Spatial Autocorrelation in Areal Data

Handed out: Thursday, April 23, 2020

Return date: Sunday, May 3, 2020 (no extensions can be granted)

Grading: This lab counts 12 % towards your final grade

Other Dates:

- The lab will be returned on Tuesday May 5, 2020
- An online Q&A session will be held on Tuesday May5, 2020, from 4:00 pm to 5:00 pm
- The final exam will be written online on Thursday, May 7, 2020, from 4:00 to 7:00 pm.

Objectives: The first part introduces you to different model structures and the concept of spatial spillovers. The second part explores properties of regression based global spatial statistics, models and their identification, while the last part explores local spatial statistics and their impact on global statistics.

Format of answer: Your answers (statistical figures and verbal description) should be submitted as ***hardcopy***. Add a running title with the following information: Lab05, your name and page numbers. You may use this document as template. Copy the requested statistical figures into your document. No trial and error answers are permitted. Label each answer properly with the bold task headings. You are expected to hand in professionally formatted answers: use a fixed pitch font, like **`Courier New`**, for any R code the use mathematical type-setting when equations are required. Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

Part I: Spatial spillovers (4 points)

Task 1 (4 points): Study the paper by Golpher , A. B., and P. R. Voss. How to Interpret the Coefficients of Spatial Models: Spillovers, Direct and Indirect Effects. *Spatial Demography* (2016) 4:175-205.

Note: In this lab the terminology of the Golpher's paper is used.

Assuming that we just deal with a simple bivariate base *SLM* model with $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$ with n spatial observations and $\varepsilon \sim N(0, \sigma^2 \cdot \mathrm{I})$.

[a] How to interpret the entries in the derivatives matrix below with regards to changes in the independent variable x_i at location i? (1 point)

$$\frac{\partial \mathrm{y}}{\partial \mathrm{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix}$$

Since there are n locations, let's denote observation for each location by i where i varies from $i = 1,2,3, \dots \dots n$.

All observations for dependent variable y_i can be represented as $y_1, y_2, y_3, \dots . y_n$ corresponding observations for independent variable x can be represented as $x_1, x_2, x_3, \dots . x_n$. Let's represent change in y_i due to changes in x_i by Δy_i and Δx_i be change in x_i. Thus $\Delta y_1, \Delta y_2, \dots . \Delta y_n$ represent change in observations $y_1, y_2, \dots \dots . . y_n$ respectively 1 due to changes in x_i. Let's say $\Delta x_1, \Delta x_2, \dots . \Delta x_n$ represent change in x_i at locations $1,2, \dots \dots n$ respectively. Thus, total change in observation y_1 can be represented as

$$\Delta y_1 = \frac{\partial y_1}{\partial x_1} \times \Delta x_1 + \frac{\partial y_1}{\partial x_2} \times \Delta x_2 + \cdots + \frac{\partial y_1}{\partial x_n} \times \Delta x_n$$

And Δy_2 represents change in observation y_2 at location 2 due to changes in x_i. Thus,

$$\Delta y_2 = \frac{\partial y_2}{\partial x_1} \times \Delta x_1 + \frac{\partial y_2}{\partial x_2} \times \Delta x_2 + \cdots + \frac{\partial y_2}{\partial x_n} \times \Delta x_n$$

Where $\frac{\partial y_1}{\partial x_1}$ represent change in y_1 for 1 unit change in x_1, $\frac{\partial y_1}{\partial x_2}$ represents change in y_1 for 1 unit change in x_1 and so on. Similarly $\frac{\partial y_2}{\partial x_1}$ represent change y_2 for 1 unit change in x_1 and $\frac{\partial y_2}{\partial x_2}$ represents change in y_2 for 1 unit change in x_2 and so on. But for simple linear regression, for a matrix represented by $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$, off-diagonal elements are zero as all observations are independent. Therefore

$\Delta y_1 = \frac{\partial y_1}{\partial x_1} \times \Delta x_1$ and $\Delta y_2 = \frac{\partial y_2}{\partial x_2} \times \Delta x_2$. But since $x_1, x_2, \dots x_n$ are observations for the same random variable x_i , $\frac{\partial y_1}{\partial x_1} = \frac{\partial y_2}{\partial x_2} = \cdots = \frac{\partial y_n}{\partial x_n}$. Therefore, all diagonal elements in the matrix above represent the value of a regression coefficient β_1. In other words, if there are changes in variable x_i at location i, it will cause change of β_1 in dependent variable y_i at location i for each 1 unit change in x_i at location i.

