442 HW4

Summary

We want to find out what are the most important demographic factors in Wisconsin that seem to be causing strong spatial pattern in supporting Trump in 2016. In addition, we want to find out that is Trumpism a urban/rural or racial phenomenon and are there any important spatial variables we did not consider? By fitting a BYM model with spatial random effect, we find that White people who live in rural area are more likely to vote for Trump. The vote for Trump is higher in rural areas with higher proportion of White people, which means Trumpism should be considered both rural and racial phenomenon. The population density is the most important demographic factors in causing spatial pattern affecting the odds of voting for Trump. The spatial variation is significant throughout Wisconsin and spatial random effect cannot be ignored. There might be other important spatial explanatory variable, such as household income and educational level, that we didn't consider which lead to a less precise conclusion.

Introduction

The U.S. Presidential Election has always been a hot topic all across the globe. In this assignment, we want to find out that what factors are most important in influencing the votes for Trump in 2016 Presidential Election. To narrow down our field of interest, we studied the spatial data set associated various factors for Wisconsin—which is a swing state. To be more specific, we want to find out that what demographic factors seems to be causing a strong spatial pattern in supporting Trump. Is Trumpism a urban/rural or racial phenomenon? Are there some spatial explanatory variables that we ignored but is important? Is the spatial variation significant across the sub-counties in Wisconsin?

Method and model model

The dataset is considered to be areal dataset and we fit a BYM model to the data with spatial random effect:

$$Y_i \sim Binomial(N_i, \rho_i)$$

Y_i is the number votes for Donald Trump in sub-county i. N_i is the total number of votes in sub-county i and ρ_i is the probability of a voter voting for Trump in sub-county i. The probability that a voter vote for Trump shouldn't be too small(i.e. >0.1), and each person will either vote for Trump or not vote for Trump, thus we consider Y_i follows Binomial distribution instead of a Poisson distribution.

$$log\left(\rho_i/(1-\rho_i)\right) = \mu + X_i\beta + U_i$$

$$U_i \sim BYM(\sigma^2, \tau^2)$$

The $\rho_i/(1-\rho_i)$ is the odds for voting for Trump and $log\left(\rho_i/(1-\rho_i)\right)$ is the log odds. μ represents the intercept and X_i represents a vector of "logPdens", "propWhite" and "propInd", where "logPdens" is the log of total population to surface area(log population density), "propWhite" is the proportion of each region which is White and "propInd" is the proportion of each region which is Indegenous. β represents the corresponding vector of parameters for X_i. U_i represents the spatial random effects(residual spatial variation), with spatial variance σ^2 and spatial independent variance(non-spatial) τ^2.

$$\theta_1 = \sqrt{\sigma^2 + \tau^2}$$

$$\theta_2 = \sigma/\sqrt{\sigma^2 + \tau^2}$$

$$Prob\left(\theta_1 > log(1.5)\right) = 0.5$$

$$Prob(\theta_2 > 0.5) = 0.5$$

θ_1 represents the standard deviation for spatial random effect(U_i) and θ_2 represents the spatial proportion(dependency parameter). U_i is on the log scale and θ_1 largely determine the characteristic of U_i, we don't want U_i to be too large or too small. By the best guess, $log(1.5)$ is big enough for standard deviation on the log scale, thus we take prior for θ_1 such that $Prob\left(\theta_1 > log(1.5)\right) = 0.5$. In addition, we do not have much experience about the prior for θ_2 and by the best guess we just take the prior median as 0.5 that $Prob(\theta_2 > 0.5) = 0.5$. Moreover, the INLA function use default prior which is not specified in the code. The normal prior for the intercept β_0 and other β's:

$$\beta_0 \sim N(0, \infty)$$

$$\beta_1 .. \beta_3 \sim N(0,1000)$$

Results

Figure 1-4 are generated from real data to visualized the proportion of voting for Trump, population density, proportion of Indegenous people and proportion of White people. In Figure 1, the dark red represents all voter votes for Trump and dark blue represents no one vote for Trump. From Figure 2, the red color means highest population density(could be interpreted as major cities) and dark blue means the lowest population density(yellow, green and blue colors could be interpreted as rural areas). In Figure 3, the dark red represents the sub-county that has about 100% Indegenous people(Indian reservation) and dark purple represents the sub-county that has almost no Indegenous people. In Figure 4, the dark red represents the sub-county that has almost 100% of White people whereas the blue color correspond to Indian reservation. Combined Figure 1-4, the conclusion we can get is that Indegenous people are unlikely to vote for Trump whereas White people who live in rural areas are more likely to vote for Trump. Thus, Trumpism is more likely both rural and racial phenomenon. Population density, proportion of Indegenous people and proportion of White people all seems to influence the Trump support.

