Shane Clement
03/18/2023

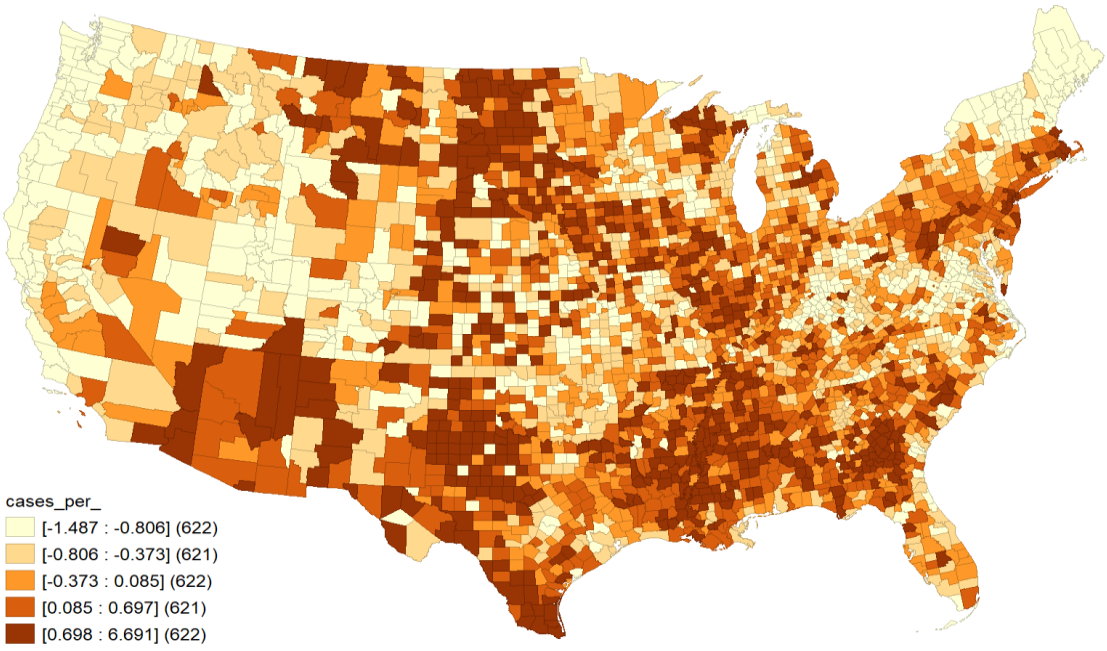
Project 5: Multiple Regression with R and GeoDa

Using the US Covid data at the county level, calculate COVID deaths per 100,000 of the population registered over the first year of the pandemic (roughly March 2020 – February 2021). The data in this file have all been normalized. Here are the meanings of the variables of interest:

population = county population in 2020
cases_per_ = Covid death rate (normalized)
pct_poc = Persons of color as share of total population (normalized)
pct_smoker = Share of population that smoke (normalized)
pct_povert = Share of population in poverty (normalized)
pct_obese = Share of population defined as obese (normalized)
pct_65plus = Share of elderly population (normalized)
per_dem = Share of Democratic registered voters (normalized)

Generate a quantile map of the Covid death rate and reporting the results:

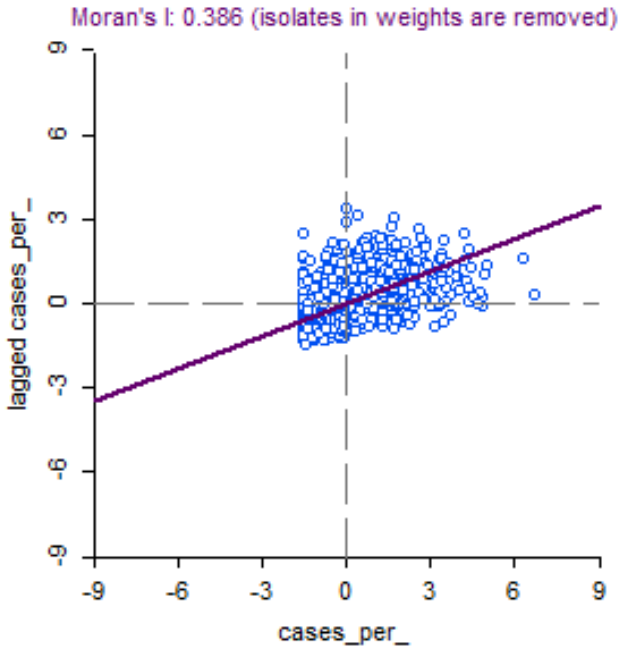
The quantile map shows evidence that Covid death rates have positive spatial autocorrelation and are non-random. There appears to be clustering in parts around the U.S. The darker areas, such as North Dakota, Texas, and Arizona, had higher Covid death rates. Covid death rates were lower in areas such as Northern California, Washington, and Maine.



