Generalized Linear Geo-statistical Model

Aichen Liu

06/12/2020

Donald Duck

Introduction

In 2016, Donald J. Trump won the US election with 304 votes compared to 227 votes for Hillary Clinton. The result of the 2016 US election was once debated. The task of this project is to explore the pattern and the structure of Trump supporters in Wisconsin. We are primarily interested in finding the most important demographic factors that seem to be causing a strong spatial pattern in Trump's support. In this project, a generalized linear geostatistical model is build to explore whether supporting Trump is an urban/rural phenomenon or a racial phenomenon, especially for White and Indigenous people. Or whether there exist some spatial explanatory variables which can explain the distribution of Trump voters throughout Wisconsin.

The vote data contains some key variables: propWhite–The proportion of White people in each region. PropInd–The proportion of Indigenous people in each region. pop–The total population in each region. area–The surface area of each region. pdens(logPdens)–The(log of)population density in each region. trump–The number of votes for Trump. Total–The total number of votes.

Model

The data contains 1853 observations and 30 variables including a spatial polygons column which contains the coordinates of the boundaries of each region within Wisconsin. Since we are interested in the number of votes for Trump, the variable trump is set to be the dependent variable. Variables logPdens, propWhite, propInd are set to be the independent variables. Since each region has different characteristics, a spatial random effect is included in the model. We assume the variable trump follows a binomial distribution; here is the mathematical equation of the model.

$$Y_i \sim Binomial(N_1, \mu_i)$$

$$logit(p) = \mu + X(s)\beta + U(s)$$

where

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

$$Cov[U(s+h), U(s)] = \sigma^2 \rho(\frac{h}{\phi}; v)$$

$Priors$:

$$P(\sigma > log(2.5)) = 0.5$$

$$P(\phi > 0.5) = 0.5$$

In this model, Y_i is the number of Trump voters in region i, which follows a binomial distribution, N_i is the number of voters in region i, $logit(p)$ represents the log of odds, it is the log of odds of voting Trump, $X(s)$

is a vector of covariates at different regions including logPdens, propWhite, propInd, βs are the parameters. And $U(s)$ is the spatial random effect that is also called the residual spatial variation that represents the difference between actual probability and what the covariates predict. it follows a Besag, York and Mollie model. $U(s)$ depends on three parameters:$\sigma, \phi, andv.\sigma$ is the variability in the residual variation, *phi* is the range parameter, and v is the shape parameter. Base on existing knowledge, we know that the prior distribution of σ and *phi* has a prior median of log(2.5) and 0.5 respectively. This model is sensible because we hypothesize that there might be some differences among different regions. However, the entries in the variable trump are all positive integers, a Poisson model would be more appropriate.

Result

After building the model, we can generate a summary table with the predicted medians and their credible intervals in Figure 1. The first column is the odds ratio and the second and third columns are the corresponding 95% credible intervals. If one is in the range of the credible interval, then the corresponding variable is not statistically significant. As indicated in figure 1, all the credible intervals don't include 1, which means all the variables are significant. If we focus on rows, we observe that the odds ratio for logPdens is 0.922. This implies that for every unit increase in logPdens, the likelihood of voting for Trump become 0.922 times the original probability. Namely, for every unit increase in the population density, the probability of voting for Trump decreases by 0.078 (1-0.922). Since rural areas usually have lower population density and urban areas have higher population density, we conclude that people in urban areas are less likely to vote for Trump compared with those in rural areas. If we consider racial factors, we can observe that the odd ratio of variable propWhite is 4.132. This means that every unit increase in the proportion of the White leads to a 3.132 (4.132-1) increase in the probability of voting for Trump. In contrast, the odds ratio for propInd is 0.454. It means that if the proportion of Indigenous increases by one unit, the probability of voting for Trump would decrease by 0.546 (1-0.454) which is more than half. Thus we know that White people are more likely to vote for Trump compare with Indigenous people. The last two rows of the summary table provide the estimation of σ and the spatial range parameter. As indicated in Figure 1, both of them are significant. Hence there exists some spatial variation for each region. To further verify the existence of spatial variation, some maps are generated to indicate the population density and the proportion of different ethnicities.

