University of Missouri

STAT 7610

Project 2

Areal data analysis

Souvik Bag

14421334

Introduction

In this project, we are given county wise Covid-19 data for the state of Missouri. First we calculated the per capita testing rate, vaccine rate and positive test rate. We further converted these estimates in ′1000 scale for ease of our work.

Next we plot the data to check if there exists any spatial dependence or not for the three above mentioned variables.

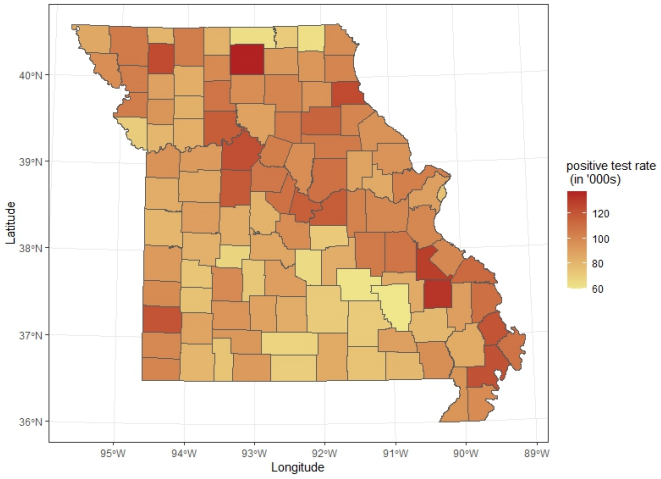

Figure 1: Per capita positive test rate in Missouri

Above plot is the county wise positive test rate plot in Missouri in the year of 2021 till the month of April. We can see some kind of spatial dependence in the data in the central and south eastern part of Columbia.

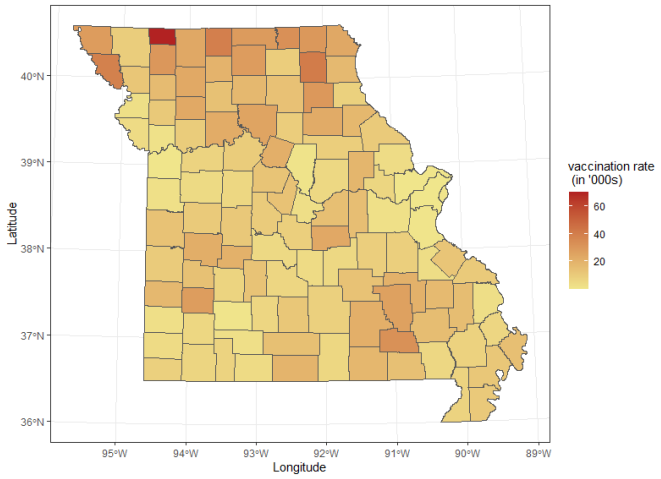

Figure 2: Per capita vaccination rate test rate in Missouri

Next we see the plot for the per capita vaccination rate. Here to our surprising, we do not see much clear picture of spatial dependence as the vaccination rate itself is very low throughout. In

this case a statistical test would be appropriate.

Q_1 : Test for spatial dependence

Now we may want to statistically check the existence of spatial dependence in our data. In this case we will use **Moran's I** statistic.

Before proceeding with Moran's I, we need to find the spatial neighbours of each county. Now it is logical to use a distance based approach to find the neighbours because we know that Covid-19 spreads within close proximity. We take 70km ($\approx$ 44 Miles) as our threshold. During covid, under lockdown situation people could only travel to nearby places so 44 miles or nearby county was a natural choice.

Next we plot the neighbours to see if our threshold value make sense or not.

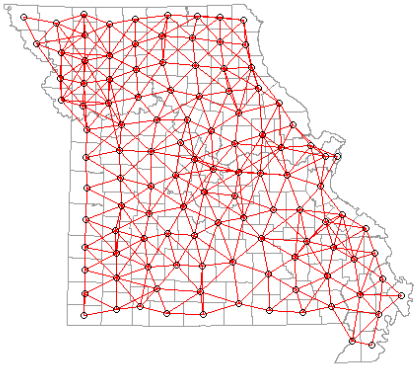

Figure 3: Neighbours of each county in Missouri

From the plot it seems to use the threshold of 70 Km as logical. As we defined the neighbours, next we assign a weight matrix corresponding to each connection band of this neighbourhood plot. We will use row-standardized weighting technique as for the covid data set they make more sense.