Determine if there is evidence of spatial autocorrelation in the Covid death rate data using a reasonable spatial weights scheme:

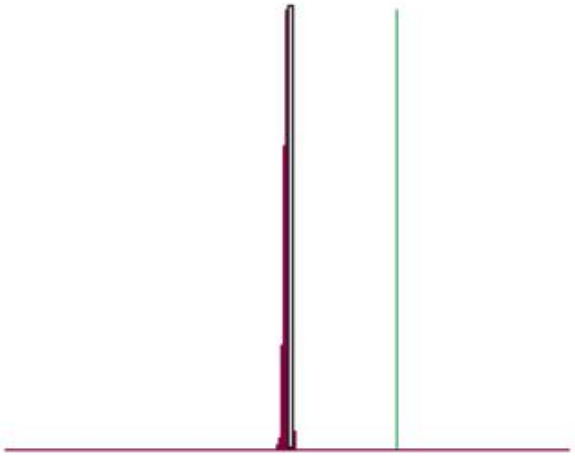
With rook contiguity, the Z-score is 34.76 which is far from 0. The farther from 0, it is more likely to reject the null hypothesis that the spatial distribution of the values of Covid death rate are random. The Moran's I is .386 also shows evidence of positive spatial autocorrelation.

Property	Value
type	rook
symmetry	symmetric
file	uscovid_rook.gal
id variable	POLY_ID
order	1
# observations	3108
min neighbors	0
max neighbors	13
mean neighbors	5.62
median neighbors	6.00
% non-zero	0.18%

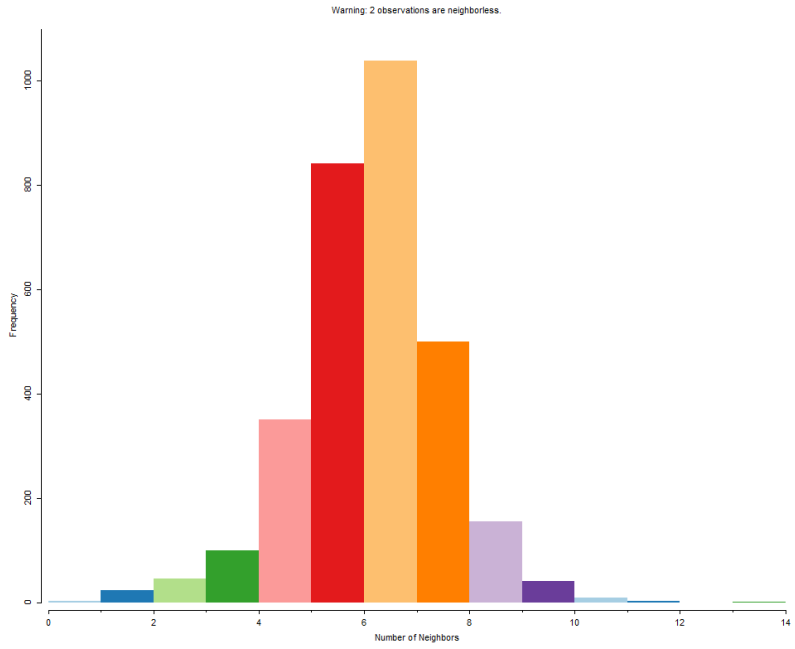


Run

permutations: 999
pseudo p-value: 0.001000

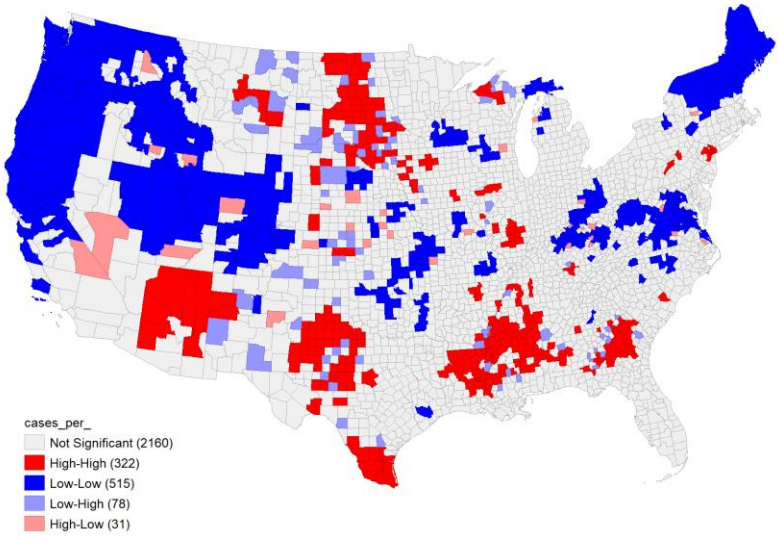


I: 0.3856 E[I]: -0.0003 mean: -0.0004 sd: 0.0111 z-value: 34.7559

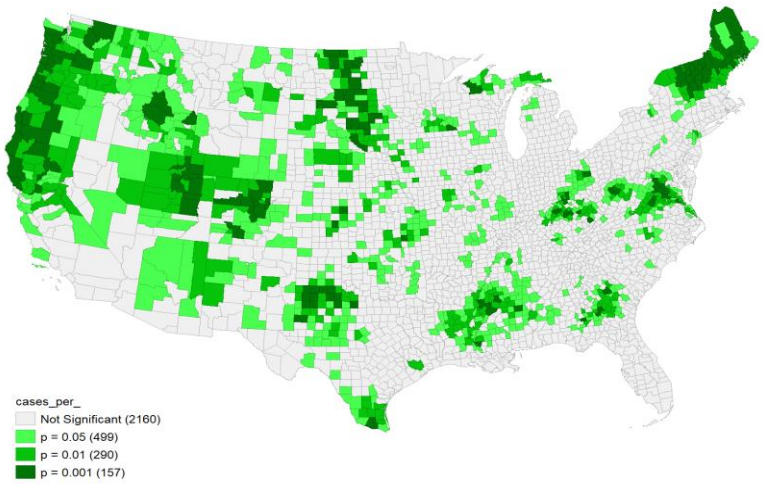


Determine is there is evidence of significant local spatial clusters of Covid death rates:

The LISA cluster map uses rook contiguity and shows evidence of clustering.
There is (dark blue) low-low clustering on the West Coast, Colorado, and the Northeast.
The (bright red) high-high clustering appears areas in such as Arizona, parts of Texas, and Louisiana.



The below map shows the areas with statistical significance. The darker green shaded areas are significant at .001, the moderate green areas are significant at .01, and the lighter green areas are significant at .05. The coast from Northern California to Washington and the Northeast are statistically significant at .001. Overall, some regions have below average cases and some are above average.



Regress the Covid death rate against the six independent variables given above. Show and interpret the regression output (partial regression coefficients, goodness of fit statistics).

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  totals_apr20_feb21
Dependent Variable  :  cases_per_  Number of Observations: 3108
Mean dependent var  :   0.0142213  Number of Variables   :    7
S.D. dependent var  :    0.995707  Degrees of Freedom    : 3101

R-squared           :    0.143199  F-statistic           :     86.3793
Adjusted R-squared  :    0.141541  Prob(F-statistic)     :           0
Sum squared residual:     2640.12  Log likelihood        :    -4156.52
Sigma-square        :    0.851378  Akaike info criterion :     8327.04
S.E. of regression  :    0.922702  Schwarz criterion     :     8369.33
Sigma-square ML     :    0.849461
S.E of regression ML:    0.921662

--------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic    Probability
--------------------------------------------------------------------------------
          CONSTANT     0.0136215      0.0165646       0.822328      0.41094
           pct_poc       0.28321        0.02737        10.3475      0.00000
        pct_smoker    -0.0194318      0.0227318      -0.854828      0.39268
        pct_povert      0.117537      0.0199066        5.90445      0.00000
        pct_obsese      0.148325       0.019553        7.58582      0.00000
        pct_65plus      0.127995      0.0196366        6.51821      0.00000
           per_dem     -0.168049       0.021479       -7.82388      0.00000
--------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   3.169286
TEST ON NORMALITY OF ERRORS
TEST                     DF           VALUE              PROB
Jarque-Bera               2          3484.9192          0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                     DF           VALUE              PROB
Breusch-Pagan test        6           127.0045          0.00000
Koenker-Bassett test      6            39.7589          0.00000
```

The regression equation: Y = .0136 + .28321 pct_poc + (-.0194) pct_smoker + .117 pct_povert + .148 pct_obese + .128 pc_65plus + (-.168) per_dem, states that if the independent variables are 0, then the Covid death rate equals to .0136.

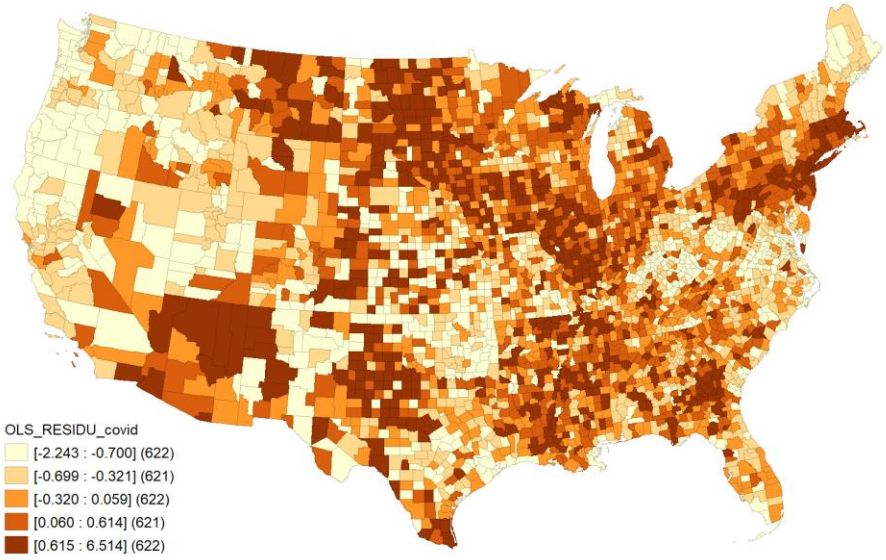
When interpreting the regression model, the partial regression coefficient shows the influence on the dependent variable, holding all other independent variables constant. Therefore, if a partial regression coefficient is positive, it positively influences the dependent variable (and vice versa).

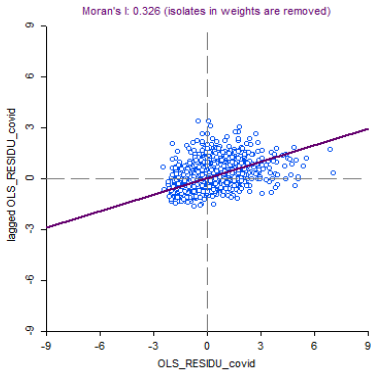
For every unit change in pct_poc, Covid death rates increase by .283. For every unit change in pct_smoker, Covid death rates decrease by .0194. For every unit in pct_poverty, Covid death rates increase by .117. For every unit change in pct_obese, Covid death rate increase by .148. For every unit change in pc_65plus, Covid death rate increases by .128., and for every change in pct_dem, Covid death rate decreases by .168. All of these independent variables are statistically significant with the exception of pct_smoker.

The goodness of fit, the R^2 of .143 reveals that the model explains about 14% variance in Covid death rates.

Capture and plot the residuals from the regression. Determine if there is evidence of spatial autocorrelation in the residuals.

The residual values are the difference between the observed and predicted values for the dependent variable, Covid death rates. The darker areas are higher values of the residuals. This displays evidence of spatial autocorrelation. The Moran's I of .326 shows that there is evidence of spatial autocorrelation where similar values are clustered in different parts of the country.

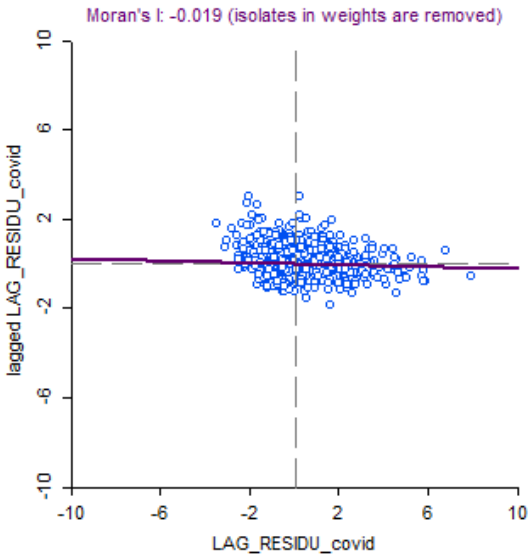




Run spatial lag and spatial error models. Interpret the output from these and show if one of these models fits the data better than the other.