Figure 1: Odd Ratio and 95% Credible Intervals–Natural Scale

	0.5quant	0.025quant	0.975quant
(Intercept)	0.570	0.437	0.743
logPdens	0.922	0.914	0.930
propWhite	4.132	3.166	5.382
propInd	0.454	0.322	0.640
sd	1.375	1.356	1.397
propSpatial	2.612	2.502	2.680

Map A indicates the proportion of Trump voters in different regions, the closer the colour to red, the higher the proportion of voting Trump, the closer the colour to blue, the lower the proportion of voting Trump. Map B, C, D each indicates the population density, proportion of Indigenous people and proportion of White people. Map E illustrates the spatial random effect and the predicted proportion of trump voters. Map F is the predicted number of Trump voters. The distribution in map F is similar to map A. Therefore the observed data is similar to the predicted values.

In map B we observed that the population density is higher around the right side of the border, and the corresponding areas in map A appear in the colour white or blue. It implies that the proportion of Trump supporters in high-population-density areas is low. On the other hand, for those areas that appear in blue or green in map B, the corresponding proportion of Trump supports are high. These arguments further

strengthen the conclusion we made that Trumpism is a rural phenomenon. People in rural areas are more likely to vote for Trump than those in urban areas.

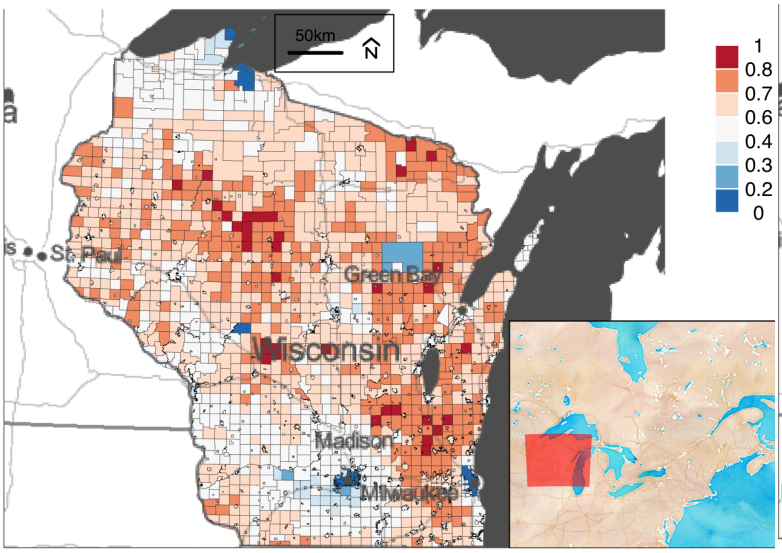
Map C shows that the majority of Indigenous people gather around Green Bay, and others are around the upper part of Wisconsin. It is obvious that the corresponding areas in map A are dark blue, which means that the majority of the Indigenous people are not Trump supports. On the other hand, map D shows the distribution of White people in Wisconsin. The distribution of Map D is very similar to Map A. Red regions in map D are also appear to be red in map A. Thus, Trumpism is also a racial phenomenon. White people are more likely to support Trump compares with Indigenous people.

Map E shows the spatial random effect. The colour appears to be blue if the random effect is negative and red if it is positive. As indicated, the random effect is negative around the bottom left corner and the upper part of Wisconsin, whereas it is positive around the bottom right corner and the middle part of Wisconsin. The differences between regions are significant, therefore the spatial random effect is significant as well. These maps proved that the spatial variation with Trump voters is large throughout Wisconsin. Since there exists a large spatial variation throughout Wisconsin, there may be some other explanatory variables that are also important to predict the support of Trump, such as income and educations that we can test in the next step.