Now for the vaccination rate we get Moran's I = 0.332 with p-value = 4.713×10^{-12}. As the p-value is less than 0.05 we reject our null hypothesis $H_0 : I = 0$ or H_0 : There is no spatial dependence in the data and conclude that there are some positive spatial dependence in the data. That means when one county has higher vaccination rates that affects the people from neighbouring counties to get vaccinated more. It may be an affect of similar kind of governmental,

NGO and media specific advertisement in support of vaccines. Similarly, for the positivity rate, the Moran's i statistic is 0.236 with p-value $= 4.75 \times 10^{-7}$. So we reject our null hypothesis and conclude that positivity rate has positive spatial dependence. In the end, for the test rate, the Moran's i statistic is 0.207 with p-value $= 8.07 \times 10^{-6}$. So we reject our null hypothesis and conclude that test rate has positive spatial dependence.

Q_2 : Spatial areal regression

Now we are interested if the variables are spatially dependent or not after removing the first order/mean effect. To show this, we take "Positivity rate" as our response variable. Now as a natural choice, "Test rate", "Vaccination rate", "Population", "Income", "House value", "Percentage of old people with age more than 65" comes to mind as explanatory variables. If the "Test rate" increases then "Positivity rate" is expected to increase. Again, as the rate of vaccination goes up, the rate of positive results starts going down as the effect of vaccines. Similarly income and house value shows about the socio-economic condition of the area and people with good income tend to get their shots earlier because of supply chain shortage and high initial price.

So our model is :

$$PR = \beta_0 + \beta_1 \times TR + \beta_2 \times \text{VR} + \beta_3 \times \text{Popultn} + \beta_4 \times \text{Income} + \beta_5 \times \text{hoval} + \beta_6 \times \text{Senior}$$

Where PR is Positivity rate, TR : Test rate, VR : Vaccination rate, Popultn : Population, hoval : House value and Senior : Percentage of old people with age more than 65.

The summary of the fitted model :

Table 1: Summary of linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	71.5661	12.7578	5.61	0.0000
hoval	-0.0001	0.0001	-1.75	0.0835
test_rate	0.0245	0.0048	5.15	0.0000
vac_rate	-0.4519	0.1674	-2.70	0.0080
Income	0.0005	0.0003	2.05	0.0431
Popultn	-0.0000	0.0000	-1.04	0.3017
senior	-0.1507	0.3628	-0.42	0.6787

We can see that (Table 1) Population and senior are statistically insignificant at 0.05% level of significance. House value is significant only at 8% level of significance. This result is quite surprising because if a county has more population then they should have more tests and sub-

sequently more positive rate which was not supported by the results from our model. Moreover "Senior" variable which represents the percentage of people older than 65 is also not significant. It raise questions on the fact that older people are more likely to get Covid.

Now we conduct a residual analysis to check if our model assumptions are met.

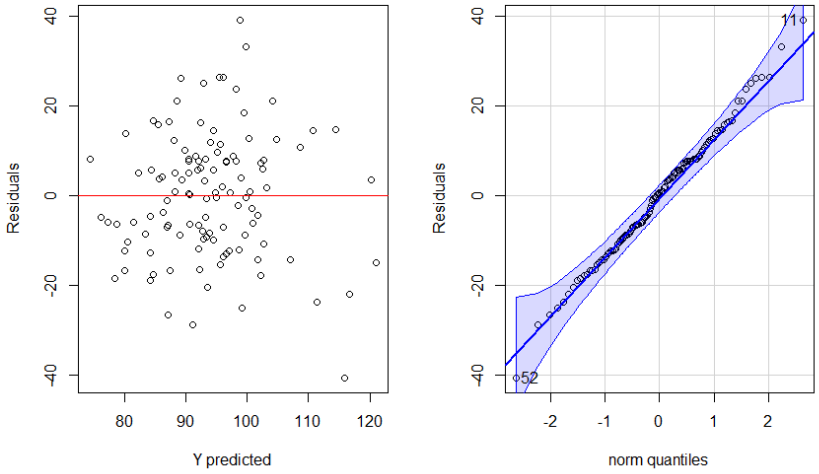

Figure 4: "Residual vs predicted" and "qqPlot of residuals"

From the predicted vs residual plot (Figure 4) we can comment that our residuals are randomly distributed as they do not exhibit any kind of pattern. Also the QQ-plot shows that the errors are approximately normally distributed.

Next we check the spatial dependency of the residuals based on graph.

