Spatial Modelling of 1980 USA Election Turnout

Brian K. Masinde

6-24-2023

Introduction

Presidential elections in the USA are a chance for the population to elect a president that aligns with their socio-economic values and transformation views. There are assumptions that socio-economic or demographic factors influence persons political interest and their voting preference (Skoutaris 2017). Voter turnout is a way of measuring the electorates apathy or interest in elections. Therefore, modelling voter turnout is a topic of interest in the US to better understand spatial socio-economic variations and policies that influence voter turn out across US counties and states (McDonald and Popkin 2001; Rosenstone 1982). In this mini-project the aim is to model voter turn out across counties in the USA in the 1980 election where the main two candidates were Ronald Reagan and Jimmy Carter. Since the data is irregular lattice, I use spatial econometric models and compare this to the linear regression model.

Objectives

The overall objective is to model the spatial economic and demographic factors influencing voter turn out.

1. Fit a linear model as the base model for comparison. Test whether, a linear model is well specified for this data.
2. Fit CAR model
3. Interpret the results

Data

The data is from 1980 elections and the 1980 census in the USA (see Figure 1). This data with 3107 counties (observations) was compiled by Pace and Barry (1997). They found it interesting that a US presidential election had coincided with a national census and saw it as opportunity to model voter turn out as an outcome of socio-economic status. The dataset contains the following variables:

1. Votes cast as a proportion of population over age of 19 and eligible to vote. This is the outcome variable of interest.
2. Population with college degrees as proportion of population over age 19 and eligible to vote.
3. Home ownership as proportion of population over age 19 eligible to vote.
4. Income per capita of population over age 19 eligible to vote.

This dataset does not contain the lattice/polygons of the counties and therefore I had to source US county data online. I did a spatial join to merge the county data to the voter turn out data. The image below shows voter turnout as a proportion of the population eligible to vote in a sample of the US counties.