[b] Explain, based on the derivative matrix $\frac{\partial \mathrm{y}}{\partial \mathrm{x}}$, what the *direct*, *indirect* and *total* effects are as well as how to interpret these effects? (1 point)

Glogher and Voss (2016) define direct effect as change in a observations of dependent variable in a given region for a 1-unit change in a particular independent variable of that region.

From that definition, direct effect of variable x_i on dependent variable y_i at location 1 i.e. y_1 can be expressed as $\frac{\partial y_1}{\partial x_1}$ whereas average direct effect of x_i across all locations can be expressed as

$$Direct\ Effect = (\frac{1}{n}) \times (\frac{\partial y_1}{\partial x_1} + \frac{\partial y_2}{\partial x_2} + \cdots + \frac{\partial y_n}{\partial x_n})$$

From the above equation, diagonal elements of matrix given in task 1(a) represent direct effect and average of these diagonal elements gives us mean direct effect.

On the other hand, indirect effect is change in a dependent variable of a region for a 1-unit change in independent variable in other regions. For dependent variable y_i at location 1 i.e. y_1 , indirect effect due to variable x_i can be expressed as $\frac{\partial y_1}{\partial x_2} \times \Delta x_2 + \frac{\partial y_1}{\partial x_3} \times \Delta x_3 + \cdots + \frac{\partial y_1}{\partial x_n} \times \Delta x_n$. Indirect effect at all other locations can be expressed likewise. Hence all non-diagonal elements of matrix in task 1(a) represent indirect effect. Matrix in task 1 (a) has total n^2 elements and out of those n elements are diagonal. Thus, total non-diagonal elements are $n^2 - n$. Thus, average indirect effect of x_i on y_i across all locations can be expressed as

$$Indirect\ Effect = \left(\frac{1}{n^2 - n}\right) \times \left(\left(\frac{\partial y_1}{\partial x_2} + \frac{\partial y_1}{\partial x_3} + \cdots + \frac{\partial y_1}{\partial x_n}\right) + \left(\frac{\partial y_2}{\partial x_1} + \frac{\partial y_2}{\partial x_3} + \cdots + \frac{\partial y_2}{\partial x_n}\right) + \cdots + \left(\frac{\partial y_n}{\partial x_1} + \frac{\partial y_n}{\partial x_2} + \cdots + \frac{\partial y_n}{\partial x_{n-1}}\right)\right)$$

Total effect is nothing but sum of the direct and indirect effect. Therefore, total effect for dependent variable y_i at location 1 i.e. y_1 can be expressed as:

$$\left(\frac{1}{n}\right) \times \frac{\partial y_1}{\partial x_1} + \frac{\partial y_1}{\partial x_2} + \frac{\partial y_1}{\partial x_3} + \cdots + \frac{\partial y_1}{\partial x_n}$$

Whereas average total effect of x_i on y_i can be expressed as:

$$\left(\frac{1}{n^2}\right) \times \left(\left(\frac{\partial y_1}{\partial x_1} + \frac{\partial y_1}{\partial x_2} + \frac{\partial y_1}{\partial x_3} + \cdots + \frac{\partial y_1}{\partial x_n}\right) + \left(\frac{\partial y_2}{\partial x_1} + \frac{\partial y_2}{\partial x_2} + \frac{\partial y_2}{\partial x_3} + \cdots + \frac{\partial y_2}{\partial x_n}\right) + \cdots + \left(\frac{\partial y_n}{\partial x_1} + \frac{\partial y_n}{\partial x_2} + \cdots + \frac{\partial y_n}{\partial x_n}\right)\right)$$

In summary, direct effect represent change in a dependent variable of region due to change in exogeneous variable in the same region whereas indirect effect represents change in dependent variable of region due to changes in exogeneous variable of other regions. Thus, direct effect is unaffected by changes in exogeneous variable in other regions whereas indirect effects are affected by changes in exogeneous variables in connected regions. Total effect represents the change in dependent variable in a region due to change in exogeneous variable of that region as well that of connected regions.

