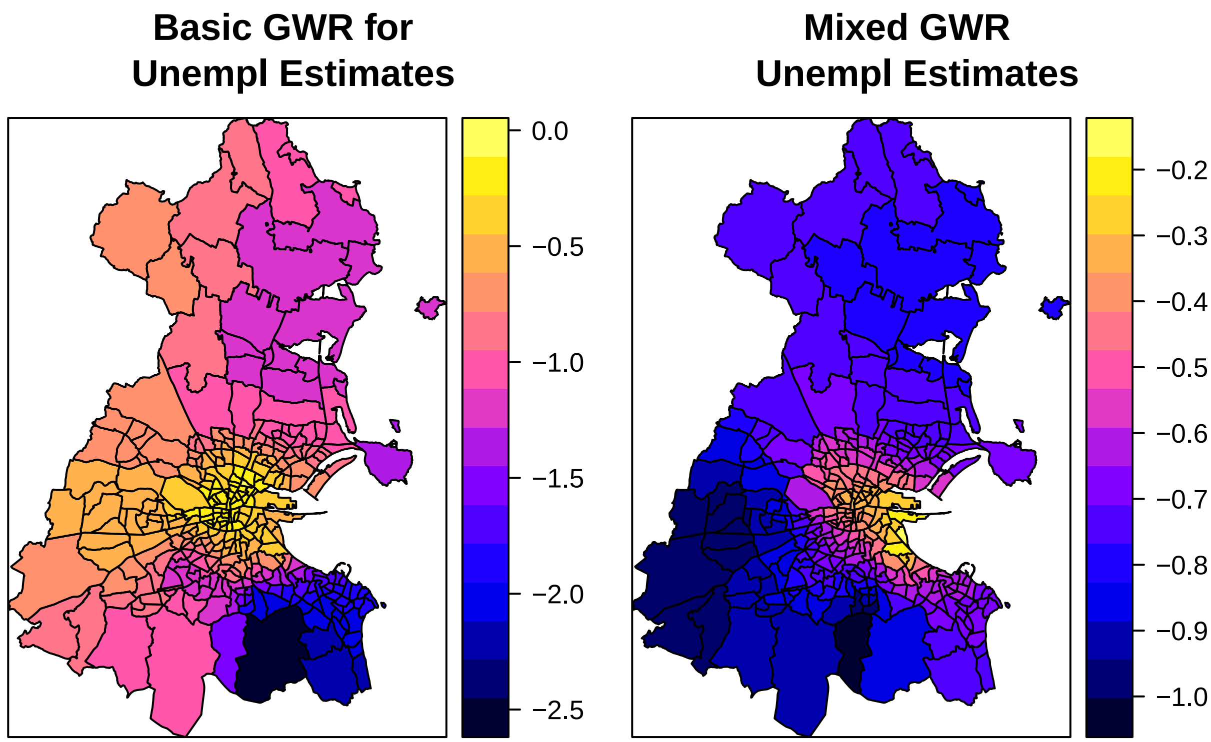
To perform a basic the bandwidth is determined to be 109 and the variables used in the model are LARent, SC1, Unempl, LowEduc, Age18_24, Age25_44 and Age45_64. The global regression model has an r^2 value of 0.63, whereas the GW regression model has an r^2 value of 0.8, meaning that the GW regression model accounts for more variability in the data.

```
##
## Tests based on the Monte Carlo significance test
##
##             p-value
## (Intercept)    0.22
## LARent         0.24
## SC1            0.00
## Unempl         0.00
## LowEduc        0.14
## Age18_24       0.10
## Age25_44       0.15
## Age45_64       0.12
```

Before computing the mixed GW regression model, which variable should be considered local and which should be considered global has to be determined. Using the Monte Carlo significance test is can be seen that only the variables SC1 and Unempl need to be considered local. The mixed model is then created using the same adaptive bandwidth as the basic model.

```
##    ***********************************************************************
##    *                       Package   GWmodel                             *
##    ***********************************************************************
##    Program starts at: 2019-05-08 00:36:33
##    Call:
##    gwr.mixed(formula = GenEl2004 ~ LARent + SC1 + Unempl + LowEduc +
##     Age18_24 + Age25_44 + Age45_64, data = Dub.voter, fixed.vars = c("LARent",
##     "LowEduc", "Age18_24", "Age25_44", "Age45_64"), intercept.fixed = TRUE,
##     bw = bw.gwr.2, kernel = "bisquare", adaptive = TRUE)
##
##    ********************Model calibration information********************
##    Mixed GWR model with local variables : SC1 Unempl
##    Global variables : Intercept LARent LowEduc Age18_24 Age25_44 Age45_64
##    Kernel function: bisquare
##    Adaptive bandwidth: 109 (number of nearest neighbours)
##    Regression points: the same locations as observations are used.
##    Distance metric: Euclidean distance metric is used.
##
##    ****************Summary of mixed GWR coefficient estimates:******************
##    Estimated global variables :
##                                   Intercept   LARent  LowEduc Age18_24
##    Estimated global coefficients:  83.66391 -0.11402  0.10223 -0.19868
##                                   Age25_44 Age45_64
##    Estimated global coefficients: -0.52139  -0.1663
##    Estimated GWR variables :
##                 Min.    1st Qu.     Median    3rd Qu.    Max.
##    SC1    -0.0093192  0.0552526  0.1933535  0.3774271  0.5664
##    Unempl -1.0039156 -0.7661523 -0.6661954 -0.5212127 -0.1793
##    ***********************Diagnostic information***********************
##    Effective D.F.:   20.12
##    Corrected AIC:   1947
##    Residual sum of squares:   6931
##
##    ***********************************************************************
##    Program stops at: 2019-05-08 00:37:11
```

The spactial variation difference between the two model can be seen below. The variable Unempl is used for that comparison as it is a local variable in both models thus the only changes are the other variables being used as global variables. (SC1 could be used also as it was also a local variable in both models.)



Conclusion

When dealing with data that has some spatial heterogeneity standard global models are not sufficient to appropriately account for the variability in the data. As is the case with the 2002 Dail Elections data for the 322 Electoral Divisions of Dublin. Steps where taken to minimise the collinearity effects on the data, which resulted in the removal of the variable DiffAdd. The remaining variables, LARent, SC1, Unempl, LowEduc, Age18_24, Age25_44 and Age45_64, where then used as predictor variables for models whose response variable was the voter turnout percentage of the population the Dublin EDs. Two types of geographically weighted models where created, PCA and Regression. These geographical weights allow the data to be expressed more locally than globally, thus producing a better discription of the voting habits for specific groups. The GW PCA model produces local proportions that are generally higher than that of the global values of 60% for the standard PCA, and the GW regression model results in an r^2 value of 0.8, higher than that of the global regression model which has an r^2 value of 0.63. The GW PCA and Regression models accounts for more variability in the data than the global models, and thus, produce a better description of voter habits.