Figure 5 shows the spatial random effects ($E(U|Y)$) and Figure 6 shows the predicted probability of voting for Trump based on the covariates($E(\lambda|Y)$). In Figure 5, the green rectangle shows the area that the model predicts that people have sort of average probability voting for Trump. However, compared to the same region in Figure 6, the random effect shows a probability of voting for Trump bigger than average(lots of dark red area which means the random effect is positive). The covariates show approximately 80% people voting for Trump but for some reason the actual probability is like 99% in Figure 1. This means the covariates do not tell the whole story, which confirms that there are spatial patterns that affect the probability of voting for Trump.

By checking Table 1, we could get more precise results. For example, the natural-scale posterior median for proportion of White people is 4.129, this means that holding other things constant, as the proportion of White people change from 0% to 100%, the odds of voting for Trump becomes 313% larger. For another example, the natural-scale posterior median for proportion of Indegenous people is 0.454. This means that holding other things constant, if the proportion of Indegenous people change from 0% to 100%, the odds of voting for Trump becomes 55% smaller. Since the table is in natural scale for the parameters, if the CI excludes 1, the β is considered to be significant—all three βs are significant. The CI for log population density and proportion of Indegenous people are all strictly smaller than 1 whereas the CI for proportion of White people is strictly larger than 1. This means that the increase in log population density and proportion of Indegenous people will lead to decrease in the odds of voting for Trump whereas the increase in proportion of White people will lead to increase in the odds of voting for Trump. The results in Table 1 correspond to the conclusion from Figure 1- 4 that White people who live in rural areas(less population density) tend to vote for Trump.

In terms of the log population density, it is not very interpretable. We also want to compare the effect of each of the three independent variables. Thus, we make IQR adjustments on the log-scale point estimates and quantiles, and provide a more comparable table in natural scale. Table 2 shows the effect of change in IQR for all three independent variables on odds of voting for Trump. For example, if the population density goes up by 1 IQR, the effect of IQR for population density reduces odds of voting for Trump by 22%. Similar interpretation could be applied for proportion of White and proportion of Indegeneous people. However, the effect of population density is the largest in affecting the odds of voting for Trump.

In addition, the point estimate of marginal standard deviation(θ_1) is 0.32 and the 95% CI all above 0. We could interpret it as: compared to a typical region with spatial random effect of 0, a region with 1 more spatial standard deviation has roughly $e^{0.32} - 1 \approx 38\%$ higher odds of voting for Trump. The point estimate of spatial proportion is 0.963, which suggest strong neighbor dependence, approximately 96% of the total variation is spatial. Since we have many sub-counties, the spatial proportion tend to be closer to 1. Figure 5 shows that U is quite spatially smooth, the areas with high spatial random effects are clustered together and the areas with low spatial random effects areas are clustered together, which is consistent with a big θ_2. The CI for spatial proportion excludes 0 and there exists significant spatial variation and the spatial random effect cannot be ignored. In conclusion, the information provided by Table 1 is consistent with figures: the vote for Trump in close areas are spatially correlated, and Trumpism is more like both rural and racial phenomenon.

