Homework 4: Two reasonably non-standard problems

Ging_Chen

13/12/2020

Question 1 Donald Duck

Background

In this report, we are going to explore the factors that influenced the voters in Wisconsin to vote for Trump in 2016. By using the provided Wisconsin.RData, a binomial model is built to explore the following question:

1. What are the most important demographic factors which seem to be causing a strong spatial pattern in trump support?
2. Is Trumpism a primarily urban or rural phenomenon?
3. Or is it rather a racial phenomenon with Trump appealing to White voters?
4. Is there other spatial explanatory variable? Or is there really very little spatial variation, with Trump voters being evenly distributed throughout Wisconsin?

First, we will explain the variables that play an important role in building the model. Then we will discuss the prior distributions, spatial correlation of the model in detail. In the result section, we will analyze the parameter table as well as the graphs.

Data

The model will be based on three important variables from the `wisconsinCsubm` data, which contains the sub-county election results from the `Wisconsin.RData` file. First, `trump` represents the number of votes for Trump, which is also the dependent variable in the model. Then, `propWhite` and `propInd` are the proportion of each region which is White and Indigenous, respectively. Lastly, `logPdens` is a variable calculated by taking the log of the ratio of `pop` (the total population) and `area` (the surface area measured in km^2).

Model

Consider the following model:

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

$$log(\frac{p_i}{1-p_i}) \sim \mu + X_i\beta + U(s)$$

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

- Y_i is the dependent variable, which is the number of votes for Trump. Y_i follows a Binomial distribution that has trial size of N_i; in our case, it is the total number of votes in the sub-county. p_i represents the probability of voting for Trump.
- X_i is a vector of the dependent variables, which includes `logPdens`, `popWhite`, and `popInd`, as described in the Data section of this report.
- Spatial random effect (also called the residual spatial variation) is included in the model as $U(s)$, where U follows a Besag, York and Mollié model. We will focus on specifically the correlation of U:

$$cov[U(s+h), U(s)] = \sigma^2 \rho(h/\phi; v)$$
$$\rho(h/\phi; v) = exp(-\alpha h)$$

- $U(s)$ depends on three important variables:
  - σ, which is the variability in residual variation, which is the variance in U.
  - ϕ the range parameter
  - v the shape parameter
- The correlation of U can be described as an exponential decay function, where as distance between two location increases, the correlation between them decreases. If we put this concept into context, we can easily see that two distant neighborhoods will less likely to have the same political preference compared to two adjacent neighborhoods. Here's when ϕ the range parameter comes in: if ϕ is small, the co-variance falls quickly.
- $\alpha = \frac{1}{\phi}$, which is a scale parameter.

prior

There are two variables that follow a prior distribution that we have included in our model: `sd` the variability in residual variation U, and `propSpatial` the range parameter of U. In R the prior distributions are:

```
prior = list(sd = c(log(2.5), 0.5), propSpatial = c(0.5, 0.5))
```

- First, we can see $p(\sigma > log2.5) = 0.5$, where $log2.5$ is the median of the exponential distribution. This implies that when U increases by one standard deviation, the odd ratio is $e^{log2.5} = 2$. To elaborate, when $U = 0$, let it have an odd of $odd1$; when $U = \sigma$, it has an odd of $odd2$. Then the ratio between the two is: $\frac{odd2}{odd1} = e^{log2.5} = 2$.
- Similarly, we can see that `propSpatial` follows a prior distribution where $p(\sigma > 0.5) = 0.5$, and that 0.5 is the median of the distribution.

