NCG613 – Assignment 2

Dublin Voter Turnout in 2002 General Elections

Rahul Jadhav (17250785)

07 May 2018

Contents

1. Introduction

Linear Regression is one of the widely used statistical modeling techniques to predict continuous response. The strength and direction of relation between response and predictor variables is given by regression coefficients. We get one coefficient per variable used as a predictor. Geographical Weighted Regression model is a bit different where **coefficients are able to vary** across the space. Spatial variation in coefficients reveals otherwise hidden and meaningful patterns in spatial data.

In this report, we are going to analyze effect of migration, public housing, social-class, unemployment, educational attainment and age groups on overall electoral participation during 2002 General Elections held in 322 Electoral Divisions (EDs) in Dublin city of Ireland.

2. Data

'Dub.Voter' dataset available in R package 'GWmodel' contains following details per ED about voter turnout during 2002 General Elections in Dublin.

ED code,
centroid coordinates (Easting and Northing) of an ED,
percent of population with different address a year ago,
percent of population renting residence from local authority,
percent of population with head of household in social class 1,
percent of population seeking work,
percent of population educated only to lower secondary level,
percent of population aged between 18-24 inclusive,
percent of population aged between 25-44 inclusive,
percent of population aged between 45-64 inclusive,
percent of voting age population who voted on election day.

3. Problems

Here, we are interested in finding which among above mentioned variables **most strongly influence** the variation in voter turnout.

We have to **explore** given data to study collinearity among variables i.e. to study if variables are associated/correlated with each other (in addition to the response variable of voter turnout itself). Unlike experimental design where correlation among variables can be controlled or avoided by careful design, field studies like this do not offer this luxury to researchers. This **multicollinearity** among variables is considered as a problem in the context of statistical

modeling. Here, we are going to check multicollinearity among variables and take necessary actions to minimize its influence on the results we will get.

As we have location coordinates available in our data, while predicting overall voter turnout we are also going see if there is any specific trend which traces its origin to the area of voting. This aspect is studied under **spatial variability**.

4. Approach & Results

(A) Fitting model without considering spatial component:

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	77.7047	3.93928	19.7256	0.00000
DiffAdd	**-0.0858289**	**0.0859356**	**-0.998757**	**0.31869**
LARent	-0.0940168	0.0176518	-5.32619	0.00000
SC1	**0.0863655**	**0.0708484**	**1.21902**	**0.22375**
Unempl	-0.721619	0.0938726	-7.68722	0.00000
LowEduc	**-0.130728**	**0.430219**	**-0.303863**	**0.76143**
Age18_24	-0.139923	0.0547955	-2.55355	0.01114
Age25_44	-0.35365	0.0745036	-4.74675	0.00000
Age45_64	**-0.0920224**	**0.090232**	**-1.01984**	**0.30859**

From above summary, it is evident that predictors **DiffAdd**, **SC1**, **LowEduc** and **Age45_64** are not significant (p-value > 0.05) in determining overall voter turnout.

If we refit the model by removing non-significant predictors, we get following model which is better than previous one as it got lower AIC (1993.962 as against 1999.15), little higher Adjusted R_Squared (63 as against 62.9) and VIFs between 1 and 2.

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	75.8919	1.66021	45.7121	0.00000
LARent	-0.092285	0.0173077	-5.33202	0.00000
Unempl	-0.757483	0.0758968	-9.98044	0.00000
Age18_24	-0.154226	0.0505523	-3.05083	0.00247
Age25_44	-0.350024	0.0464538	-7.53487	0.00000

Thus, **LARent, Unempl, Age18_24 and Age25_44** are the predictors which **most strongly** influence voter turnout.

If 2 predictors are collinear (i.e. can be represented as linear combination of each other), we get unstable regression estimates with high standard error. To avoid this let us explore data to and check if multicollinearity exist.

From scatterplots, we observe that there is a positive linear trend between DiffAdd & Age25_44 with r = 0.7 and negative trend between Age24_44 & Age45_64 with r = -0.69. However, we observe that VIFs are within range of 1 to 3, we will check collinearity using principal component analysis **(PCA)**.