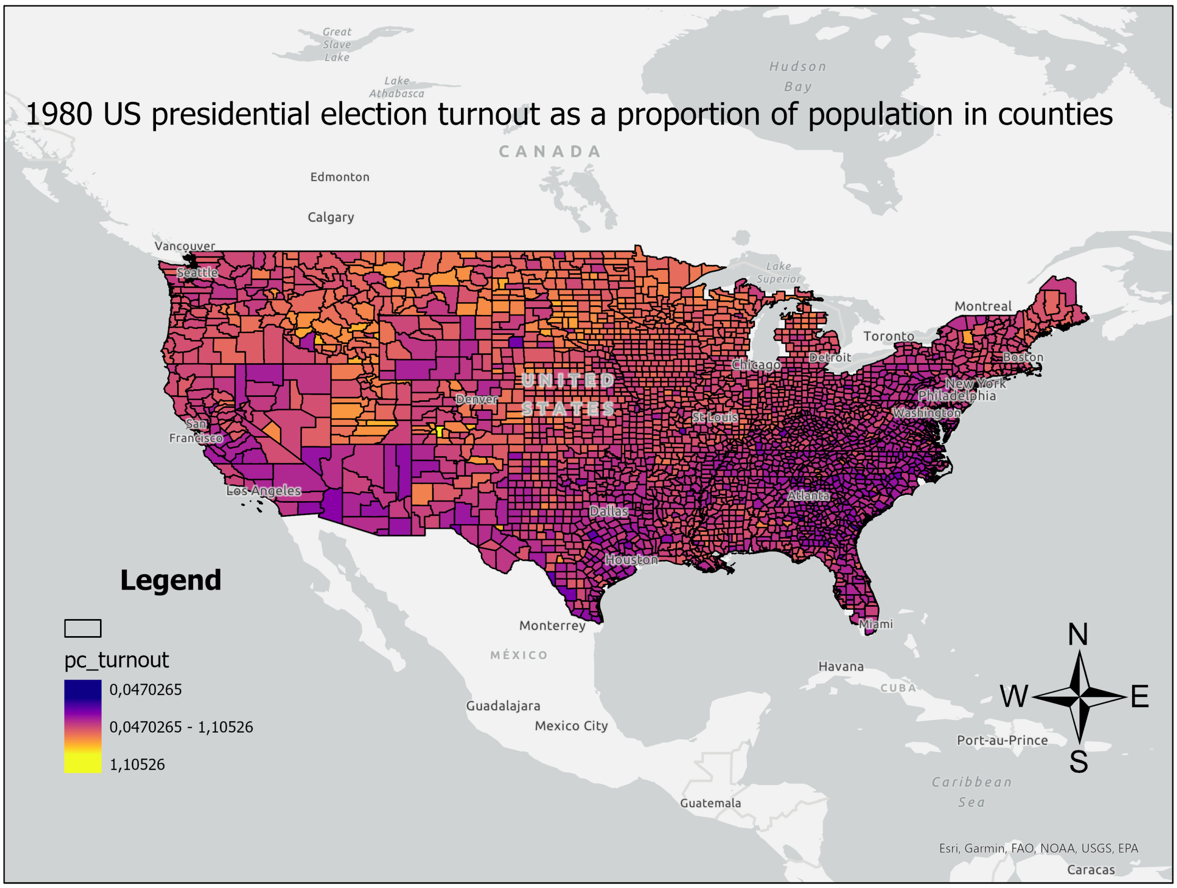

Figure 1: 1980 US presidential election turnout

Methods

Linear regression

Linear regression model fits a linear relationship between the dependent variable and the independent variables. The residuals in a linear regression model must be independent and identically distributed $\sim N(0, \sigma^2)$. For our data the linear model is specified as:

$$voter\ turnout = \beta_0 + \beta_1\ College + \beta_2\ home - ownership + \beta_3\ aggregate\ income + \epsilon$$

Spatial regression

There are two main models for modelling irregular lattice data. Spatial autoregressive (SAR) models and conditional autoregressive models (CAR). For this data and objectives I use the CAR model because it only takes into account the local spatial influence of neighbors of a particular lattice while SAR is global (i.e., effect of all latices is taken into account) (De Smith 2018). My choice is backed by the belief that election apathy/interest would be more of a local issue and thus sentiments in further states may not be able to influence sentiments in further states (i.e., it is not equivalent to a diffusion/interaction/transfer process to warrant a SAR). I use the Queen weights which is symmetrical. Queen weights consider the neighbors as sharing both a vertices and edges, as an example, a regular lattice would have 8 neighbors (Anselin, Syabri, and Kho 2009). The general CAR model is:

$$\begin{aligned} E\big[y(s_i)|y(s_j), j \neq i\big] &= \mu(s_i) + \sum_{j \neq i} C_{ij}\big(y(s_j) - \mu(s_j)\big) \\ &= x_i\beta\ + \sum_{j \neq i} C_{ij}\big(y(s_j) - x_j\beta\big) \end{aligned}$$

Where:

$$C_{ij} = \rho_s W_{ij}$$

W_{ij} are the spatial weights which indicate a neighbor relationship while ρ_s are the "strength of the spatial dependence" (LeSage and Pace 2009).

I also use Moran's I and Lagrange multiplier tests to check for spatial correlation.

Results

Linear model results

The linear model fitted is as follows:

$$voter\ turnout = \beta_0 + \beta_1\ College + \beta_2\ home - ownership + \beta_3\ aggregate\ income + \epsilon$$

The results show that all 3 variables are significant in explaining voter turn out. College attendance and home-ownership have a positive relationship with voter turn out meaning if a person and has attended college and is a homeowner they were likely to vote in the election. However, income has a negative estimate meaning that high income is associated with political disinterest or apathy during the 1980 election.

Table 1: Linear model results

coefficients	Estimates	P value	Significance
Intercept	0.078344	1.84e-07	***
college	0.693881	2e-16	***
homeowner	0.895777	2e-16	***
income	-0.020219	2e-16	***

Plotting and observing residuals vs leverage informs where the residuals meet the IID assumption. In this case there is a slight patter which means there's some other process underlying the data generating process.