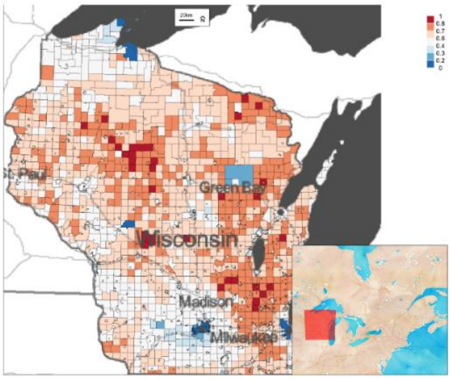

Figure 1: Probability of voting for Trump

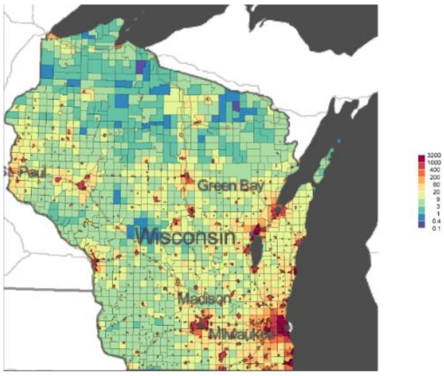

Figure 2: Population density

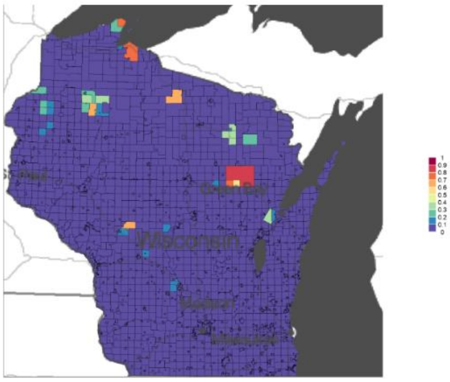

Figure 3: Proportion of Indegenous

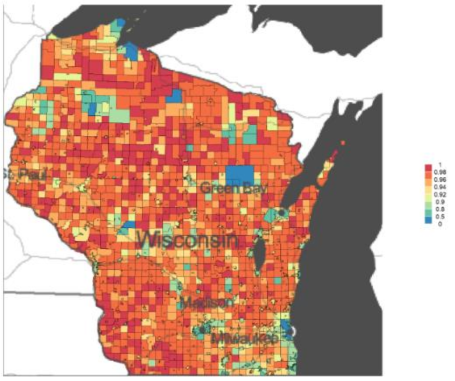

Figure 4: Proportion of White

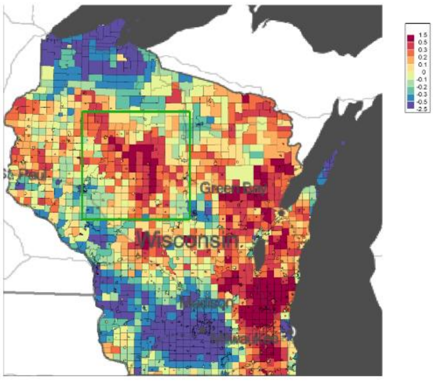

Figure 5: Random

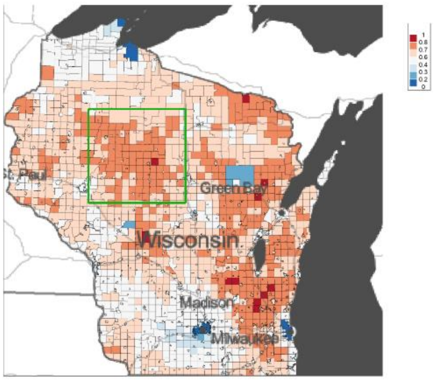

Figure 6: Fitted

Table 1: Posterior median and quantiles for odds ratios in voting Trump(natural scale), standard deviation(not in natural scale) and spatial proportion(not in natural scale) of random effect

	est	0.025 quant	0.975 quant
Intercept(natural scale)	0.57	0.438	0.744
logPdens(natural scale)	0.922	0.914	0.930
propWhite(natural scale)	4.129	3.163	5.379
propInd(natural scale)	0.454	0.322	0.640
sd(not natural scale)	0. 320	0.307	0.337
propSpatial(not natural scale)	0.963	0.924	0.987

Table 2: The comparison of effects of three independent variables on odds of voting for Trump(use IQR to make adjustments so that we could make comparisons)

	est(shows changes to odds of voting for Trump)	2.5 quant	97.5 quant
logPdens	-22.322	-24.407	-20.181
propWhite	4.846	3.918	5.775
propInd	-0.648	-0.931	-0.367