Result and Dicussion

What are odds and odds ratio from the parameter table

Variable	0.5 quant	exp.Est	0.025 quant	0.975 quant
Intercept	-0.56276	1.75551	-0.82716	-0.29674
logPdens	-0.08105	0.92215	-0.08979	-0.07232
propWhite	1.41879	4.13212	1.15241	1.68307
propInd	-0.78943	0.45410	-1.13430	-0.44628
sd	0.31830	1.37479	0.30419	0.33446
propSpatial	0.96016	2.61211	0.91715	0.98591

From our model:

$$log(\frac{p_i}{1-p_i}) \sim \mu + X_i\beta + U(s)$$

We know that when `logPdens` increase by one unit, the odds of voting for Trump will decrease by a factor of 0.92215 providing that other factors are unchanged. In other words, the population density and the preference to vote for Trump has an inverse relation. This is based on the following calculation:

Let `logPdens` be x, then : $odd1 = \frac{p}{1-p} = e^{\beta_0+\beta_1(x)}$

Let `logPdens` increase by 1, x + 1, then: $odd2 = \frac{p}{1-p} = e^{\beta_0+\beta_1(x+1)}$

Then: $\frac{odd2}{odd1} = e^{\beta_1} = e^{-0.08105} = 0.92215$

This phenomenon can be seen from the graphs below: From the Figure 1, we see that places with dark red, which represents high support for Trump, corresponds with the greem/lemon areas (low density areas) from the Figure 2. From the graphs and the statistical estimates above, we see that Trumpism is primarily a rural phenomenon.

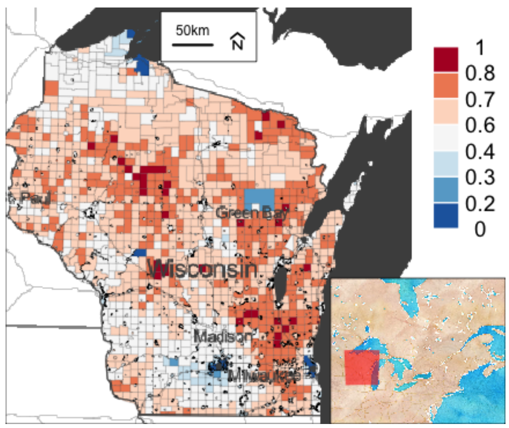

Figure 1: Support for Trump

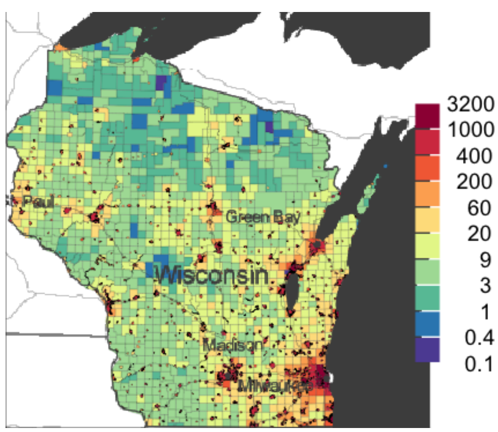

Figure 2: Population Density

Next we will look at how ethnicity affects the preference to vote for Trump. From the parameter table, we see that as the proportion of White in the area increases, the probability of voting for Trump increases by a factor of 4.13212. Note that this conclusion can be calculated in the same way as shown before with the `logPdens` case. On the other hand, as the proportion of Indigenous increases, the probability of voting for

Trump decreases by a factor of 0.45410. From these numbers, we can conclude that Trumpism is also a racial phenomenon appealing to mostly White voters. This racial phenomenon can also be visualized through the Figure 3: High proportion of White residence area corresponds with high support for Trump. Similarly, a high density of Indigenous population reflect a very low level of support for Trump, as see in Figure 4.

