---
title: "STA442 HW4"
output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = F)
```

Donald

This report is constructed for investigating on factors that influenced the voters in Wisconsin to vote for Trump in 2016, specifically on demographic factors, including rural/urban, race etc., that may cause a strong spatial pattern in Trump support.

```{r}
load(file = "wisconsin.RData")
```

Model specification:

$$
\begin{aligned}
Y_{i} & \sim \text { Binomial }\left(N_{i}, \rho_{i}\right) \\
\log \left[\rho_{i} /\left(1-\rho_{i}\right)\right] &=\mu+X_{i} \beta+U_{i} \\
U_{i} &=\operatorname{BYM}\left(\sigma^{2}, \phi\right)
\end{aligned}
$$

Given the fact that each vote has a binary output, logistic is the most appropriate model to test out multiple factors that may influence dichotomous outputs (Yes = 1, No = 0) with binomial distribution. In the logistic model, the response variable is represented by the logarithmic of odds in Wisconsin to vote for Trump in 2016, whereas the odds is represented by the chance of voting for Trump over the chance of not voting.

Here, Y_i represents the voting data in each individual region specified by boxes in graphs below, which has a binomial distribution explained above; whereas N_i and p_i represents the total number of voters and probability of voting in each region. Moreover, X_i indicates the vector matrix for various demographic factors that we intend to test out, including “propWhite” and “propInd” as the proportion of each region which is White and Indigenous respectively; “pop”, “area”, “pdens” and “logPdens’ as the total population, surface area in square kilometers, ratio of the two, and log of the ratio; “trump” and “Total” as the number of votes for Trump and total number of votes respectively. U_i is used to indicate the spatial variation using BYM model (Bayesian Hierarchical Spatial model) for spatial population when specifying region.

Considering prior distributions, the penalized complexity prior is selected for both prior median and spatial proportion, as the intended research questions are to compare demographic factors before and after modeling that influenced the voters in Wisconsin to vote for Trump, such that:

$$
\begin{array}{l}
p r\left(\sigma > \log(2.5) \right)=0.5 \\
p r\left(\phi > 0.5\right)=0.5
\end{array}
$$

Plots using original data:

Four plots using the original data set are first displayed below, which include: (a). the proportion of voting Trump; (b). population density in square kilometers; (c). population density for indigenous group; (d). population density for white group.

Also, two additional plots using model results are displayed as well, which include: (a). random effects for spatial variation, U_i; (b). fitted value for the estimated proportion of voting Trump. The table results for desired demographic factors' effect on Trump voting is given below.

```{r}
theColTrump = mapmisc::colourScale(wisconsinCsubm$propTrump,
col = "RdBu", breaks = sort(unique(setdiff(c(0, 1, seq(0.2,
0.8, by = 0.1)), 0.5))), style = "fixed", rev = TRUE)
theColPop = mapmisc::colourScale(wisconsinCsubm$pdens, col = "Spectral",
breaks = 11, style = "equal", transform = "log", digits = 1,
rev = TRUE)
theColWhite = mapmisc::colourScale(wisconsinCsubm$propWhite,
col = "Spectral", breaks = c(0, 0.5, 0.8, 0.9, seq(0.9,
1, by = 0.02)), style = "fixed", rev = TRUE)
theColInd = mapmisc::colourScale(wisconsinCsubm$propInd,
col = "Spectral", breaks = seq(0, 1, by = 0.1), style = "fixed",
rev = TRUE)
theBg = mapmisc::tonerToTrans(mapmisc::openmap(wisconsinCm,
fact = 2, path = "stamen-toner"), col = "grey30")
theInset = mapmisc::openmap(wisconsinCm, zoom = 6, path = "stamen-watercolor",
crs = mapmisc::crsMerc, buffer = c(0, 1500, 100, 700) *
1000)
library("sp")
mapmisc::map.new(wisconsinCsubm, 0.85)
sp::plot(wisconsinCsubm, col = theColTrump$plot, add = TRUE,
lwd = 0.2)
raster::plot(theBg, add = TRUE, maxpixels = 10^7)
mapmisc::insetMap(wisconsinCsubm, "bottomright", theInset,
outer = TRUE, width = 0.35)
mapmisc::scaleBar(wisconsinCsubm, "top", cex = 0.8)
mapmisc::legendBreaks("topright", theColTrump, bty = "n",
inset = 0)
mapmisc::map.new(wisconsinCsubm, 0.85)
plot(wisconsinCsubm, col = theColPop$plot, add = TRUE, lwd = 0.2)
plot(theBg, add = TRUE, maxpixels = 10^7)
mapmisc::legendBreaks("right", theColPop, bty = "n", inset = 0)
mapmisc::map.new(wisconsinCsubm, 0.85)
plot(wisconsinCsubm, col = theColInd$plot, add = TRUE, lwd = 0.2)
plot(theBg, add = TRUE, maxpixels = 10^7)
mapmisc::legendBreaks("right", theColInd, bty = "n", inset = 0)
mapmisc::map.new(wisconsinCsubm, 0.85)
plot(wisconsinCsubm, col = theColWhite$plot, add = TRUE,
lwd = 0.2)
plot(theBg, add = TRUE, maxpixels = 10^7)
mapmisc::legendBreaks("right", theColWhite, bty = "n", inset = 0)
```

Plots using model results:

```{r}
load(file = "resWisconsin.RData")

theColRandom = mapmisc::colourScale(resTrump$data$random.mean,
col = "Spectral", breaks = 11, style = "quantile", rev = TRUE,
dec = 1)
theColFit = mapmisc::colourScale(resTrump$data$fitted.invlogit,
col = "RdBu", rev = TRUE, breaks = sort(unique(setdiff(c(0,
1, seq(0.2, 0.8, by = 0.1)), 0.5))), style = "fixed")
mapmisc::map.new(wisconsinCsubm, 0.85)
plot(resTrump$data, col = theColRandom$plot, add = TRUE,
lwd = 0.2)
plot(theBg, add = TRUE, maxpixels = 10^7)
mapmisc::legendBreaks("topright", theColRandom)
mapmisc::map.new(wisconsinCsubm, 0.85)
```

```
plot(resTrump$data, col = theColFit$plot, add = TRUE, lwd = 0.2)
plot(theBg, add = TRUE, maxpixels = 10^7)
mapmisc::legendBreaks("topright", theColFit)
```

Though observing plots above, there are three significant observations marked by red, yellow and blue. For red marks, we can notice that in urban (downtown) areas with populations density more than 3200 per square kilometers and with 0.94 to 0.96 density for white group, it shows a relatively high proportion of voting Trump, between 0.7 to 0.8. However, in regions with only indigenous group marked in yellow, the proportion of voting Trump is significantly low, between 0.1 to 0.2. And for blue marks with the highest proportion of voting Trump, it locates in urban area with relatively low population density between 9 to 20 per square kilometers and significantly high population density for white group, between 0.98 to 1. These results can also be shown in odds ratio for desired demographic factors' effect on Trump voting in the table below, which is transformed from logarithmic of odds.

Table of model results:

```{r}
knitr::kable(resTrump$parameters$summary[, paste0(c(0.5,
0.025, 0.975), "quant")], digits = 5)
```

Considering the credible intervals for all three intended variables, "pdens", "propWhite" and "propInd", all intervals do not include 1, meaning that variables are significant for interpretation. For population density, "pdens", while keeping all other variables fixed, it's odds ratio is 0. 92215, which is smaller than 1. This means that, when population density increases by 1 percent, the odds of voting Trump will increase by 0. 92215 times, indicating there is less chance of voting Trump. For population density for white, "propWhite", has an odds ratio of 4. 13212. So when population density for white increase by 1 percent, the voting Trump will drastically increase by 0. 92215 times, meaning that there is a greater chance of voting Trump. Similarly, the odds ratio for population density for indigenous is 0.45410, meaning that there is less chance of voting Trump as well. Note that "sd" indicates the part of residual that the model cannot explain, whereas "propSpatial" indicates the spatial random effect for spatial variations, which can be further shown in graphs below.

For the first plot, spatial variation (U_i) tries to explain residuals from statistics above using geographical location and environments factors, which can be quantified as spatial random effect, as large value and darker shade in the graph indicates greater variation. We can see that the spatial variation in downtown area marked in red is specifically large, whereas south regions with less population density shows less spatial variation. For the second plot, the fitted value is transformed by inverse logit to present a more precise estimate proportion of voting Trump, which is similar to what is observed in the actual proportion of voting trump in 1.(a), as there are nearly no supporters for Trump in southwest, more supporters in either southeast and northwest with an aggregation in downtown area specifically.

Therefore, through observing set of plots and statistical table, we can conclude that, demographic factors indeed place an influence in Wisconsin to vote for Trump in 2016, as there are more votes associated with urban area and area with more white group, whereas there are less votes observed for rural area and area with indigenous group.

Question 2

This report is constructed for investigating whether exposure to certain factors that make individuals more susceptible to COVID-19. Specifically, we focus on three factors in influencing COVID-19 susceptibility, ambient air pollution that puts stress on the lungs; locations with high unemployment, as such areas tend to have high deprivation and low access to health care; and ethnic minorities who are more likely to live in large, multi-generational households, work in high-risk occupations, and structural racism making access to health care harder.

Load original dataset.