[c] Why does the spatial error model *SEM* have only direct effect predictions $E[\mathbf{y}|\mathbf{x}] = \beta_0 + \beta_1 \cdot \mathbf{x}$? (0.5 points)

In SAR or SAC models, dependent variable y is dependent upon the value of exogeneous variable in the same region as well as the value of dependent variable in connected regions and hence as a result whereas in SDEM and SLX models, the value of dependent variable in a region is dependent on value of exogeneous variable in the same as well as the value of exogeneous variable in connected regions. As a result, in these models, dependent variable have indirect effects either from endogenous or

exogeneous variables. Contrary, in a spatial error model, we assume that only residuals are spatially autocorrelated. However, since with the equation, $E[\mathrm{y}|\mathrm{x}] = \beta_0 + \beta_1 \cdot \mathrm{x}$, we just estimate the value of y and not the errors associated with it, spatial error model has only direct effects.

[d] The spatial error model *SEM* can be rewritten as (see page 179)

$$(\mathbf{I} - \lambda \cdot \mathbf{W}) \cdot \mathbf{y} = \beta_0 \cdot (\mathbf{I} - \lambda \cdot \mathbf{W}) \cdot \mathbf{1} + \beta_1 \cdot (\mathbf{I} - \lambda \cdot \mathbf{W}) \cdot \mathbf{x} + \boldsymbol{\varepsilon} \text{ with } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \cdot \mathbf{I}).$$

Argue why the estimates b_0 and b_1 of ordinary least squares and this general least square model are the same. Thus, the *OLS* estimates are unbiased, i.e., $E(b_0) = \beta_0$ and $E(b_1) = \beta_1$. (0.5 points)

For SLM, the expected value of y is given as:

$$E[\mathrm{y}|\mathrm{x}] = b_0 + b_1 \cdot \mathrm{x}$$

In the given equation for SEM, if we divide both sides by $(\mathrm{I} - \lambda \cdot \mathrm{W})$, then the Equation of SEM reduces to-

$$\mathrm{y} = \beta_0 + \beta_1 . \mathrm{x} + (\mathrm{I} - \lambda \cdot \mathrm{W})^{-1} . \varepsilon$$

Thus, the expected value of dependent variable SEM can be expressed as:

$$\mathrm{E}[\mathrm{y}|\mathrm{x}] = \beta_0 + \beta_1 . \mathrm{x}$$

Since the equation for expected value of dependent variable in both SEM and SLM have same form,

$$E(b_0) = \beta_0 \text{ and } E(b_1) = \beta_1$$

[e] Discuss for Table 2 and the models SAR & SAC with $\rho = 0.2$ and at $\rho = 0.5$ how and why the spillover effects distribute around the central region 7 and the peripheral region 3. (1 point)

You may also use for your argument the property $(\mathbf{I} - \rho \cdot \mathbf{W})^{-1} = \mathbf{I} + \rho \cdot \mathbf{W} + \rho^2 \cdot \mathbf{W}^2 + \rho^3 \cdot \mathbf{W}^3 + \cdots$ (see page 180). Note that $\mathbf{W}^2$ consists of 2-steps links and $\mathbf{W}^3$ predominately of 3-steps links.

In a SAR and SAC model, value of the dependent variable is the results of direct effect of value of exogeneous variable in the region under consideration as well as the values of exogeneous variable in the connected regions. The coefficient associated with exogeneous variable in both these models is expressed as: $\beta . (\mathbf{I} - \rho \cdot \mathbf{W})^{-1}$. However, $(\mathbf{I} - \rho \cdot \mathbf{W})^{-1} = \mathbf{I} + \rho \cdot \mathbf{W} + \rho^2 \cdot \mathbf{W}^2 + \rho^3 \cdot \mathbf{W}^3 + \cdots$

Since $\mathbf{W}^2$ consists of 2-steps links and $\mathbf{W}^3$ predominately of 3-steps links and so on, this power series leads to effect of change in endogenous variable in one region to change in endogenous variable in all regions resulting in a global spillover. But since the values of ρ are less than 1, with each 1 unit increasing power, the value of coefficient associated with link matrix decreases exponentially. Hence as the power of the link matrix increases, the associated effect on the coefficient also decreases.