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
36.0844	25.5869	11.9196	10.5303	6.8905	3.6798	3.1114	2.1967

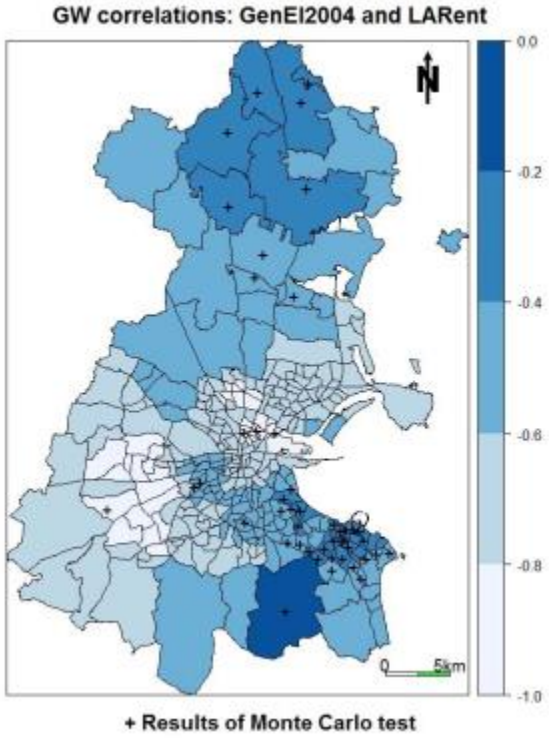
From above PCA components we can see that first 3 components explain almost **73% of variability**, suggesting global collinearity among predictors.

If relationships among variable are estimated as global across the space, then we use non-spatial regression model taking into consideration spatial autocorrelation(calculated using **Moran's I**). In this case where spatial autocorrelation do exist, we go for spatial model i.e. Geographically Weighted Regression model.

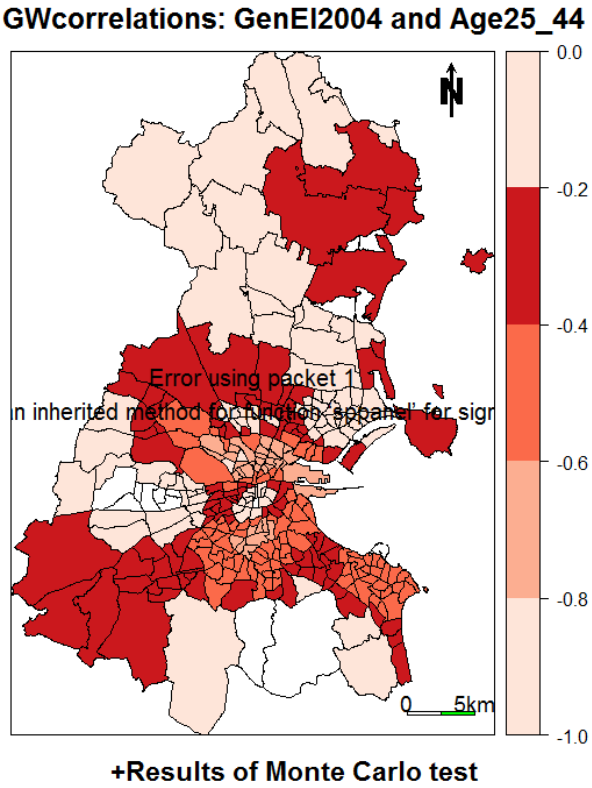
(B) Fitting GWR Model:

The residuals from a non-spatial predictive model may exhibit spatial pattern, so it is worthwhile to consider a spatial component.

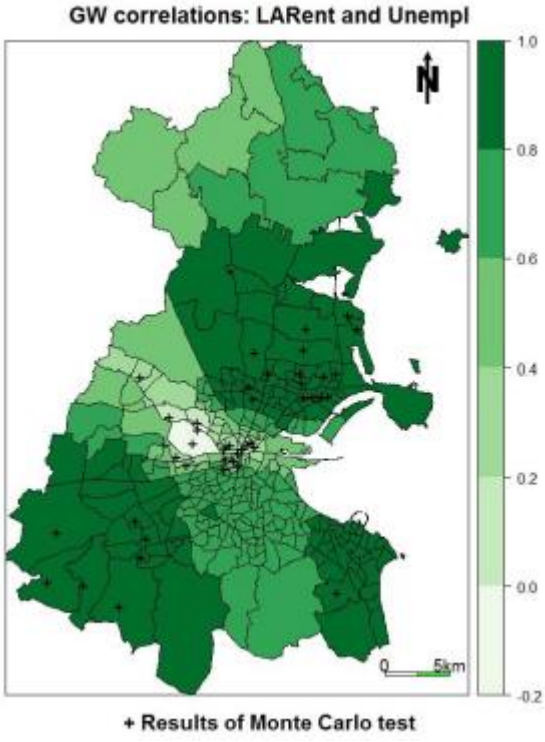
Let us check **GW correlations** to study local relationship between voter turnout and LARent (one of the significant predictors). Here, we will use bi-square kernel with adaptive bandwidth say = 48. We will also perform corresponding Monte Carlo test for correlation specifications.