```{r}
load(file = "England_shp.RData")
UK_shp$logExpected = log(UK_shp$E)
UK2 = UK_shp[grep("Wight", UK_shp$Name, invert = TRUE),]
```

Load model results.

```{r}
load(file = "englandRes.RData")
```

Table of model results.

```{r}
knitr::kable(englandRes$parameters$summary[, paste0(c(0.5,
0.025, 0.975), "quant")], digits = 5)
```

Model specification:

Since the susceptibility of COVID-19 associated with different factors follow the Poisson distribution, which is a discrete and countable distribution, a statistical model corresponding to the data set can be expressed as below.

$$
\begin{aligned}
Y_{i} & \sim \text { Poisson }\left(E_{i} \lambda_{i}\right) \\
\log \left(\lambda_{i}\right) &=\mu+X_{i} \beta+U_{i} \\
U_{i} & \sim \operatorname{BYM}\left(\sigma^{2}, \tau^{2}\right) \\
\theta_{1} &=\sqrt{\sigma^{2}+\tau^{2}} \\
\theta_{2} &=\sigma / \sqrt{\sigma^{2}+\tau^{2}}
\end{aligned}
$$

Here, Y_i represents the COVID-19 data in each individual region in graphs below, which has a Poisson distribution with rate parameter of COVID-19 incident, $λ_i$. Moreover, X_i indicates the vector matrix for various factors that we intend to test out, including “pm25modelled” as the concentrations of fine particulate matter (PM 2.5) in the health authority; “cases” and “E” as the number of COVID-19 cases up to 15 October 2020 and expected number computed from population data and known incidence rates; “Unemployment” and “Ethnicity” as the percent of individuals who are unemployed and ethnic minorities respectively. $β$ describes the log scale $λ_i$ difference, which can be accessed differently in three factors we intend to consider. U_i is used to indicate the spatial variation using BYM model (Bayesian Hierarchical Spatial model) for spatial population when specifying region, whereas $σ^2$ indicates the spatial variation and $ϕ^2$ indicates all other remaining variations in the model. Therefore, through calculations, $θ_1$ represents the total standard deviation and $θ_2$ represents the proportion of spatial standard deviation out of the total standard deviation.

Considering prior distributions, the penalized complexity prior is selected as well for both $θ_1$ and $θ_2$, meaning that there are half of the probability to be larger than the median, and the intended research questions are to compare various factors in influencing COVID-19 susceptibility, such that:

$$
\begin{array}{l}
\operatorname{pr}\left(\theta_{1}>0.5\right)=0.5 \\
\operatorname{pr}\left(\theta_{2}>0.5\right)=0.5
\end{array}
$$
Five plots using the original data set are first displayed below, which include: (a). Cases of COVID-19; (b). Expected number computed from population data and known incidence rates;

(c). concentrations of fine particulate matter measuring air pollution; (d). percentage of ethnic minorities group; (e). percentage of unemployed group.

Also, two additional plots using model results are displayed as well, which include: (a). random effects for spatial variation, U_i; (b). Poisson rate parameter of COVID-19 incident, λ_i. The table results for desired demographic factors' effect on Trump voting is given below.