Golgher and Voss (2016) represent example of 7 regions on page 190. Table 2 provides the spillover in other regions due to 10 units of increase in human capital in region 7 and region 3. In both the examples, sum of spillover in all regions remains same.

In the given example region 7 is at the periphery and is surrounded by total 2 regions, region 6 and 8. Therefore the spillover effect in regions 6 and 8 is far larger as compared to all other regions. Values of the spillover effect in farther region are much less than the values of spillover effects in region 6 and 8.

For rho =0.2, among region 6 and 8, region 8 has slightly greater spillover than region 8 because region 8 has a greater number of immediate neighbors than region 6 and hence it draws the spillover from several regions at higher orders of rho. But since the value of rho is small, at higher orders rho, this spillover effect diminishes rapidly, and hence only small increment is seen in region 8 than region 6 at rho =0.2. Likewise, smaller spillover effect has been observed in farther regions. After regions 6 and 8, the value of spillover effect is greater in region 5 which is closest non-immediate neighbor is. Region 3 and 4 are farther from the region7 than the region 5 and hence smaller spillover is observed in these regions compared to region 5. But again region 3 has more immediate neighbors and hence greater spillover in this region than region 4. Region 1 and 2 are farthest from region 7 and have same spillover values. Because of the smaller value of rho and their larger distance from region 7, any connectivity differences in these two regions do not make any difference in the spillover values in these regions. For rho=0.5, the order of spillover effect in regions remains almost same, with higher values for region 6 and 8 followed by region 5,3 and 4. The only difference is for rho =0.2, region 1 and 2 received same spillover effect but at rho =0.5, region 2 has greater spillover effect. This is because in this case value of rho is large enough to not dampen the effect of larger connectivity region 2 than region 1.

Central region 3 is surrounded by total 5 regions, regions 1,2,4,5 and 8. Therefore spillover is stronger in these immediately connected region. But in region 7, the spillover values were much larger in neighboring regions. This is because contrary to region 7 example above, in this case the spillover effect gets distributed in larger number of neighbors. Among immediate neighbors, for rho=0.2, region 2 has greatest spillover effect followed by region 8, region 4, region 1, region 5. Although region 8 has more neighbors, at rho=0.2, it does not increase spillover effect in this region. But at rho=0.5, greater connectivity of region 8 makes difference to receive larger spillover value. Similar effect can be observed for region 2 and 4. For rho =0.2, despite larger connectivity, region 4 has less spillover effect compared to region 4 but at rho=0.5, it receives larger spillover value. Region 6 and 7 are farther from the region 3 but among them region 7 is farthest and hence region 7 has least spillover effect than region 6 in both the cases: rho =0.2 and rho=0.5.

Part II: Global autocorrelation (4 points)

Throughout Parts II and III you will be working with the **`tractShp`** dataset in the package **`TexMix`**. Make sure to ***exclude*** the Love Field and the DFW airport tracts

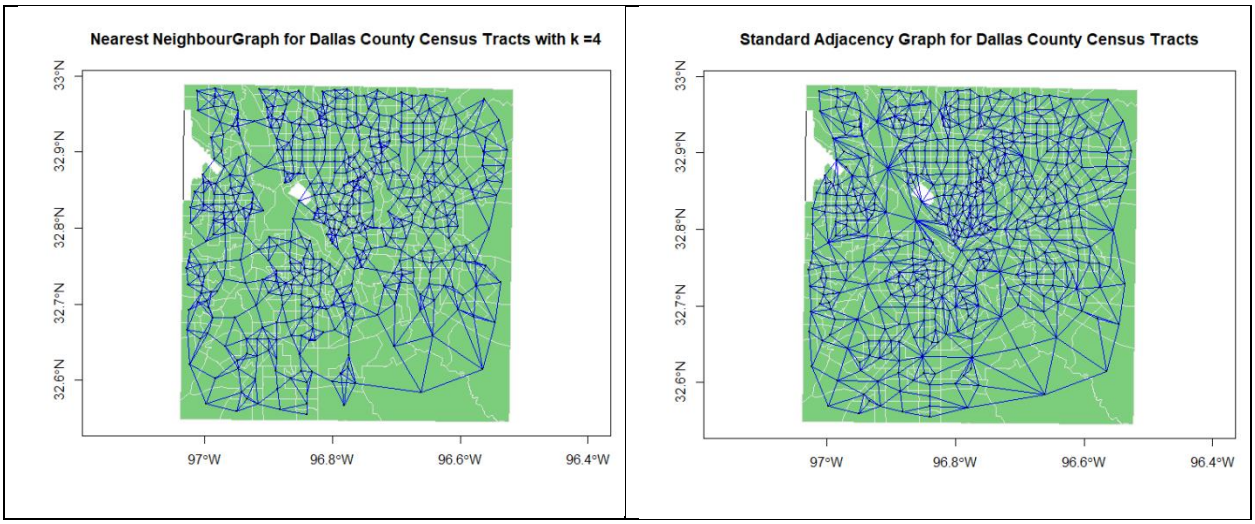
Task 2 (1 points): Calculate the $k = 4$ nearest neighbor graphs for the census tracts and the standard adjacency graph. (see Bivand *et al.* page 269)