From above plot, we see that relationship between voter turnout and LARent is strongest in central Dublin confirming it as a **local relationship**. Here, turnout is low where local authority renting (LARent) is high.

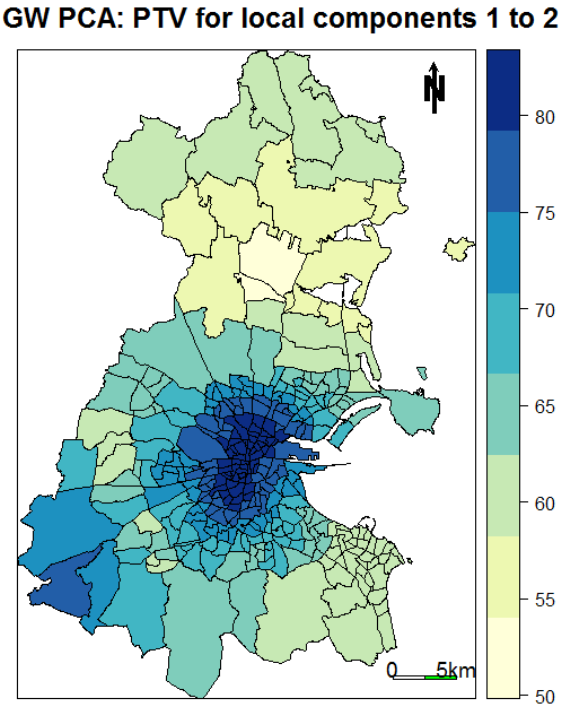


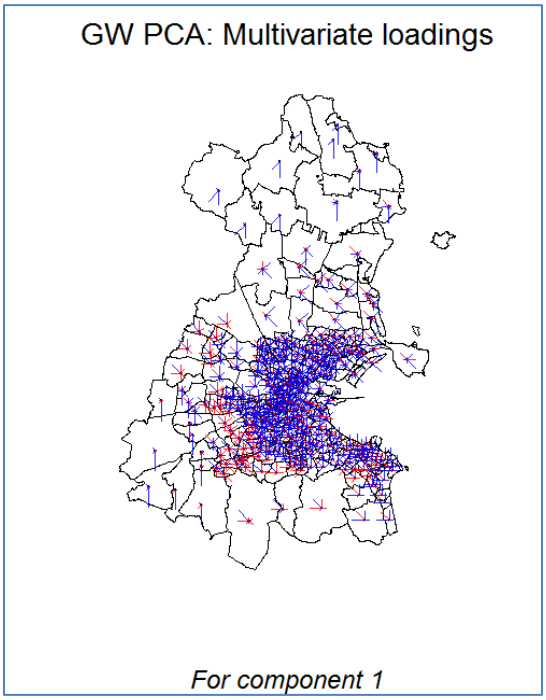
From above plot (ignoring error), we see that relationship between voter turnout and Age25_44 is strongest in north Dublin. This relationship is also **local**. Here, turnout is high for population between 25 and 44.