```
>>03/08/23 15:46:21
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : totals_apr20_feb21
Spatial Weight      : uscovid_rook
Dependent Variable  :  cases_per_  Number of Observations: 3108
Mean dependent var  :   0.0142213  Number of Variables   :    8
S.D. dependent var  :    0.995707  Degrees of Freedom    : 3100
Lag coeff.   (Rho)  :    0.517375

R-squared           :    0.331940  Log likelihood        :   -3858.41
Sq. Correlation     : -            Akaike info criterion :    7732.83
Sigma-square        :    0.662337  Schwarz criterion     :    7781.16
S.E of regression   :    0.813841

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
      W_cases_per_      0.517375      0.0203428       25.4329      0.00000
          CONSTANT     0.0066824      0.0146127      0.457302      0.64745
           pct_poc      0.182918      0.0247959       7.37694      0.00000
        pct_smoker     0.0111768      0.0200585      0.557207      0.57739
        pct_povert      0.088136      0.0176873       4.98302      0.00000
        pct_obsese     0.0750323      0.0174003       4.31212      0.00002
        pct_65plus     0.0925496      0.0174238       5.31169      0.00000
           per_dem    -0.0912074      0.0192919      -4.72775      0.00000
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      6       137.8184     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : uscovid_rook
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1       596.2119     0.00000
```

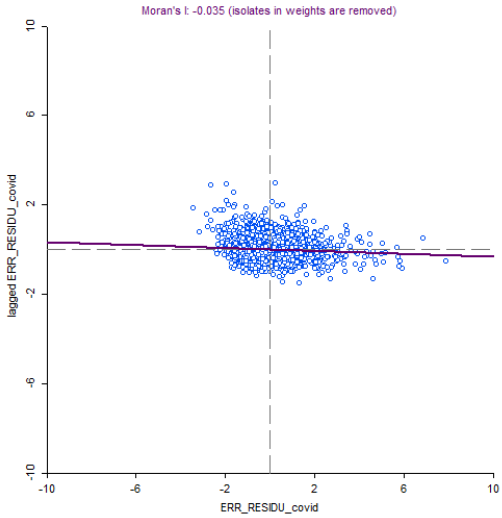