Map and interpret both graphs.

Why are binary nearest neighbor graphs not necessarily symmetric?

```
library(spatstat)
library(maptools)
library(colorspace)
library(sp)
library(spdep)
library(TexMix)
library(car)
library(CancerSEA)

setwd("E:\\Courses\\Spring2020\\TexMix_0.5.1\\TexMix\\data")
load("E:\\Courses\\Spring2020\\TexMix_0.5.1\\TexMix\\data\\tractShp.RData")
plot(tractShp, border="black", lwd=1,  axes=T)
mod_tractShp <- tractShp[!row.names(tractShp) %in% c("153", "247"),]
coords <- coordinates(mod_tractShp)
IDs <- row.names(mod_tractShp)
# k =4 nearest neighbhours
Dallas4_nb <- knn2nb(knearneigh(coords, k=4), row.names = IDs)
plot(mod_tractShp, col="palegreen3" ,border=grey(0.9), axes=T, add=T)
plot(Dallas4_nb, coords, pch=19, cex=0.1,col="blue", add=T)
title("Nearest NeighbourGraph for Dallas County Census Tracts with k =4")
box()
# standard adjacency graph
tracts.link <- poly2nb(mod_tractShp, queen=F)
```

```
plot(mod_tractShp,col="palegreen3" ,border=grey(0.9), axes=T, add=T)
plot(tracts.link,coords=coords, pch=19, cex=0.1,col="blue", add=T)
title("Standard Adjacency Graph for Dallas County Census Tracts")
box()
```



In the nearest neighbor graph, each polygon, including polygons at the edges have at least 4 neighbors. The knearneigh function makes sure that each polygon has at least 4 neighbors. Because of this constraint some polygons seem to have more than 4 connections, but that is because one polygon might have 4 neighbors, but additional connection is required to satisfy the condition of 4 neighbors for the adjacent polygon. Therefore, although visually more than 4 connections are observed for some polygons, the function knearneigh returns indices of only nearest 4 neighbors.

The standard adjacency graph displayed above follows the rook's contiguity. That means as long as the two polygons share edge, they are neighbors. In this graph, polygons in the middle of the study area have more connections because they are surrounded by other polygons from all sides whereas polygons at the edges have less connections because they do not have polygons on at least one of their sides. Also, some larger polygons share edge with several smaller polygons, such polygons are also observed to have greater number of connections in the standard adjacency graph. Contrary to this graph, number of neighbors for each polygon in the nearest neighbor graph remain 4 or just slightly more than 4. No drastic difference between number of neighbors among polygons are observed for nearest neighbor graph in contrast to standard adjacency graph.

Binary nearest neighbors are not necessarily symmetric because one polygon can be nearest neighbor of another polygon but the vice versa may not be always true and therefore binary nearest neighbors are not necessarily symmetric.