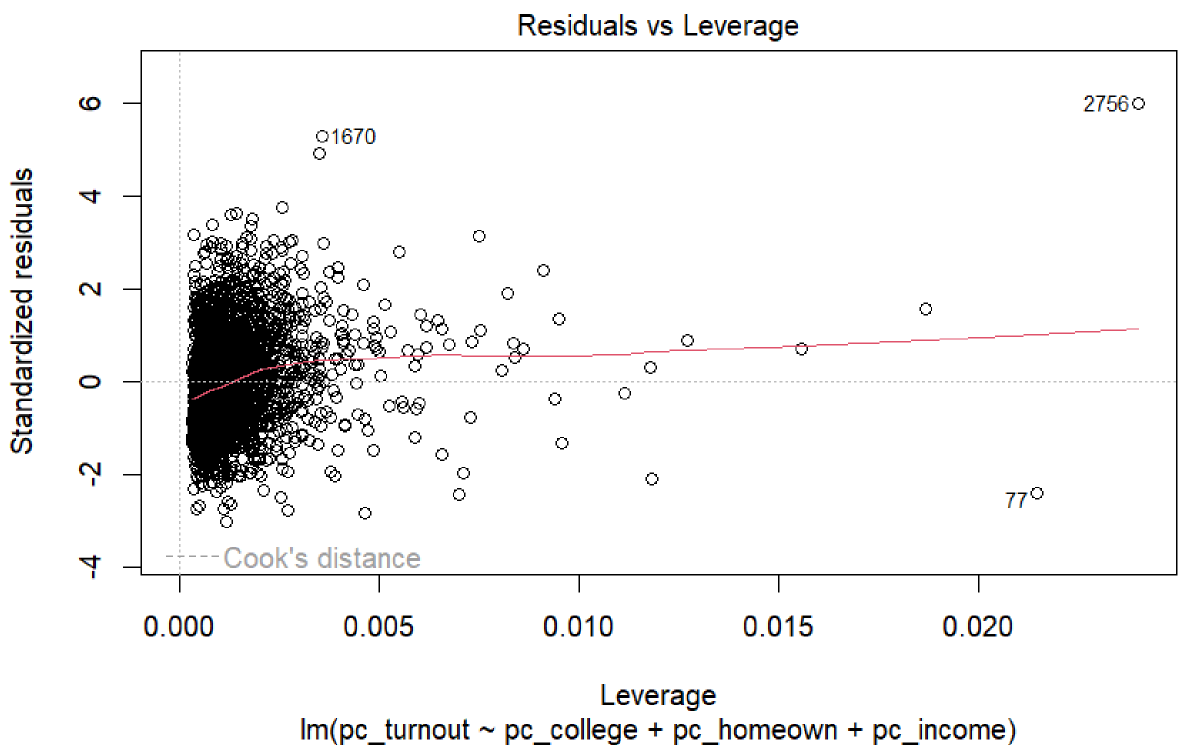

Figure 2: Residuals vs leverage to check IID assumptions

Spatial analysis results

Spatial clustering and dependence

Here I present the results of the Moran's I test and Lagrange multiplier tests based on the Queen weights (see Figure 3, appendix for queen weights plot).

Table 2: Moran's I results

Morans I statistic	P value
0.6114038375	2.2e-16

Table 3: Lagrange Multiplier test

Tests	Statistic	P value	Significance
LMerr	1740.8	2.2e-16	***
LMlag	1310.6	2.2e-16	***

Moran's I test results (Table 2), we reject the null hypothesis of no spatial dependence.

Lagrange multiplier test results (Table 3) we also reject the null hypothesis of no spatial dependence. Furthermore, the Lagrange multiplier test shows that there is spatial dependence in the errors, and the spatial lag models.