In the United States, Covid death rates are spatially related to neighboring areas. An independent variable is added to the standard model to represent a spatial lag of the dependent variable. The spatial lag of Covid death rates is .517, which is significantly different than 0 at the .01 level. The spatial lag exerts influence over the dependent variable, Covid death rates. Additionally, the Likelihood Ratio Test 596.22 is the value for the t-statistic that states how much the spatial lag is capturing spatial dependence in the data. The Moran's I for the spatial lag is -.019.

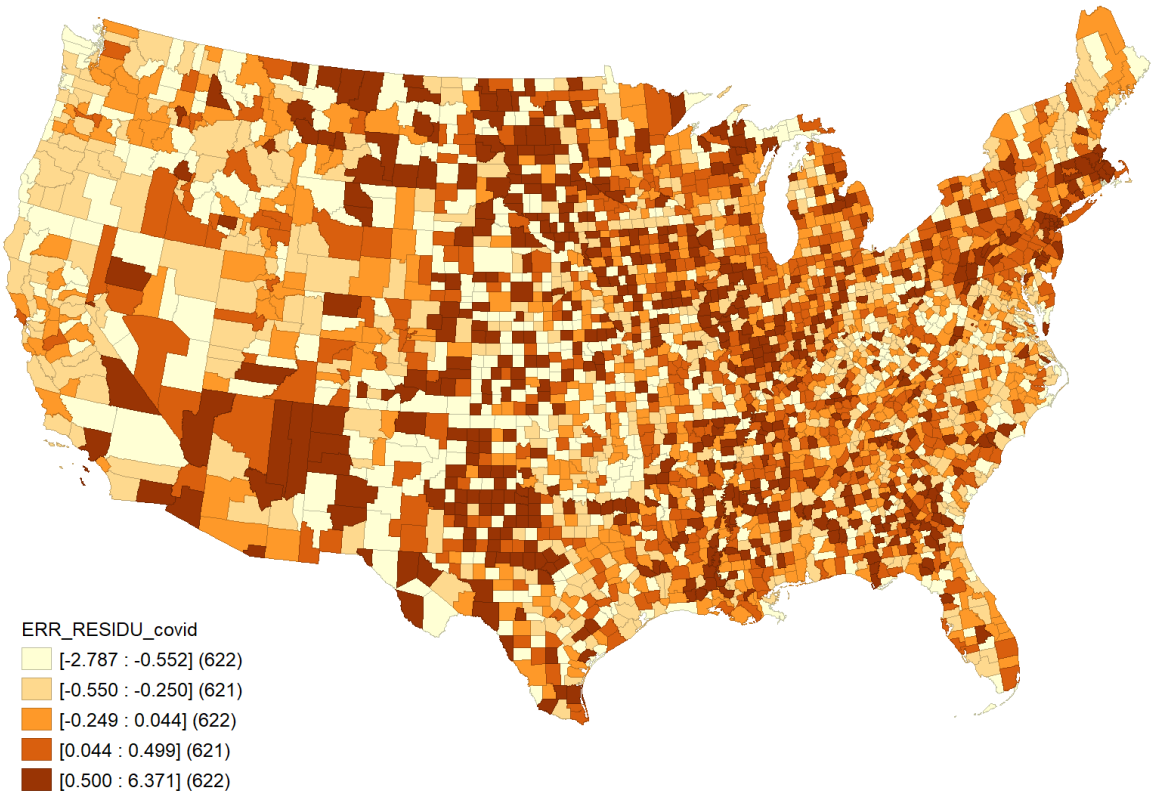


The following are the results from running the spatial error model. Spatial errors result from measurement issues or from the influence of spatially autocorrelated variables that are absent from the model but have influence on the other variables included in the model.

LAMBDA is the spatial error term and is .538 in this model. This value is significant and a higher value that the spatial lag error of .517. The Maximum Likelihood Ratio Test is 619. Therefore, the spatial error should be used instead of the spatial lag. The Moran's I on the spatial error model is -.035 which is not statistically significant. Meaning that the spatial dependence has been captured by the spatial error test. Lastly, the map of the residuals of the spatial errors shows that spatial autocorrelation has been removed and can be applied to the model.

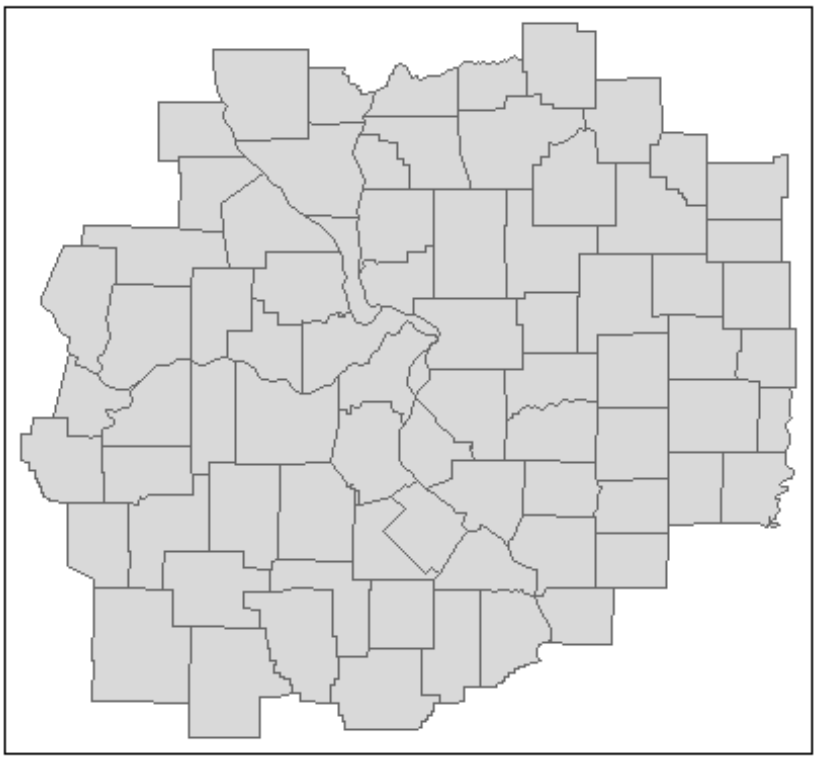
```
>>03/08/23 16:08:04
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : totals_apr20_feb21
Spatial Weight      : uscovid_rook
Dependent Variable  :  cases_per_  Number of Observations: 3108
Mean dependent var  :    0.014221  Number of Variables   :    7
S.D. dependent var  :    0.995707  Degrees of Freedom    : 3101
Lag coeff. (Lambda) :    0.538533