```{r}
casesCol = mapmisc::colourScale(UK2$cases, dec = -3, breaks = 12,
col = "Spectral", style = "quantile", rev = TRUE)
Ecol = mapmisc::colourScale(UK2$E, breaks = casesCol$breaks,
col = casesCol$col, style = "fixed")
pmCol = mapmisc::colourScale(UK2$modelledpm25, breaks = 9,
dec = 0, style = "quantile")
ethCol = mapmisc::colourScale(UK2$Ethnicity, breaks = 9,
digits = 1, style = "quantile")
uCol = mapmisc::colourScale(UK2$Unemployment, breaks = 12,
dec = 0, style = "quantile")
rCol = mapmisc::colourScale(englandRes$data$random.mean,
breaks = 12, dec = -log10(0.25), style = "quantile")
fCol = mapmisc::colourScale(englandRes$data$fitted.exp,
breaks = 9, dec = 1, style = "quantile")
insetEngland1 = mapmisc::openmap(UK2, zoom = 3, fact = 4,
path = "waze", crs = CRS("+init=epsg:3035"))
library("raster")
insetEngland = raster::crop(insetEngland1, extend(extent(insetEngland1),
-c(25, 7, 4, 9.5) * 100 * 1000))
library("sp")
mapmisc::map.new(UK2)
mapmisc::insetMap(UK_shp, "topright", insetEngland, width = 0.4)
plot(UK2, col = casesCol$plot, add = TRUE, lwd = 0.2)
mapmisc::legendBreaks("left", casesCol, bty = "n")
mapmisc::map.new(UK2)
plot(UK2, col = Ecol$plot, add = TRUE, lwd = 0.2)
mapmisc::legendBreaks("left", casesCol, bty = "n")
mapmisc::map.new(UK2)
plot(UK2, col = pmCol$plot, add = TRUE, lwd = 0.2)
mapmisc::legendBreaks("left", pmCol, bty = "n")
mapmisc::map.new(UK2)
plot(UK2, col = ethCol$plot, add = TRUE, lwd = 0.2)
mapmisc::legendBreaks("left", ethCol, bty = "n")
mapmisc::map.new(UK2)
plot(UK2, col = uCol$plot, add = TRUE, lwd = 0.2)
mapmisc::legendBreaks("left", uCol, bty = "n")
mapmisc::map.new(UK2)
plot(UK2, col = rCol$plot, add = TRUE, lwd = 0.2)
mapmisc::legendBreaks("left", rCol, bty = "n")
mapmisc::map.new(UK2)
plot(UK2, col = fCol$plot, add = TRUE, lwd = 0.2)
mapmisc::legendBreaks("left", fCol, bty = "n")
```

Though observing plots above, we can obtain a general idea of how air pollution, unemployment and ethic group affect COVID-19 susceptibility. First, considering the hypothesis that ambient air pollution that puts stress on the lungs, which might cause more COVID-19 susceptibility potentially, we can make a comparison between graphs 2.1.(a) and 2.1.(c). Even though there is a large amount of pollution and COVID-19 cases occur in the southeast of England, we can still observe cases more than 21000 in the northwest part of English, which has barely no impact of air pollution. Also, in areas located in the center part of England, where pollution is also significant, they also have intermediate level of cases (4000-6000) and even low level of cases (2000-3000) of COVID-19. Therefore, air pollution should be unrelated with COVID-19 susceptibility.

Then considering the hypothesis that we would expect to see more COVID-19, where there is high unemployment, as such areas tend to have high deprivation and low access to health care, we can make a comparison between graphs 2.1.(a) and 2.1.(e). Even though we can't observe significant unemployment in the southeast of England, where COVID-19 shows a huge outbreak, we can still see that other regions, especially northeast, indicate a trend that high unemployment rate is related to more COVID-19 cases. Therefore, we expect an influence from unemployment toward COVID-19 susceptibility.

Lastly, considering the hypothesis that ethnic minorities who are more likely to live in large, multi-generational households, work in high-risk occupations, and structural racism making access to health care harder, we can make a comparison between graphs 2.1.(a) and 2.1.(d). We can clearly see that ethic groups aggregate in the southeast and northwest of England, which are two of the most occurring regions of COVID-19, showing more than 21000 cases. Areas in central region where there is an intermediate level of aggregation of ethic group also show a corresponding level of COVID-19 cases, 6000-8000. Therefore, we expect influence from ethnic group toward COVID-19 susceptibility too.

The model graphs 2.2.(a) and 2.2.(b) shows the spatial random effect (U_i) and COVID-19 incident rate parameter ($λ_i$). Since higher random effect is correlated with a greater spatial variation, we can observe the greatest variation in the northwest of England, which also has the highest incident rate of COVID-19. Central England also has certain variation, while variation gradually decrease in southwest, and incident rate of COVID-19 shows the same trend. Note that all these conclusions can also be shown in lambda ratio for desired factors' effect on COVID-19 susceptibility in the table below, which is transformed from logarithmic of lambda.

Considering the credible intervals for all three intended variables, "Ethnicity", "modelledpm25" and "Unemployment", only "modelledpm25" include 1, meaning that variables except "modelledpm25" are significant for interpretation, which corresponds to our result previously. For "Ethnicity", while keeping all other variables fixed, it's lambda ratio is 1.01212, which is greater than 1. This means that, when ethnicity proportion increases by 1 percent, the lambda ratio of COVID-19 susceptibility will increase by 1.01212 times, indicating there is more chance of COVID-19 susceptibility. For "Unemployment", while keeping all other variables fixed, it's lambda ratio is 1.11987, which is greater than 1 as well. This means that, when unemployment proportion increases by 1 percent, the lambda ratio of COVID-19 susceptibility will increase by 1.11987 times, indicating there is more chance of COVID-19 susceptibility. Note that "sd" indicates the part of residual that the model cannot explain, whereas "propSpatial" indicates the spatial random effect for spatial variations.

Therefore, through observing set of plots and statistical table, we can conclude that, Unemployment and Ethnicity indeed place an influence in COVID-19 susceptibility, while the condition of air pollution is considered as insignificant from results above.