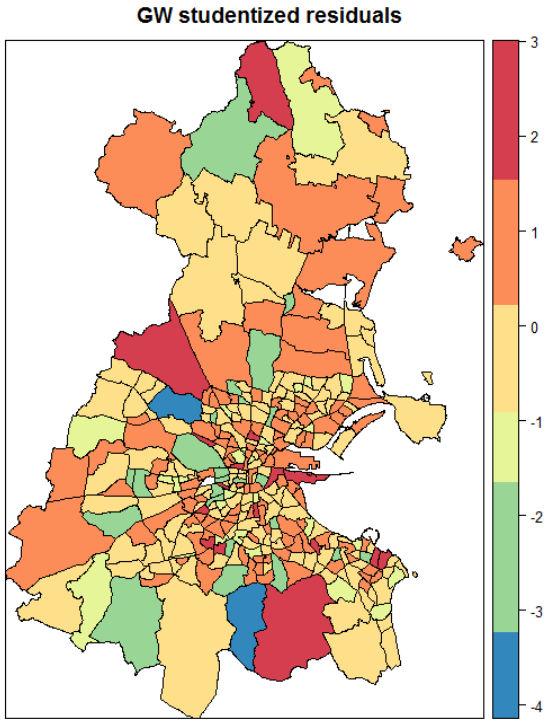


From above plot of LARent vs Unempl, we see strong positive correlation in 3 district areas of Dublin. This collinearity among predictors is a matter of concern.

Both of the following PCA plots show clear spatial variation. We can see higher percentages are found in central Dublin while lower percentages in north.







Above plot shows GW studentized residuals for basic GW model, which shows spatial pattern in voter turnout.

5. Conclusion

Kavanagh and team for their conclusions in paper “Turnout or turned off? Electoral participation in Dublin in the early 21st century“ applied basic GW regression to Dublin voter data. We have observed from various techniques above that there exists **considerable multicollinearity** among predictors. Thus, inferences made by model considering multicollinearity are going to be statistically more accurate than those made by Kavanagh and team using basic GW model.

6. References

Most of the concepts/code in this writing are referred from following paper:

The GWmodel R package: Further Topics for Exploring Spatial Heterogeneity using Geographically Weighted Models - Binbin Lu, Paul Harris, Martin Charlton & Chris Brunsdon