CAR model

For the CAR model i fit two models, one with only the four independent variables (income, homeowner, and college educated) and the second model I added dummy variable as to whether the counties belonged to West, North East, Mid-West, and South of the US. The results are in Table 4 and 5 respectively.

```
# snippet of the code for the first model
# I set the option zero.policy = TRUE to solve the problem of Islands
## in the polygons
e80_CAR <- spautolm(pc_turnout ~ pc_college + pc_homeown + pc_income,
                    data = voter_turn,
                    listw = nb2listw(queenNB, zero.policy = TRUE),
                    family = "CAR",
                    zero.policy = TRUE
                    )
summary(e80_CAR)
```

Table 4: CAR model results

Coefficients	Estimates	P value	Significance
Intercept	1.581024	2.2e-16	***
college	-0.276703	2.2e-16	***
homeowner	1.1985873	2.2e-16	***
income	0.0164498	2.2e-16	***

Under the CAR model 1 (Table 4), all the explanatory variables are significant in explaining voter turn out in the US counties. Interestingly, college attendance has a negative relationship with voter turn out. This means, taking into account the spatial dependence in the data, college educated were less likely to vote during the 1980 US election, and this can interpreted as disapproval of the policies of the running candidates. In contrast home owners and high income earners were likely to vote and this could mean that they were optimistic in their candidate of choice winning the election.

Table 5: CAR model 2 results with regional dummies

Coefficients	Estimates	P value	Significance
Intercept	0.27815632	2.2e-16	***
college	0.17514987	8.967e-11	***
homeowner	0.94169813	2.2e-16	***
income	-0.00073883	0.5714	-
North East	-0.11973614	6.661e-14	***
Mid-West	-0.07505448	8.000e-12	***
South	-0.14594108	2.2e-16	***
West	NA	NA	-

Under CAR model 2 (Table 5) with regions with taken into account all explanatory variables are significant in explaining voter turnout in the US except aggregate income. College attendance in this model shows a positive relationship with voter turnout meaning that college educated were likely to vote during the 1980 US elections. This is in contrast to results in Table 4. The regions (North East, Mid-West, South) show a negative relationship with voter turnout. No statistics are obtained for the West region and this maybe attributed to the Islands in the dataset that result in an asymmetric weight (this is also listed as a limitation of the analysis).

Conclusion

In summary, while linear regression model had significant coefficient estimates, the model violates IID assumption. Importantly, the dataset being an irregular lattice data the appropriate methods to implement are spatial methods. Using CAR (which takes into account local spatial dependence) the analysis shows that income, home-ownership, and college education can be used as proxies to explain voter turnout among the sampled counties in the USA.

Limitations

The weights are not symmetrical and I attribute this to the data processing challenges. The data had many islands. Therefore the model above may not be robust enough.

Bibliography

Anselin, Luc, Ibnu Syabri, and Youngihn Kho. 2009. "GeoDa: An Introduction to Spatial Data Analysis." In *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, 73–89. Springer.

De Smith, Michael John. 2018. *Statistical Analysis Handbook*. The Winchelsea Press.

LeSage, J., and R. Pace. 2009. *Introduction to Spatial Econometrics*. CRC Press.

McDonald, Michael P, and Samuel L Popkin. 2001. "The Myth of the Vanishing Voter." *American Political Science Review* 95 (4): 963–74.

Pace, R Kelley, and Ronald Barry. 1997. "Quick Computation of Spatial Autoregressive Estimators." *Geographical Analysis* 29 (3): 232–47.

Rosenstone, Steven J. 1982. "Economic Adversity and Voter Turnout." *American Journal of Political Science*, 25–46.

Skoutaris, Christoforos. 2017. "Modeling Election Results as a Function of Geodemographical and Lifestyle Variables."