R-squared           :    0.340613  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-3846.565610
Sigma-square        :    0.653737  Akaike info criterion :     7707.13
S.E of regression   :     0.80854  Schwarz criterion     :     7749.42

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
          CONSTANT     0.0121682      0.0314134       0.387356     0.69849
           pct_poc      0.313535      0.0367834        8.52383     0.00000
        pct_smoker  -0.000836979      0.0286871     -0.0291762     0.97672
        pct_povert      0.131971      0.0218024        6.05307     0.00000
        pct_obsese     0.0917403      0.0274862        3.33769     0.00084
        pct_65plus      0.129458      0.0212653        6.08778     0.00000
           per_dem     -0.176706      0.0277512       -6.36751     0.00000
            LAMBDA      0.538533       0.020654         26.074     0.00000
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      6       149.5453     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : uscovid_rook
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1       619.9093     0.00000
```


Moran's I: -0.035 (isolates in weights are removed)
lagged ERR_RESIDU_covid
ERR_RESIDU_covid
10
6
2
-2
-6
-10
-10
-6
-2
2
6
10


ERR_RESIDU_covid
[-2.787 : -0.552] (622)
[-0.550 : -0.250] (621)
[-0.249 : 0.044] (622)
[0.044 : 0.499] (621)
[0.500 : 6.371] (622)

Using a stlouis.shp file and the packages spgwr, rgdal, and tmap for spatial analysis and mapping.

```
library(sf)
library(spgwr)
library(rgdal)
library(spdep)
library(tmap)
setwd("C://Unit 5/data/stlouis")
stlouis2 <- readOGR("C://Unit 5/data/stlouis/stlouis.shp")
#summary(stlouis2)
qtm(stlouis2)
```



Run a non-spatial regression of the homicide rate (HR8893) on police expenditure (PE87) and a local deprivation index (RDAC90) and interpret the output.

The regression model shows if the coefficients on police expenditure and the deprivation index were 0, the homicide rate would be .0375, but it's more important to focus on the partial regression coefficients. The model shows that for every unit change in police expenditure, homicide rate increases on average by 1.567, holding the deprivation index constant. For every unit change on the deprivation index, the homicide rate increases on average by 5.29, holding police expenditure constant. Both independent variables are statistically significant. With a p-value of 0.000000003194, we reject the null hypothesis that independent variables do not have a significant impact on the dependent variable. The R^2 shows that 46% of variance in model is explained by the dependent variable, homicide rate. Also, the residual standard error is 4.692. The smaller the value, the better the fit.

```
stl2_model <-lm(HR8893 ~ PE87 + RDAC90,data=stlouis2)
summary(stl2_model)

##
## Call:
## lm(formula = HR8893 ~ PE87 + RDAC90, data = stlouis2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8749 -2.7648 -0.6719  2.1715 20.2329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03748    1.54263   0.024    0.981
## PE87         1.56705    0.37128   4.221 6.75e-05 ***
## RDAC90       5.29091    0.82492   6.414 1.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.692 on 75 degrees of freedom
## Multiple R-squared:  0.4065, Adjusted R-squared:  0.3906
## F-statistic: 25.68 on 2 and 75 DF,  p-value: 3.194e-09
```

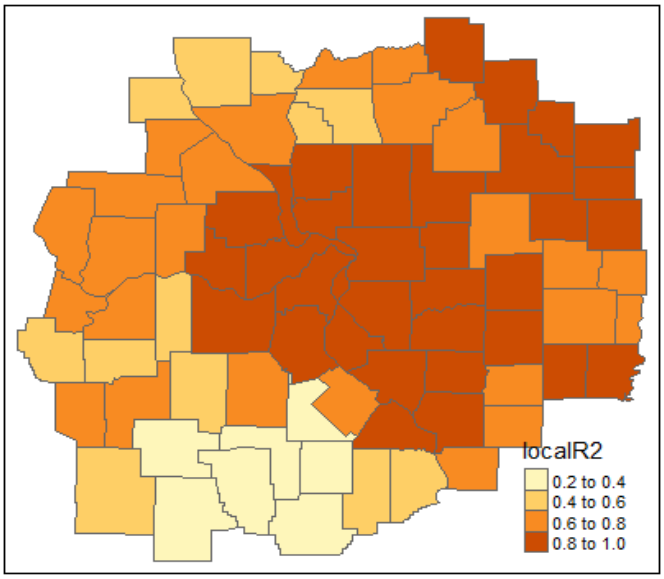
Using the R package spgwr, show how the R2 and the partial regression coefficients in the model vary across the study region. Report and interpret the output.