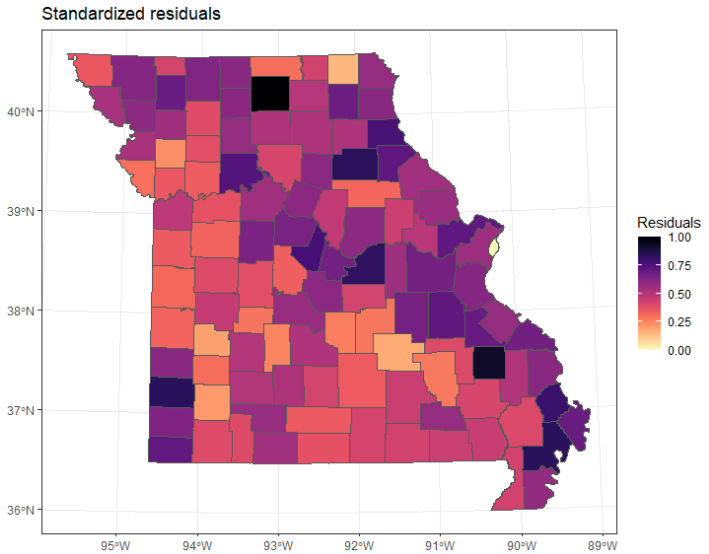

Figure 5: Map of standardized residuals in Missouri

Now from (Figure 5) it is clearly visible to have spatial dependence in the residuals. As the residuals which are close to zero tends to be clustered together, whereas the residuals that are close to 1 are clustered together.

Now wa want to check statistically if the errors are dependent or not or if they have any spatial dependence or not. So we calculate the Moran's I statistic which gave us value less than 0.05. Which means the residuals are spatially dependent.

Now we plot the proximity matrix W (Figure 6)to see if it is sparse or not.

Now after playing with different variables and model, we came up with the Conditional autoregressive model. The estimated coefficients are:

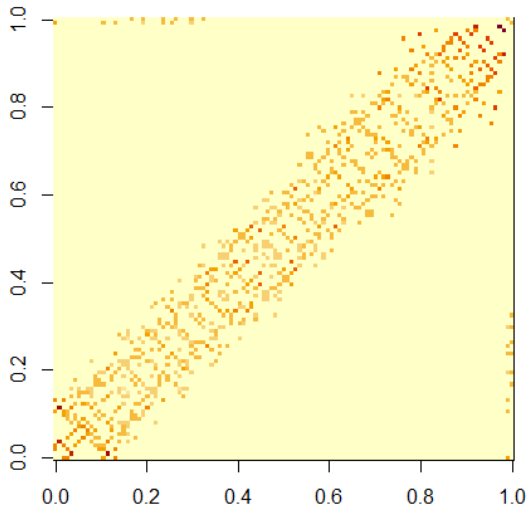

Figure 6: Image of W (non-zero elements are in dark shade)

Table 2: Summary of CAR model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	72.27	12.39	5.83	0.00
hoval	-0.00	0.00	-1.86	0.06
test_rate	0.02	0.00	5.29	0.00
vac_rate	-0.46	0.16	-2.82	0.00
Income	0.00	0.00	2.10	0.04
Popultn	-0.00	0.00	-1.04	0.30
senior	-0.15	0.35	-0.42	0.68

Also the AIC value that we get is **948.38**. Which is not very good. But if we see the R-squared value of the fitted linear model this seems to be reasonable.

Q_3 : Spatial map of the county level estimates

We not find the fitted values and plot (Figure 7) them county wise on the map of Missouri.

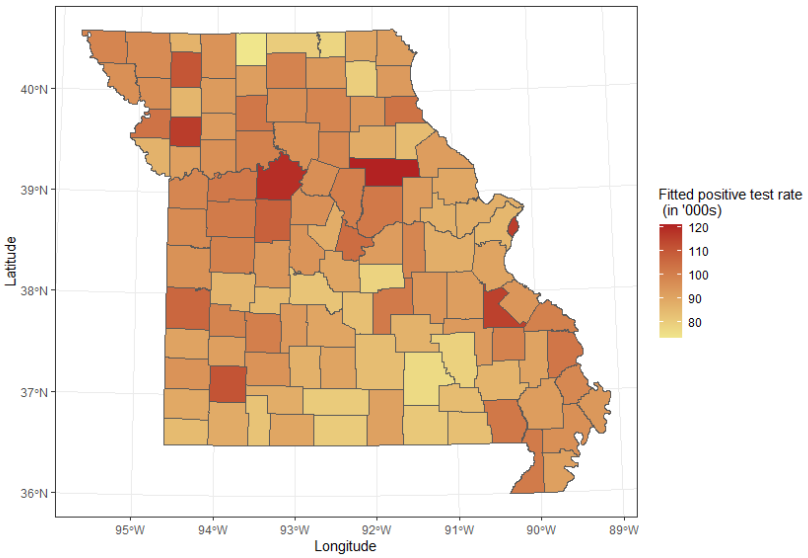

Figure 7: Per capita fitted positive test rate in Missouri

The fitted plot looks very similar to the original plot. In the middle parts of Missouri there are more spatial dependence than the extreme north or extreme south parts. Kansas city, St. Louis, Lawrence, Audrain and Saline were some hotspots. Also the variability is not that much in terms of numbers because it is varying from $80 - 120$ which is quite less. So as a suggestion to the MO dept of health, our suggestion is to focus more on the red coloured counties as they may cause more spread of Covid than the light colour ones.