Task 3 (1 points): Based the neighbor's adjacency link structure in the *W*-coding scheme evaluate the standard deviate $z\left(I^{observed}\right) = \frac{I^{observed} - E(I|H_0)}{\sqrt{Var(I|H_0)}}$ of the global Moran's statistic for the residuals of the sequence of regression models:

Model No.	Model	Moran's I standard deviate
Model1	**PCTNOHINS~1**	19.772
Model2	**PCTNOHINS~PCTHISPAN**	14.707
Model3	**PCTNOHINS~PCTHISPAN+PCTUNIVDEG**	13.706
Model4	**PCTNOHINS~PCTHISPAN+PCTUNIVDEG+PCTFAMPOV**	10.234

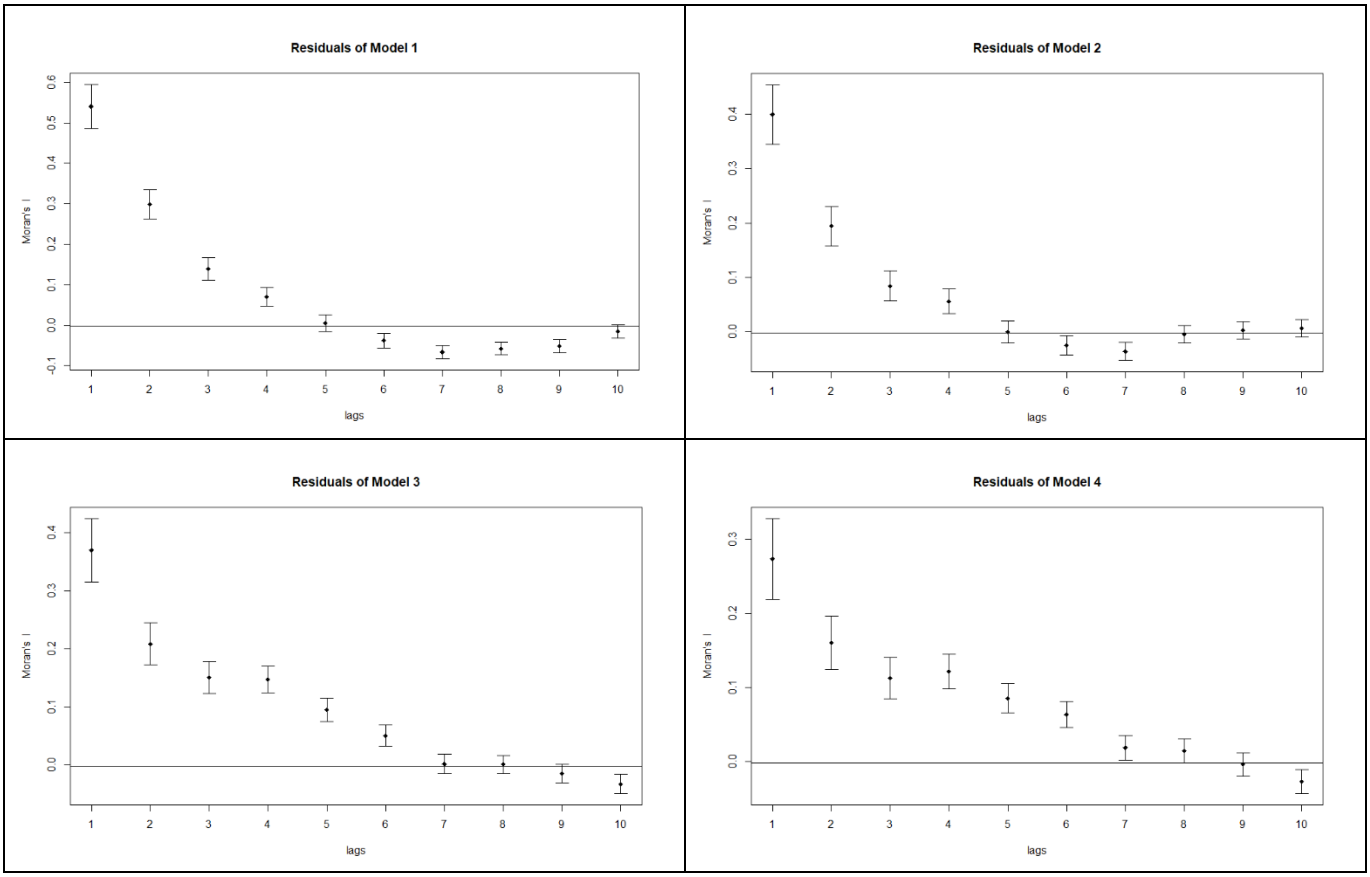
Explain why usually the autocorrelation level shrinks as additional exogeneous variables are added to the model.