```
GWRbandwidth <-gwr.sel(HR8893 ~ PE87 + RDAC90,data=stlouis2, adapt=T)

## Adaptive q: 0.381966 CV score: 2462.794
## Adaptive q: 0.618034 CV score: 2493.324
## Adaptive q: 0.236068 CV score: 2309.391
## Adaptive q: 0.145898 CV score: 2094.679
## Adaptive q: 0.09016994 CV score: 1394.082
## Adaptive q: 0.05572809 CV score: 1032.717
## Adaptive q: 0.03444185 CV score: 903.0778
## Adaptive q: 0.02128624 CV score: 840.6013
## Adaptive q: 0.01315562 CV score: 843.9937
## Adaptive q: 0.01808047 CV score: 837.933
## Adaptive q: 0.01803978 CV score: 837.9085
## Adaptive q: 0.0161742 CV score: 837.8115
## Adaptive q: 0.01701687 CV score: 837.5713
## Adaptive q: 0.01705756 CV score: 837.5728
## Adaptive q: 0.01697618 CV score: 837.5709
## Adaptive q: 0.01666985 CV score: 837.6052
## Adaptive q: 0.01693549 CV score: 837.5716
## Adaptive q: 0.01697618 CV score: 837.5709

gwr.model <-gwr(HR8893 ~ PE87 + RDAC90,data=stlouis2, adapt=GWRbandwidth, hat
matrix = TRUE,se.fit=TRUE)

gwr.model

## Call:
## gwr(formula = HR8893 ~ PE87 + RDAC90, data = stlouis2, adapt = GWRbandwidt
h,
##     hatmatrix = TRUE, se.fit = TRUE)
## Kernel function: gwr.Gauss
## Adaptive quantile: 0.01697618 (about 1 of 78 data points)
## Summary of GWR coefficient estimates at data points:
##                   Min.  1st Qu.   Median  3rd Qu.     Max. Global
## X.Intercept. -1.86007  1.34844  3.38473  5.33793 18.08941 0.0375
## PE87         -3.31961 -0.18401  0.44314  1.00753  2.69020 1.5670
## RDAC90       -9.17972  0.57759  1.87845  6.50702 11.94509 5.2909
## Number of data points: 78
## Effective number of parameters (residual: 2traceS - traceS'S): 47.66878
## Effective degrees of freedom (residual: 2traceS - traceS'S): 30.33122
## Sigma (residual: 2traceS - traceS'S): 2.672647
## Effective number of parameters (model: traceS): 36.96118
## Effective degrees of freedom (model: traceS): 41.03882
## Sigma (model: traceS): 2.297678
```

```
## Sigma (ML): 1.66663
## AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 452.7336
## AIC (GWR p. 96, eq. 4.22): 338.001
## Residual sum of squares: 216.6572
## Quasi-global R2: 0.9221223

results <-as.data.frame(gwr.model$SDF)
names(results)

##  [1] "sum.w"               "X.Intercept."        "PE87"
##  [4] "RDAC90"              "X.Intercept._se"     "PE87_se"
##  [7] "RDAC90_se"           "gwr.e"               "pred"
## [10] "pred.se"             "localR2"             "X.Intercept._se_EDF"
## [13] "PE87_se_EDF"         "RDAC90_se_EDF"       "pred.se.1"