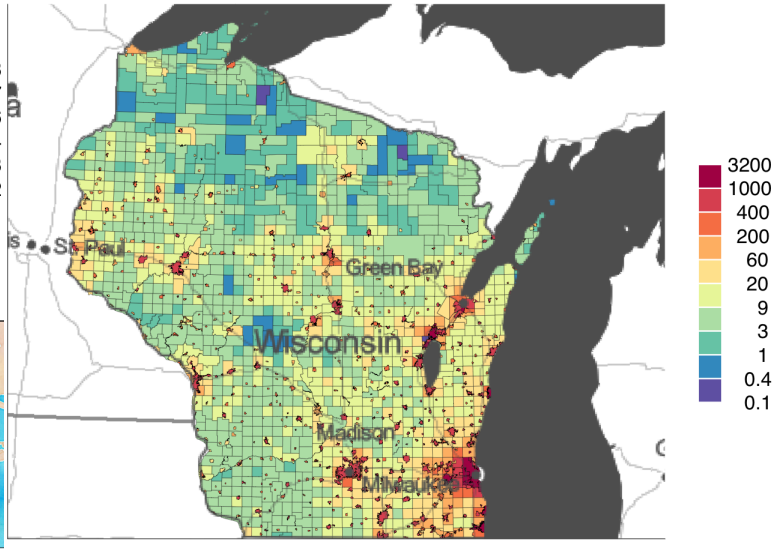
Summary

To sum up, the spatial variation is large within Wisconsin. Some important demographic factors that causing a strong spatial pattern in Trump's support are population density and ethnicities. Base on the results from the model and the maps, we can conclude that Trumpism is a rural phenomenon. People who live in rural areas are more likely to vote for Trump than people who live in urban areas. On the other hand, Trumpism is also a racial phenomenon. White people are more likely to vote for Trump compare with Indigenous people.