Appendix

Plots

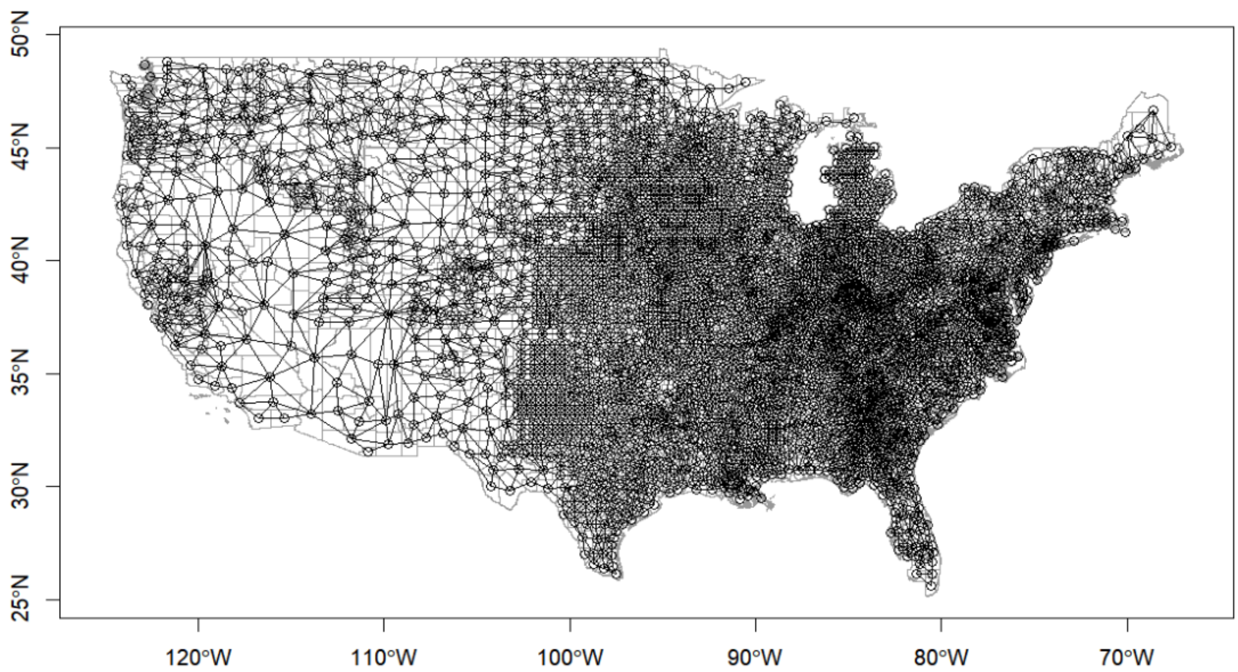

Figure 3: Queen neighbors plot

Code

```
# packages
library(sp)
library(rgdal)
library(classInt)
library(spdep)
library(ggplot2)
library(corrplot)
library(dplyr)
library(spatialreg)
#library(png)


# Loading data ________________________________________________________________
path2Data <- "~/PhD Track/courses i take/spatial statistics/data/elect80_sj.shp"
voter_turn <- readOGR(path2Data)


# Data exploration ____________________________________________________________

## Adding regions to the data
nrtheast <- c("Connecticut", "Maine", "Massachusetts", "New Hampshire",
```

```
               "Rhode Island", "Vermont", "New Jersey", "New York", "Pennsylvania")

mdwest <- c("Illinois", "Indiana", "Michigan", "Ohio",
            "Wisconsin", "Iowa", "Kansas", "Minnesota", "Missouri", "Nebraska",
            "North Dakota", "South Dakota")

sth <- c("Delaware", "Florida", "Georgia", "Maryland", "North Carolina",
         "South Carolina", "Virginia", "Washington DC", "West Virginia",
         "Oklahoma", "Texas",  "Tennessee", "Kentucky", "Alabama",
         "Mississippi", "Louisiana", "Arkansas")
wst <- c("Arizona", "Colorado", "Idaho", "Montana",  "Nevada", "New Mexico",
         "Utah", "Wyoming", "Alaska", "California", "Hawai", "Oregon", "Washington")

regions_lst <- list("nrtheast" = nrtheast, "mdwest" = mdwest, "sth" = sth, "wst" = wst)



# create new columns in the data fill with NA
voter_turn@data$region <- rep(NA, nrow(voter_turn@data))


# loop throw each row check whether obsevaration is in the region
for (i in 1:nrow(voter_turn@data)) {
  if (voter_turn@data$STATENAM[i] %in% regions_lst$nrtheast == TRUE) {
    voter_turn@data$region[i] <- 1
  } else if (voter_turn@data$STATENAM[i] %in% regions_lst$mdwest == TRUE) {
    voter_turn@data$region[i] <- 2
  } else if (voter_turn@data$STATENAM[i] %in% regions_lst$sth == TRUE) {
    voter_turn@data$region[i] <- 3
  } else if (voter_turn@data$STATENAM[i] %in% regions_lst$wst == TRUE) {
    voter_turn@data$region[i] <- 4
  }
}

# convert it to a long table
voter_turn@data$nrtheast <- rep(NA, nrow(voter_turn@data))
voter_turn@data$mdwest <- rep(NA, nrow(voter_turn@data))
voter_turn@data$sth <- rep(NA, nrow(voter_turn@data))
voter_turn@data$wst <- rep(NA, nrow(voter_turn@data))

for (i in 1:nrow(voter_turn@data)) {
  if (voter_turn@data$region[i] == 1) {
    voter_turn@data$nrtheast[i] <- 1
    voter_turn@data$mdwest[i] <- 0
    voter_turn@data$sth[i] <- 0