Also interpret the correlogram up to lag 10.

```
tracts.linkW <- nb2listw(tracts.link, style = "W")

lm1 <- lm(PCTNOHINS~1,data = mod_tractShp)

lm.morantest(lm1, tracts.linkW)

lm2 <- lm(PCTNOHINS~PCTHISPAN,data = mod_tractShp)

lm.morantest(lm2, tracts.linkW)

lm3 <- lm(PCTNOHINS~PCTHISPAN+PCTUNIVDEG,data = mod_tractShp)

lm.morantest(lm3, tracts.linkW)

lm4 <- lm(PCTNOHINS~PCTHISPAN+PCTUNIVDEG+PCTFAMPOV,data =
mod_tractShp)

lm.morantest(lm4, tracts.linkW)

# correlogram

plot(sp.correlogram(tracts.link, residuals(lm1), order=10, method="I",
style = "W"), main = "Residuals of Model 1")

plot(sp.correlogram(tracts.link, residuals(lm2), order=10, method="I",
style = "W"), main = "Residuals of Model 2")

plot(sp.correlogram(tracts.link, residuals(lm3), order=10, method="I",
style = "W"), main = "Residuals of Model 3")

plot(sp.correlogram(tracts.link, residuals(lm4), order=10, method="I",
style = "W"), main = "Residuals of Model 4")
```

In the table above, with each added exogeneous variable, the standard deviate of global Moran's statistics decreased. That is because added variables started capturing the data generating process of variable PCTNOHINS.

Usually the spatial autocorrelation in residuals decreases as number of exogeneous variables are increased. That is because usually any spatial phenomenon has certain degree of spatial autocorrelation present in it. If we fail to explain the spatial processes that are responsible for causing this spatial autocorrelation via regression models, it reflects in the residuals. But as we add variables which may explain the underlying data generating process, the spatial autocorrelation in the residuals decreases as we already some of it by adding appropriate variables. As we get closer and closer to including most of the variables causing spatial variation in the data, these variables explain the variability and thus that variability no longer get added to residuals. Therefore, as we increase the number of exogeneous variables, the spatial autocorrelation in the residuals decreases.



The above 4 figures display correlogram upto lag 10 for model1, model2, model3 and model4 in the table. In all 4 correlograms, the spatial autocorrelation in the residuals has consistently decreased upto lag 7. The residuals in model1 and model2 found to show increasing spatial autocorrelation after lag 7.

Residuals in the model1 has highest spatial autocorrelation followed by model2, model3 and model4. As we increase number of exogeneous variables in the model that can explain the spatial variation in the % of population without health insurance, it results in a reduced spatial autocorrelation in the residuals.

For example, in a census tract with higher % of population with university degree chances of not having health insurance is less since people with university degree will have better income to afford health insurance. Secondly census tracts with more % of families below poverty threshold may not have means to pay for the health insurance. Thus, as we add these variables which may explains part of underlying spatial data generating process, the more and more variation in the dependent variables gets captured and less of it gets added in the resulting residuals. Consequently, less spatial autocorrelation in the residuals. Model1 has no exogeneous variable, model2, model3 and model4 we keep on adding meaning variables and hence as we move from model with no exogeneous variable to model with most exogeneous variables, spatial autocorrelation in the residuals decreases. The same can be seen in the correlogram displayed above.

Task 4 (2 points): Calibrate a SAR model with **`errorsarlm( )`** and a LAG model with **`lagsarlm( )`**.

Compare the estimated regression coefficients of both models with those of the OLS model.

Test the residuals of the selected spatial autocorrelation model with the function **`moran.mc( )`**.

Why are the residuals expected to be spatially uncorrelated for a ***properly*** specified model?