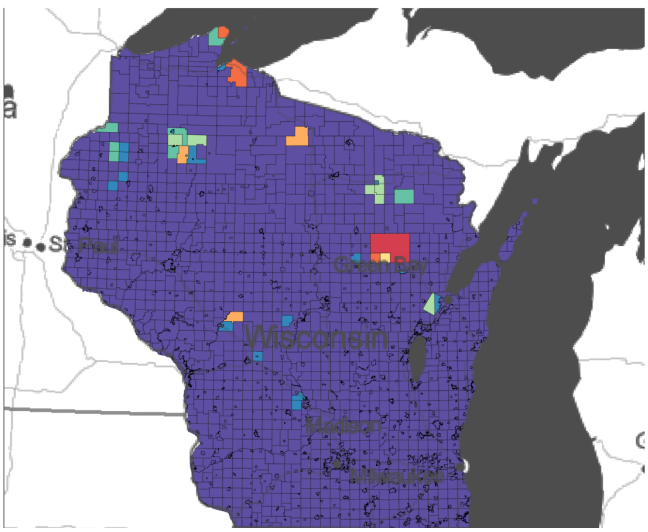
Map A



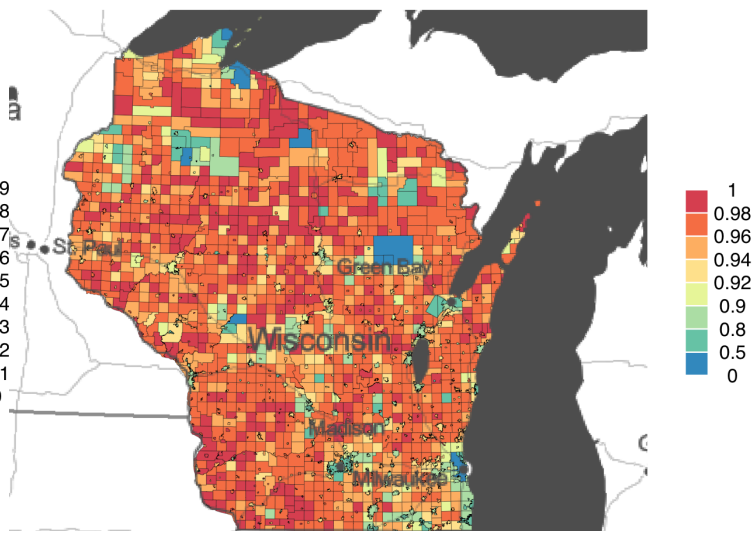
Map B



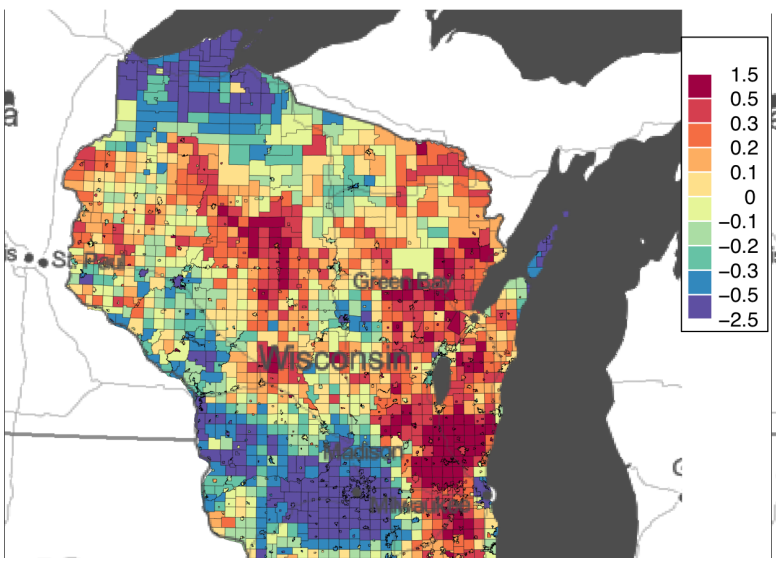
Map C



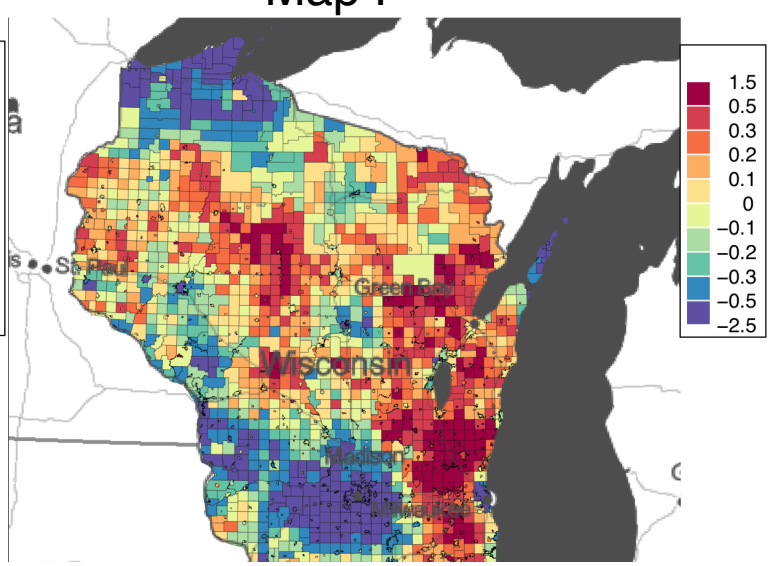
Map D



Map E



Map F



COVID-19 in England

Introduction

COVID-19, a new infectious disease, causes respiratory afflictions that have affected people all around the world. During the research of this new virus, some scientists believe that air pollution (PM2.5) is an important factor that affects the number of local cases because exposure to ambient air pollution makes individuals more susceptible to COVID-19. On the other hand, some believe that more COVID-19 cases are expected to see in regions where there is a high unemployment rate since such areas tend to have high deprivation and low access to health care. But others believe that more COVID-19 cases should be expected in regions with many ethnic minorities since ethnic minorities are more likely to live in large, multi-generational households and work in high-risk occupations. The objective of this research is to build a generalized linear geostatistical model and test whether the above hypotheses are true.