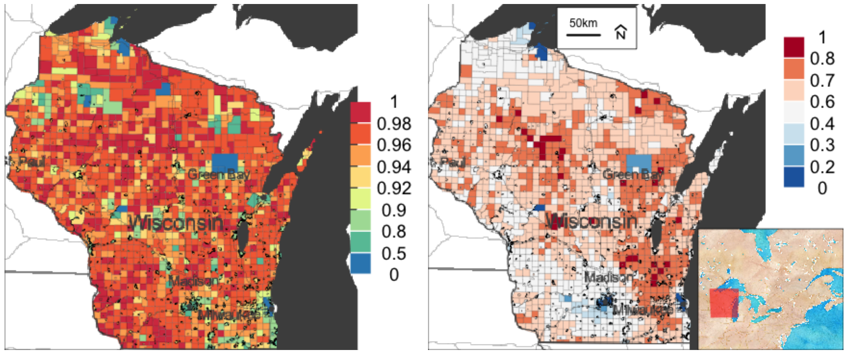

Figure 3: High White population area reflect a high level of support for Trump

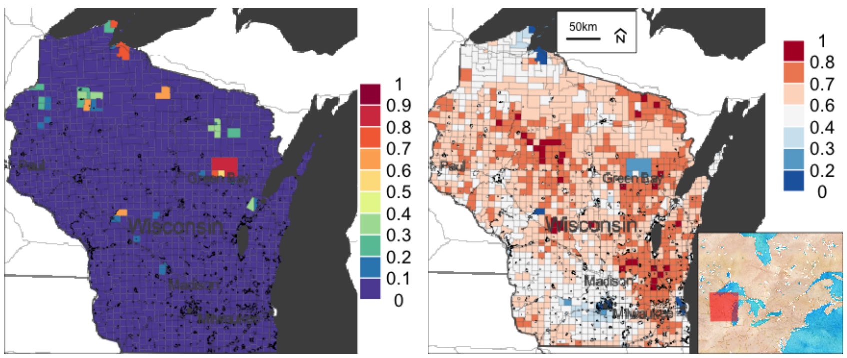

Figure 4: Indegenous population area reflect a very low level of support for Trump

Lastly, `propSpatial` which represents the range parameter, has an exponential estimate of 2.61211, which implies that there isn't a very little spatial variation; Trump voters are not evenly distributed. Moreover, `sd` has an exp. estimate of 1.37479. These two parameters make the correlation between areas very small; as distance between two arbitrary location increases, co-variance falls quickly. This can be seen from Figure 5. In the fitted distribution, we see the distribution for Trump support based on the fixed independent variables: `logPdens`, `popWhite`, and `popInd`. We again see the same trend as shown in Figure 1, 2, 3 and 4.

Conclusion

From the Geo-statistic model we can see that there is quite a spatial variation within Wisconsin. Ethnicity is an important demographic factor which seem to be causing a strong spatial patter in Trump support. The level of population density also plays a role in affecting the level of support for Trump. In conclusion, the Trump voters in Wisconsin is not evenly distributed because of demographic factors including ethnicity and the population density.

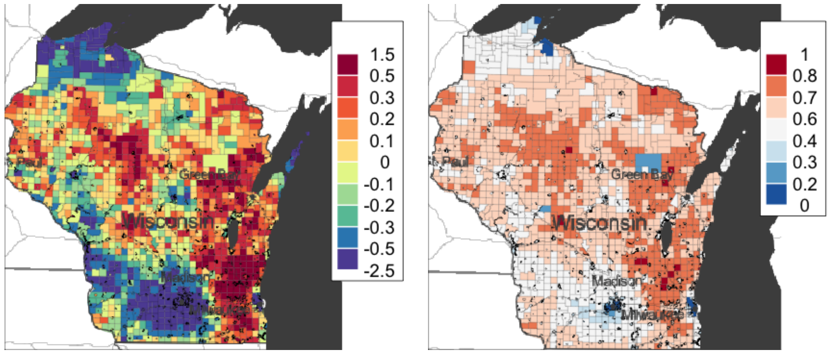

Figure 5: Random Effects (left) and Fitted (right)

Question 2 COVID-19 in England

Background

This section of the report is going to discuss the whether exposure to ammbient air pollution makes individuals more susceptible to COVID-19. We will also look at other factors such as ethnicity and rate of unemployment that influence individuals to be more susceptible to COVID-19.