    voter_turn@data$wst[i] <- 0
  } else if (voter_turn@data$region[i] == 2) {
    voter_turn@data$nrtheast[i] <- 0
    voter_turn@data$mdwest[i] <- 1
    voter_turn@data$sth[i] <- 0
    voter_turn@data$wst[i] <- 0
  } else if (voter_turn@data$region[i] == 3) {
    voter_turn@data$nrtheast[i] <- 0
```

```
    voter_turn@data$mdwest[i] <- 0
    voter_turn@data$sth[i] <- 1
    voter_turn@data$wst[i] <- 0
  } else if (voter_turn@data$region[i] == 4) {
    voter_turn@data$nrtheast[i] <- 0
    voter_turn@data$mdwest[i] <- 0
    voter_turn@data$sth[i] <- 0
    voter_turn@data$wst[i] <- 1
  }
}

# Linear regression ____________________________________________________________
lm_elect80 <- lm(formula = pc_turnout ~ pc_college + pc_homeown + pc_income,
                 data = voter_turn@data)

summary(lm_elect80)

# checking plots to see if IID assumptions are met
lm_elect80_2 <- lm(formula = pc_turnout ~ pc_college + pc_homeown + pc_income,
                 data = voter_turn2)
plot(lm_elect80_2)


# Spatial Analysis _____________________________________________________________

## Creating queen weights
sf_use_s2(FALSE)
queenNB <- poly2nb(voter_turn, queen = TRUE)

## plotting neighbor weights
plot(voter_turn, border="grey60", axes=TRUE)
plot(queenNB, coordinates(voter_turn), add=TRUE)

## Moran's I test using queen weights
elect80MoranQ <- moran.test(voter_turn@data$pc_turnout,
                            listw = nb2listw(queenNB, zero.policy = TRUE),
                            zero.policy = TRUE)
elect80MoranQ

## Lagrange multiplier test (Queen)
OLSResMoranQ <- lm.LMtests(lm_elect80,
                               listw = nb2listw(queenNB, zero.policy = TRUE),
                             test = c("LMerr","LMlag"),
                             zero.policy = TRUE)
OLSResMoranQ


## CAR model 1
e80_CAR <- spautolm(pc_turnout ~ pc_college + pc_homeown + pc_income,
                    data = voter_turn,
                    listw = nb2listw(queenNB, zero.policy = TRUE),
                    family = "CAR",
                    zero.policy = TRUE
```

```r
                    )
e80_CAR

## CAR model 2
e80_CAR_2 <- spautolm(pc_turnout ~ pc_college + pc_homeown + pc_income + nrtheast +
                        mdwest + sth + wst,
                    data = voter_turn,
                    listw = nb2listw(queenNB, zero.policy = TRUE),
                    family = "CAR",
                    zero.policy = TRUE
                    )
summary(e80_CAR_2)
```