The dataset from public health in England is used to address these hypotheses. The data contains 149 observations and 76 variables. Each observation in the dataset represents a public health region in England. Some key variables in this dataset include: pm25modelled–The concentration of fine particulate matter PM2.5 in the health authority cases. E– the number of COVID-19 cases up to 15 October 2020 and expected number. Unemployment – the percentage of unemployed individuals. Ethnicity– is the percentage of individuals who are ethnic minorities.

Model

The data contains a spatial polygons column which contains the coordinates of the boundaries of each region within England. Since we are interested in the number of cases in England, the variable cases are set to be the dependent variable. Variables pm25modelled, Ethnicity, Unemployment is set to be the independent variables. Since the variance of this model is greater than its mean, an offset (logExpected) is added to control the overdispersion. We believe that each region has different characteristics, a spatial random effect is included in the model. Since the number of cases must be a positive integer, thus the dependent variable follows a Poisson distribution. The mathematical equation of the model is stated below.

$$Y_i \sim Poisson(E_i\lambda_i)$$

$$log(\lambda_i) = \mu + X(s)\beta + U(s)$$

where

$$U(s) = W_i + V_i$$

$$V_i \sim i.i.d.\ N(0, \tau^2)$$

$$W_i \mid (W_j; j \neq i) \sim N(mean(W_j; j \sim i), \sigma^2/ \mid j \sim i \mid)$$

if

$$\sigma > \tau : U\ is\ smooth$$

$$\sigma < \tau : U\ is\ rough$$

In this model, Y_i is the number of cases of COVID-19 in region i which follows a Poisson distribution. E_i is the offset term that represents the expected count in region i. $X(s)$ is a vector of covariates at different regions including pm25modelled, Ethnicity and Unemployment. βs are the parameters. $U(s)$ is the spatial random effect. In the Poisson linear spatial model, the spatial random effect can be divided into two parts; W_i, an improper GMRF and V_i an independent noise. $U(s)$ depends on three parameters:$\sigma, \phi, andv.\sigma$ is the variability in the residual variation, *phi* is the range parameter, and v is the shape parameter. Base on existing knowledge, we know that the prior distribution of σ and *phi* both have a prior median of 0.5.

Result

After building the model, a summary table is generated in figure 3. The first column represents the posterior means of the parameters and the second and third column are the corresponding 95% credible intervals. If one is in the range of the credible interval, then the corresponding variable is not statistically significant. All credible intervals don't contain one, except modelledpm25's. It implies that the conclusion we made about the relationship between the concentration of PM2.5 and the number of cases of COVID-19 is not significant. Besides, the posterior mean for variable Ethnicity is 1.012, which means for every unit increases of Ethnicity, the number of COVID-19 cases increases by 0.012 (1.012-1). Namely, if the proportion of people who are ethnic minorities increases by one, the number of cases increases by 0.012. Hence there exist a positive relationship between the proportion of ethnic minority people and the number of COVID-19 cases. Similarly, the posterior mean for variable Unemployment is 1.12. It implies that for every unit increase in the proportion of unemployed people, the number of COVID-19 cases increases by 0.12 (1.12-1). Thus, we would expect to see more COVID-19 where there is a high unemployment rate. The last two rows of the summary table provide the estimation of σ and the spatial range parameterϕ. They are 1.342 and 2.455. Since their credible intervals don't contain one, both of them are significant. Hence there exists some spatial variation for each region. Some maps are generated to further verify the spatial variation of each variable.

Figure 3: Parameter Posterior Means and 95% Credible Interval–Natural Scale

	mean	0.025quant	0.975quant
(Intercept)	0.365	0.218	0.610
Ethnicity	1.012	1.008	1.016
modelledpm25	1.057	0.996	1.123
Unemployment	1.120	1.059	1.184
sd	1.342	1.295	1.399
propSpatial	2.455	2.155	2.652

The following seven maps illustrate the variables in each region. Map G shows the number of COVID-19 cases throughout England. If the colour is close to red, then the number of cases is large. If the colour is close to blue or green, then the number of cases is small. Similarly, map H shows the expected cases of COVID-19. Comparing map G and map H, we found that the expected number of cases is more than the observed number of cases in the right-border regions. The reality is better than what we expected.