Consider the following hypotheses:

- Main Hypothesis: air pollution puts stress on the lungs and respiratory tract. Therefore, it should be expected that there are more COVID-19 cases where air pollution is high.
- Hypothesis 2: we would expect to see more COVID-19 cases where there is high unemployment, as such areas tend to have high deprivation and low access to health care.
- Hypothesis 3: areas with many ethnic minorities have more COVID-19 cases because they are more likely to live in large households and work in high-risk occupations. Moreover, structural racism makes it challenging for the minorities to access health care.

In the Variable section, we will explain the variables that play an important role in building the model. Then we will discuss the prior distributions, spatial correlation of the model in detail. In the result section, we will analyze the parameter table as well as the graphs.

Data

From the Data `England_shp.RData` we obtain the data set for each public health region in England. Variables that we will be using for the model are as follows:

- `pm25modelled` concentrations of fine particulate matter (PM 2.5) in the healthy authority
- `cases, E` number of COVID-19 cases up to 15 October 2020 and expected number (computed from population data and known incidence rates)
- `Unemployment` percent of individuals who are unemployed
- `Ethnicity` percent of individuals who are ethnic minorities

Model

Consider the following model

$$Y_i \sim Poisson(E_i\lambda_i)$$
$$log(\lambda_i) = \mu + X_i\beta + U_i$$
$$U_i = W_i + V_i$$
$$V_i \sim i.i.d.N(0, \tau^2)$$

- Y_i is the dependent variable, in this context, Y_i represents the number of COVID-19 cases up to 15 October 2020 across public health region in England. Y_i follows a *Poission* distribution that has two parameters: E_i and λ_i. Where E_i is the expected count of COVID-19 cases in the region.
- X_i is a vector of the dependent variables, which includes `Ethnicity`, `modelledpm25`, and `Unemployment`, as described in the Data section of this report.
- U_i represents the spatial random effect that contains both the saptial correlation W_i and over-dispersion offset term V_i
- V_i follows a normal distribution with variance of τ^2.

prior

The prior included in this model is coded as follows:

```
prior = list(sd = c(0.5, 0.5), propSpatial = c(0.5, 0.5))
```

- There are two variables that follow a prior distribution that we have included in our model: `sd` the variability in residual variation W, and `propSpatial` the range parameter of W.
- We can see that both `sd` and `propSpatial` follows a prior distribution where $p(\sigma > 0.5) = 0.5$, and $p(\phi > 0.5) = 0.5$, and that 0.5 is the median of the distribution.

Result and Discussion

The parameter table

Variable	mean	Exp.Mean	0.025 quant	0.975 quant
Intercept	-1.00752	0.36512	-1.52329	-0.49379
Ethnicity	0.01205	1.01212	0.008097	0.01600
Modelledpm25	0.05578	1.05737	-0.0044	0.11611
Unemployment	0.11321	1.11987	0.057647	0.168733
sd	0.29402	1.34181	0.25872	0.33554
propSpatial	0.89801	2.45471	0.76785	0.97547

From the parameter table, we can see that as `Modeledpm25` increases by one unit, $E(y)$ increases by a factor of $e^{0.05578} = 1.05737$. However, such relation between the concentrations of fine particulate matter and the expected number of COVID-19 cases is not statistically significant, as shown from the parameter table (CI includes 0). When we observe the geographic distribution of air pollution and the density of COVID-19 cases, we see little correlation between the two (Figure 6).