gwr.map <-cbind(stlouis2, as.matrix(results))
```

The GWR model uses a method like Kernal Density Estimation to examine areas across the region. In the Local R^2 map, the darker areas represent where variance of the homicide rate is better explained in the model than the lighter areas. For example, the central part of St Louis has an R^2 ranging from .8 to 1, meaning that these areas are explaining from 80% to 100% of the model. The lighter shaded areas have an R^2 ranging from .20 to .40 and have less significance in explaining the homicide rate in the model.

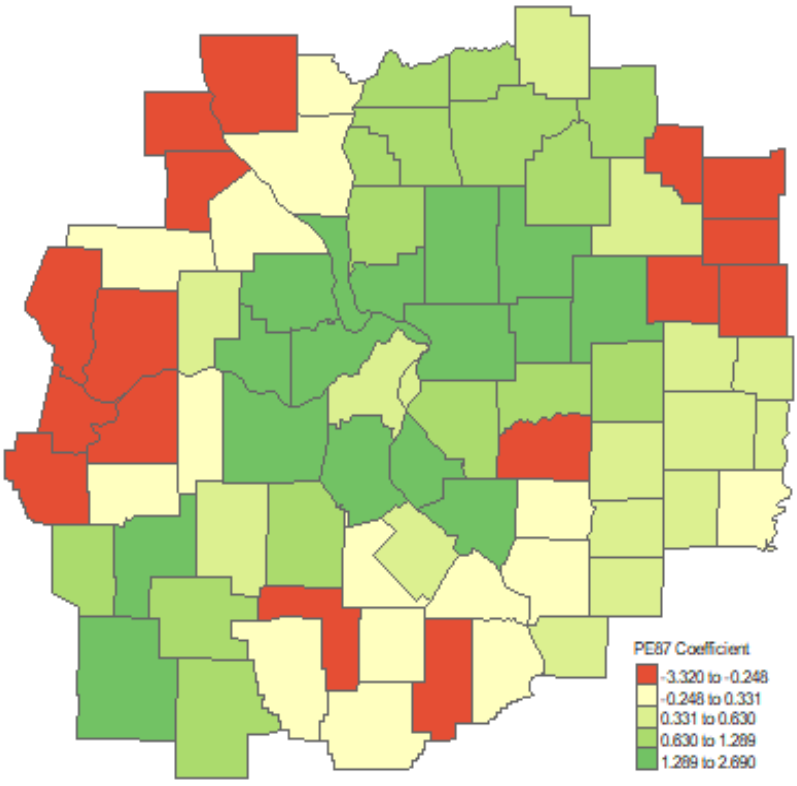
```
qtm(gwr.map, fill="localR2")
```



The below maps of police expenditure and a deprivation index show the impacts for each of the independent variables on the dependent variable. The areas with the higher and positive coefficients have more of an impact on the homicide rate. For example, in the first map, the central area that is shaded dark green have higher, positive coefficients, meaning that the

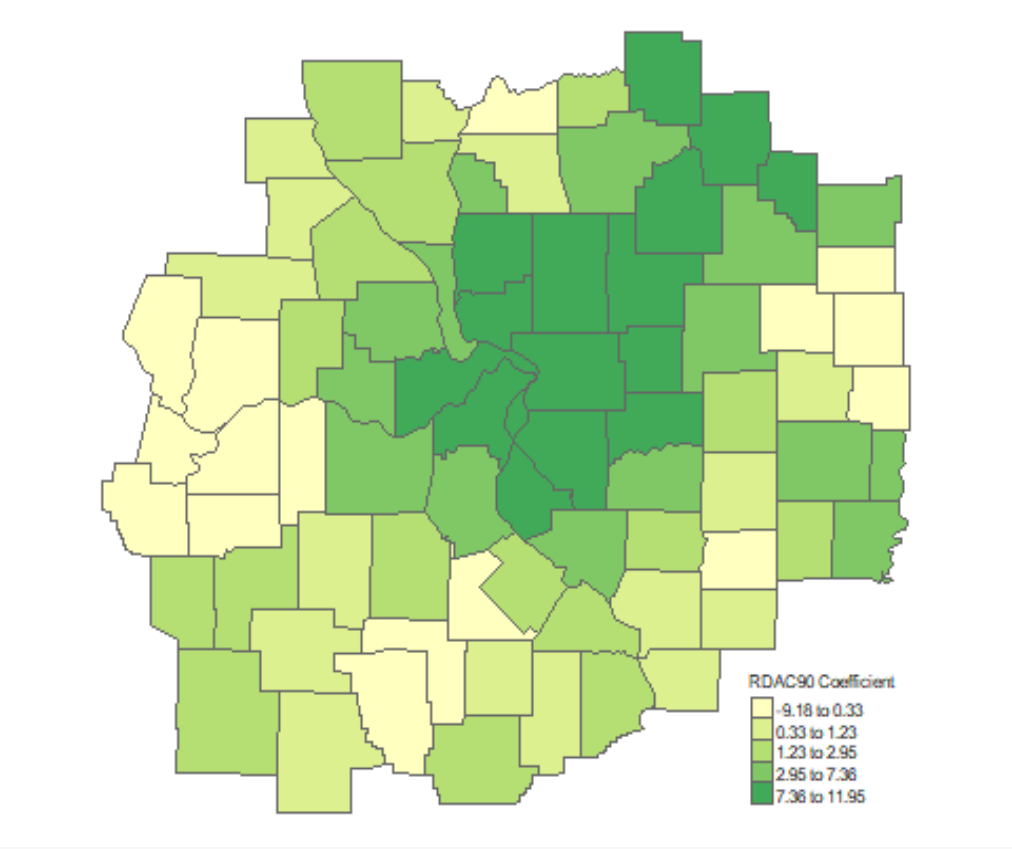
homicide rate is more impacted by police expenditure in these areas. The areas that are shaded red have coefficients that are negative, and therefore, the homicide rate is negatively impacted by police expenditure. For the deprivation index map, the central area, also shaded dark green, are positively and significantly impacted by the deprivation index.

```
stl_map2 <- tm_shape(gwr.map) +tm_fill("PE87.1", n=5,style='quantile',title="
PE87 Coefficient")+tm_borders()+tm_layout(frame=FALSE,legend.text.size=.5,leg
end.title.size =.6)
stl_map2

## Variable(s) "PE87.1" contains positive and negative values, so midpoint is
set to 0. Set midpoint = NA to show the full spectrum of the color palette.
```



```
stl_map3 <- tm_shape(gwr.map) +tm_fill("RDAC90.1", n=5,style='quantile',title
="RDAC90 Coefficient")+tm_borders() +tm_layout(frame=FALSE,legend.text.size=.
5,legend.title.size =.6)
stl_map3
```

```
 ## Variable(s) "RDAC90.1" contains positive and negative values, so midpoint
is set to 0. Set midpoint = NA to show the full spectrum of the color palette
```



.