7. Appendix (R Code)

```
# (A) Fitting Model without considering Spatial Component

library('GWmodel')
library(MASS)
library('RColorBrewer')
library('olsrr')
library(spdep)
library(leaps)
require('gstat')
require('sp')

data(DubVoter)
summary(Dub.voter)

# fitting Multiple Linear Regression
pairs(Dub.voter[4:11]) # scatterplot to check collinearity
fit1 <- lm(GenEl2004 ~ DiffAdd + LARent + SC1 + Unempl + LowEduc + Age18_24 + Age25_44 +
Age45_64,data=Dub.voter)
ols_vif_tol(fit1) # to calculate Variance Inflation Factor to study multicollinearity
summary(fit1) # Adjusted R_Squared = 62.9
cor(Dub.voter$DiffAdd,Dub.voter$Age25_44)
cor(Dub.voter$Age45_64,Dub.voter$Age25_44)
AIC(fit1) # 1999.15

# refitting model moving insignificant parameters
fit2 <- lm(GenEl2004 ~  LARent + Unempl + Age18_24 + Age25_44,data=Dub.voter)
ols_vif_tol(fit2) # to calculate Variance Inflation Factor to study multicollinearity
summary(fit2) # Adjusted R_Squared = 63
AIC(fit2) # 1993.962

# PCA
Data.scaled <- scale(as.matrix(Dub.voter@data[,4:11]))
pca <- princomp(Data.scaled, cor=F)
(pca$sdev^2/sum(pca$sdev^2))*100
#Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
#36.084 25.586 11.919 10.530 6.890 3.679 3.111 2.196
pca$loadings

# Morans I Calculations
allfits = regsubsets(GenEl2004~.,data=Dub.voter[,4:12])
summary(allfits)$which
summary(allfits)$cp
```

```
which.min(summary(allfits)$cp)
names(which(summary(allfits)$which[which.min(summary(allfits)$cp),]==TRUE)[-1])
f4cp = lm(GenEl2004~Unempl+Age25_44+LARent+Age18_24,data=Dub.voter)
summary(f4cp)
AIC(f4cp)
f3cp = lm(GenEl2004~Unempl+Age25_44+LARent,data=Dub.voter)
AIC(f3cp)

f8 = lm(GenEl2004~.,data=Dub.voter[,4:12])
summary(f8)
AIC(f8)

turnout.step <- stepAIC(f8)
AIC(turnout.step)
summary(turnout.step)

Dub.nb = poly2nb(Dub.voter) # create a list of neighbours for each ED
Dub.listw = nb2listw(Dub.nb) #add weights - default is equal weight to every neighbour

moran.test(Dub.voter$GenEl2004, Dub.listw) # complete Moran test - H0 : I=0 - random pattern
# p << .05 => reject null hyp that I = 0 and conclude that pattern is not random

lm.morantest(f4cp,Dub.listw) # morantest on the residuals
# p << .05 => reject H0 and conclude that residuals are autocorrelated
lm.morantest(turnout.step,Dub.listw) # morantest on the residuals

# GW Studentized Residuals
# Bandwidth
library(raster)
Dubvoter.bw1 <-
bw.gwr(GenEl2004~DiffAdd+LARent+SC1+Unempl+LowEduc+Age18_24+Age25_44+Age45_64,
                data=Dub.voter,approach="AIC",kernel="bisquare",adaptive=T)
Dubvoter.bw1 # 109

# Basic GWR Model
Dubvoter.gw1 <-
gwr.basic(GenEl2004~DiffAdd+LARent+SC1+Unempl+LowEduc+Age18_24+Age25_44+Age45_64,
                  data = Dub.voter,bw=Dubvoter.bw1,kernel="bisquare",adaptive=T)

mypalette<- rev(brewer.pal(11,"Spectral"))
spplot(Dubvoter.gw1$SDF,"Stud_residual",key.space = "right",cuts=10,col.regions=mypalette,main="GW
studentized residuals")
library(classInt)
nBrks <- 6
brks <- classIntervals(Dubvoter.gw1$SDF$Stud_residual, n=nBrks,style="hclust")$brks
```

```
VerySmallNumber <- 2 * (.Machine$double.eps)
brks[1]<-brks[1] - VerySmallNumber
brks[length(brks)] + VerySmallNumber
mypalette<-rev(brewer.pal(length(brks)-1,"Spectral"))
spplot(Dubvoter.gw1$SDF,"Stud_residual",key.space = "right",at=brks, col.regions=mypalette,
main="GW studentized residuals")

# (B) GWR Spatial Model
#Checking GW Correlation among few predictors

gwss.1 <- gwss(Dub.voter,vars = c("GenEl2004", "LARent", "Unempl", "Age25_44"),kernel="bisquare",
adaptive=TRUE, bw=48)
gwss.mc <- montecarlo.gwss(Dub.voter,vars = c("GenEl2004", "LARent","Unempl", "Age25_44"),
kernel="bisquare", adaptive=TRUE, bw=48)
gwss.mc.data <- data.frame(gwss.mc)
gwss.mc.out.1 <- ifelse(gwss.mc.data$Corr_GenEl2004.LARent < 0.975 &
gwss.mc.data$Corr_GenEl2004.LARent > 0.025 , 0, 1)
gwss.mc.out.2 <- ifelse(gwss.mc.data$Corr_LARent.Unempl < 0.975 &
gwss.mc.data$Corr_LARent.Unempl > 0.025 , 0, 1)
gwss.mc.out.3 <- ifelse(gwss.mc.data$Corr_GenEl2004.Age25_44 < 0.975 &