Furthermore, map I shows the concentration of PM2.5 in different regions. As indicated, the concentration of PM2.5 is the highest at the right corner. The closer to the border, the lower the concentration. Comparing map I with map G, we can see that some areas with a high concentration level have a low number of COVID-19 cases. Whereas at the top of the map, some areas with a low concentration level, have a high concentration level of COVID-19. The pattern is ambiguous. Therefore the relationship between the concentration of PM2.5 and the number of cases of COVID-19 is not significant.

Map J shows the proportion of the ethnic minority people in each region. The majority of the ethnic minority people gathered around the right corner and the middle left part of England. Comparing map D and map G, some areas where the proportion of the ethnic minority people is high have a low number of COVID-19 cases. However, some are not. Furthermore, some areas with a large number of COVID-19 cases have no ethnic minority people at all. Above in the summary table, we stated that there exists a positive relationship between the proportion of ethnic minority people and the number of COVID-19 cases. However, based on the map, a contradiction is observed.

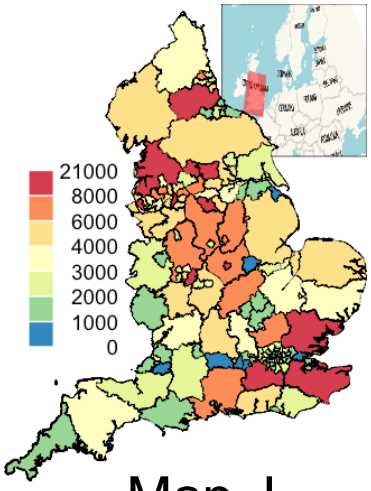
The distribution of the unemployment rate is illustrated in map K. As indicated, the unemployment rate is higher in the middle and the upper parts. The corresponding areas in map G also have a larger number of COVID-19 cases. Thus, we would expect to see more COVID-19 where there is a high unemployment rate.

Map L is the spatial random effect. Colour red means the random effect is positive and yellow means it is negative. As indicated, the upper part of England all have positive random effects, but the majority of the lower regions have small or negative random effects. The differences between regions are significant therefore the spatial random effect is significant as well. The spatial variation with the number of cases throughout England is large.

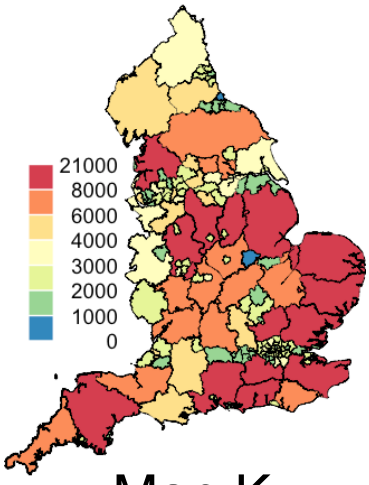
Summary

After building a well-constructed model, we now have a broad understanding of the pattern between some demographic factors and the number of COVID-19 cases. Base on data from public health England, the result shows that being exposed to ambient air polluted environment doesn't make individuals more susceptible to COVID-19. The factor that affects the infection is unemployment. Regions, where there is a high unemployment rate tend to have low access to health care. Therefore, areas with high unemployment rates tend to have more cases of COVID-19. On the other hand, there may be a positive relationship between the proportion of ethnic minority people, but we can't be sure because we observed contradictory results from the model and the map. Further analysis is required.