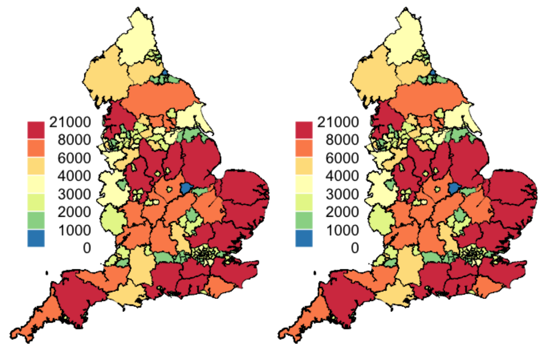

Figure 6: Expected COVID-19 cases (left) and air poluttion severity (right)

Does the rate of unemployment affect the number of cases of COVID-19? From the parameter table, we can see that the rate of unemployment affects positively to the growth of COVID-19 cases: as the rate of Unemployment increases by one unit, the expected number of COVID-19 cases increases by a factor of 1.11987. This phenomenon can be visualized in Figure 7, where the areas with medium to high unemployment rate corresponds with the areas with high density of COVID-19 cases.

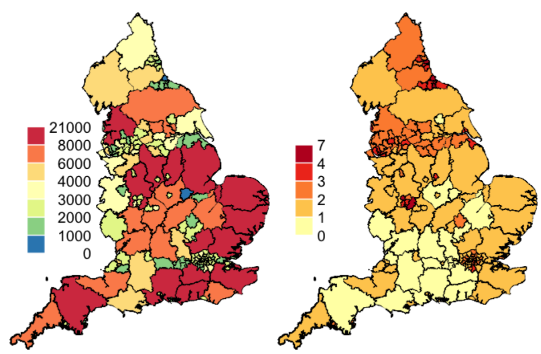

Figure 7: Expected COVID-19 cases (left) and rate of unemployment (right)

Similarly, as the percent of individuals who are ethnic minorities increases, the expected number of COVID-19 cases increases by a factor of 1.01212. This implies that the minority groups are more likely to live in large households and work in high-risk occupations. In addition, structural racism makes it challenging for the minorities to access health care. This unfortunate phenomenon can be seen by comparing the density of COVID-19 case counts with the distribution of minority groups (Figure 8).

Lastly, ϕ which is the range parameter is relatively small (2.45471) in this model. When the range parameter is small, we know that co-variance between location falls quickly. This implies that the correlation between individuals are very small in general. When we observe the graphs between expected number of COVID-19 cases and the random effect distribution, we see that places with high spatial correlation will also also result in higher count of COVID-19 cases (Figure 9). That being said, overall the spatial correlation is low.

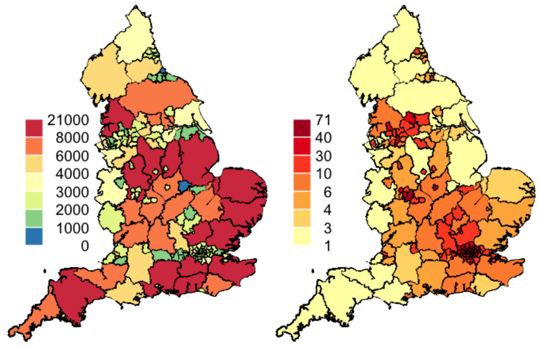

Figure 8: Expected COVID-19 cases (left) and distribution of minority (right)

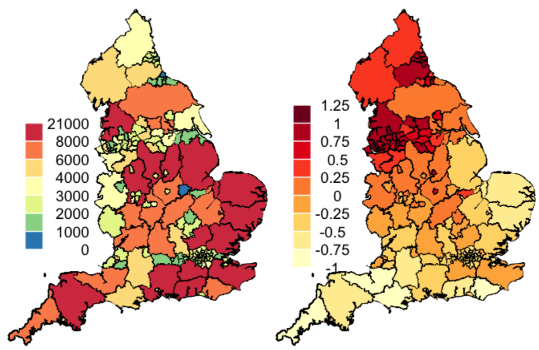

Figure 9: Expected COVID-19 cases (left) and spatial random effects (right)

Conclusion

Based on the `England_shp.RData` and the *Poisson* model, we see that

- Air pollution puts stress on the lungs and respiratory tract, but it is not necessary that there are more COVID-19 cases where air pollution is high.
- There are more COVID-19 cases where there is high unemployment; and
- Areas with many ethnic minorities have more COVID-19.