gwss.mc.data$Corr_GenEl2004.Age25_44 > 0.025 , 0, 1)
gwss.mc.out <- data.frame(Dub.voter$X, Dub.voter$Y, gwss.mc.out.1,gwss.mc.out.2,gwss.mc.out.3)
gwss.mc.out.1.sig <- subset(gwss.mc.out, gwss.mc.out.1==1, select =c(Dub.voter.X, Dub.voter.Y,
gwss.mc.out.1))
gwss.mc.out.2.sig <- subset(gwss.mc.out, gwss.mc.out.2==1, select =c(Dub.voter.X, Dub.voter.Y,
gwss.mc.out.2))
gwss.mc.out.3.sig <- subset(gwss.mc.out, gwss.mc.out.3==1, select =c(Dub.voter.X, Dub.voter.Y,
gwss.mc.out.3))
pts.1 <- list("sp.points", cbind(gwss.mc.out.1.sig[,1],gwss.mc.out.1.sig[,2]), cex=2, pch="+", col="black")
pts.2 <- list("sp.points", cbind(gwss.mc.out.2.sig[,1],gwss.mc.out.2.sig[,2]), cex=2, pch="+", col="black")
pts.3 <- list("sp.points", cbind(gwss.mc.out.3.sig[,1],gwss.mc.out.3.sig[,2]), cex=2, pch="+", col="black")

mypalette.gwss.1 <-brewer.pal(5,"Blues")
mypalette.gwss.2 <-brewer.pal(6,"Greens")
mypalette.gwss.3 <-brewer.pal(4,"Reds")

map.na <- list("SpatialPolygonsRescale", layout.north.arrow(),offset = c(329000,261500), scale = 4000,
col=1)
map.scale.1 <- list("SpatialPolygonsRescale", layout.scale.bar(),offset = c(326500,217000), scale = 5000,
col=1, fill =c("transparent", "green"))
map.scale.2 <- list("sp.text", c(326500,217900), "0", cex=0.9,col=1)
map.scale.3 <- list("sp.text", c(331500,217900),"5km",cex=0.9,col=1)
map.layout.1 <-list(map.na,map.scale.1,map.scale.2,map.scale.3,pts.1)
map.layout.2 <-list(map.na,map.scale.1,map.scale.2,map.scale.3,pts.2)
map.layout.3 <-list(map.na,map.scale.1,map.scale.2,map.scale.3,pts.3)
```

```
X11(width=10,height=12)
spplot(gwss.1$SDF,"Corr_GenEI2004.LARent",key.space = "right",col.regions = mypalette.gwss.1,at=c(-1,-0.8,-0.6,-0.4,-0.2,0),par.settings = list(fontsize=list(text=15)), main = list(label="GWcorrelations: GenEI2004 and LARent", cex=1.25), sub=list(label="+Results of Monte Carlo test", cex=1.15), sp.layout=map.layout.1)

X11(width=10,height=12)
spplot(gwss.1$SDF,"Corr_LARent.Unempl",key.space = "right",col.regions=mypalette.gwss.2,at=c(-0.2,0,0.2,0.4,0.6,0.8,1),par.settings=list(fontsize=list(text=15)), main=list(label="GWcorrelations: LARent and Unempl", cex=1.25), sub=list(label="+Results of Monte Carlo test", cex=1.15), sp.layout=map.layout.2)

X11(width=10,height=12)
spplot(gwss.1$SDF,"Corr_GenEI2004.Age25_44",key.space = "right",col.regions = mypalette.gwss.3,at=c(-1,-0.8,-0.6,-0.4,-0.2,0),par.settings = list(fontsize=list(text=15)), main = list(label="GWcorrelations: GenEI2004 and Age25_44", cex=1.25), sub=list(label="+Results of Monte Carlo test", cex=1.15), sp.layout=map.layout.1)

# to find an optimal adaptive bandwidth using a bi-square kernel using k=3
Coords <- as.matrix(cbind(Dub.voter$X,Dub.voter$Y))
Data.scaled.spdf <- SpatialPointsDataFrame(Coords,as.data.frame(Data.scaled))
bw.gwpca.1 <-bw.gwpca(Data.scaled.spdf,vars = colnames(Data.scaled.spdf@data), k=3, adaptive=TRUE)


# The GW PCA fit
gwpca.1 <- gwpca(Data.scaled.spdf, vars = colnames(Data.scaled.spdf@data), bw=bw.gwpca.1, k=8, adaptive=TRUE)

# visualisations
prop.var <- function(gwpca.obj, n.components) {
return((rowSums(gwpca.obj$var[,1:n.components])/rowSums(gwpca.obj$var))*100)}
var.gwpca <- prop.var(gwpca.1,2)
Dub.voter$var.gwpca <- var.gwpca
mypalette.gwpca.1 <-brewer.pal(8,"YlGnBu")
map.layout.3 <- list(map.na,map.scale.1,map.scale.2,map.scale.3)

X11(width=10,height=12)
spplot(Dub.voter,"var.gwpca",key.space = "right", col.regions = mypalette.gwpca.1, cuts=7, par.settings =list(fontsize=list(text=15)),main=list(label="GW PCA: PTV for local components 1 to 2", cex=1.25),sp.layout=map.layout.3)

loadings.1 <- gwpca.1$loadings[,,1]
X11(width=10,height=12)
plot(Dub.voter)
glyph.plot(loadings.1,Coords,r1=20,add=T,alpha=0.85)
```

```
title(main=list("GW PCA: Multivariate loadings", cex=1.75,col="black", font=1), sub=list("For component 1", cex=1.5,col="black", font=3))

gwpca.mc <-montecarlo.gwpca.2(Data.scaled.spdf, vars = colnames(Data.scaled.spdf@data), k=3, adaptive=TRUE)
X11(width=8,height=5)
plot.mcsims(gwpca.mc)
```