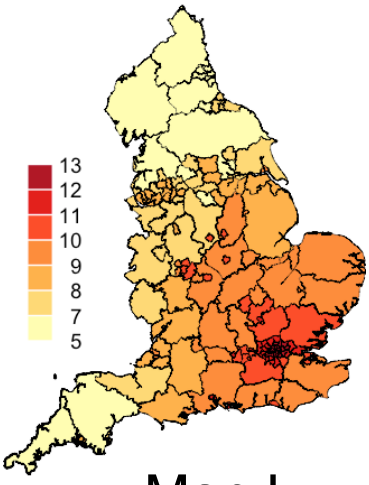
Map G



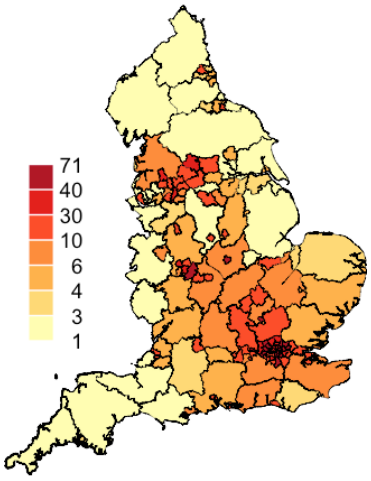
Map H



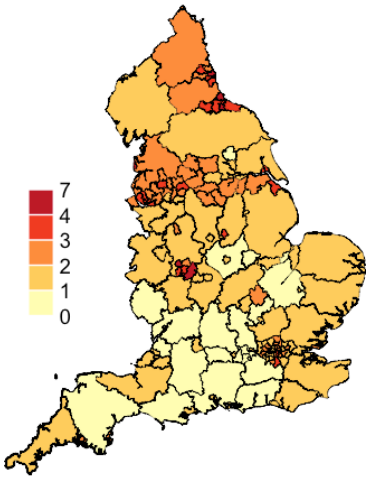
Map I



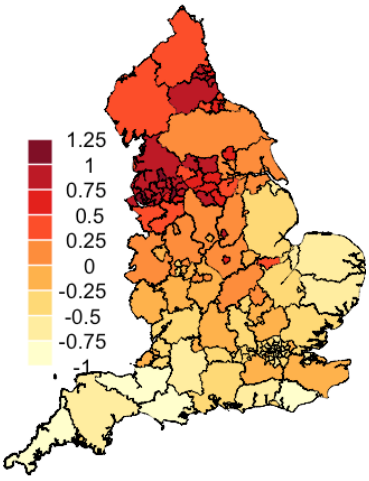
Map J



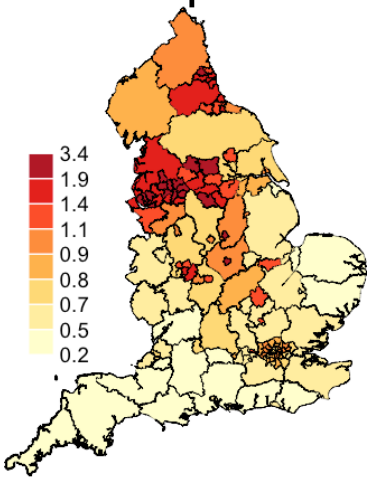
Map K



Map L



Map M



```{r,echo=FALSE, results="hide",message=FALSE}
(load("wisconsin.RData"))
(load("resWisconsin.RData"))
```
```{r,echo=FALSE, results="hide",message=FALSE}
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)
```
```{r,echo=FALSE}
knitr::kable(exp(resTrump$parameters$summary[, paste0(c(0.5,
0.025, 0.975), "quant")]), digits = 3,caption  = "Figure 1: Odd Ratio and 95% Credible Intervals--Natural Scale")
```
```{r,echo=FALSE}
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)
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)
```
```{r,echo=FALSE, results="hide",message=FALSE}
(load("England_shp.RData"))
(load("englandRes.RData"))
```
```{r,echo=FALSE}
knitr::kable(exp(englandRes$parameters$summary[, c(1,3,5)]), digits = 3,caption  = "Figure 3: Parameter Posterior Means and 95% Credible Interval--Natural Scale")
```