Variable	Model Regression Coefficients		
	SAR (errorsarlm)	LAG (lagsarlm)	OLS
INTERCEPT	14.80	8.3	11.35
PCTHISPAN	0.22	0.20	0.22
PCTUNIVDEG	-0.17	-0.083	-0.10
PCTFAMPOV	0.24	0.30	0.33

In both SAR and OLS model, there is not much difference between the value of intercept. The intercept of LAG model is less than SAR and OLS models. However, in all three cases the intercept is + ve. Similarly, the sign for coefficients associated with all other variable is same in all three models and the values of these coefficients differ only slightly. Thus, it can be deduced that coefficients of spatial models do not differ substantially from aspatial OLS models in this particular example.

```
B <- spdep::poly2nb(mod_tractShp, queen=F)

## Spatial autoregressive model SEM (SAR)

tracts.SAR <- errorsarlm(PCTNOHINS~PCTHISPAN+PCTUNIVDEG+PCTFAMPOV,data
= mod_tractShp,
                         listw=nb2listw(B,style="W"))

summary(tracts.SAR)
```

```
# Spatial LAG Model

tracts.LAG <- lagsarlm(PCTNOHINS~PCTHISPAN+PCTUNIVDEG+PCTFAMPOV,data =
mod_tractShp,

                                    type="lag",
listw=nb2listw(B,style="W"))

summary(tracts.LAG)

# OLS Model

summary(lm4)

moran.mc(residuals(tracts.SAR), tracts.linkW, nsim=9999)
```

Code Output:

```
> summary(tracts.SAR)

Call:errorsarlm(formula = PCTNOHINS ~ PCTHISPAN + PCTUNIVDEG + PCTFAMPOV,
    data = mod_tractShp, listw = nb2listw(B, style = "W"))

Residuals:
      Min        1Q    Median        3Q       Max
-14.92615  -3.27695  -0.20668   2.63428  18.90294

Type: error
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 14.806409   1.525175  9.7080 < 2.2e-16
PCTHISPAN    0.221679   0.017927 12.3655 < 2.2e-16
PCTUNIVDEG  -0.174783   0.022671 -7.7096 1.266e-14
PCTFAMPOV    0.240880   0.030303  7.9490 1.776e-15

Lambda: 0.53451, LR test value: 85.156, p-value: < 2.22e-16
Asymptotic standard error: 0.049252
    z-value: 10.853, p-value: < 2.22e-16
Wald statistic: 117.78, p-value: < 2.22e-16

Log likelihood: -1610.366 for error model
ML residual variance (sigma squared): 24.744, (sigma: 4.9743)
Number of observations: 527
Number of parameters estimated: 6
AIC: 3232.7, (AIC for lm: 3315.9)


> summary(tracts.LAG)

Call:lagsarlm(formula = PCTNOHINS ~ PCTHISPAN + PCTUNIVDEG + PCTFAMPOV,
    data = mod_tractShp, listw = nb2listw(B, style = "W"), type = "lag")

Residuals:
      Min        1Q    Median        3Q       Max
-14.09627  -3.23164  -0.60113   3.28093  21.51331

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  8.308039   1.458287  5.6971 1.218e-08
```

```
PCTHISPAN     0.201739   0.015099 13.3613 < 2.2e-16
PCTUNIVDEG   -0.083958   0.018264 -4.5970 4.286e-06
PCTFAMPOV     0.307311   0.028800 10.6706 < 2.2e-16

Rho: 0.16449, LR test value: 15.111, p-value: 0.00010137
Asymptotic standard error: 0.042653
    z-value: 3.8564, p-value: 0.00011508
Wald statistic: 14.872, p-value: 0.00011508

Log likelihood: -1645.388 for lag model
ML residual variance (sigma squared): 29.997, (sigma: 5.477)
Number of observations: 527
Number of parameters estimated: 6
AIC: 3302.8, (AIC for lm: 3315.9)
LM test for residual autocorrelation
test value: 71.751, p-value: < 2.22e-16
```

```
> summary(lm4)

Call:
lm(formula = PCTNOHINS ~ PCTHISPAN + PCTUNIVDEG + PCTFAMPOV,
    data = mod_tractShp)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4413  -3.2325  -0.5993   3.1441  21.1011

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.35903    1.19900   9.474  < 2e-16 ***
PCTHISPAN    0.22087    0.01438  15.361  < 2e-16 ***
PCTUNIVDEG  -0.10427    0.01716  -6.075 2.39e-09 ***
PCTFAMPOV    0.33174    0.02913  11.390  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.592 on 523 degrees of freedom
Multiple R-squared:  0.7704,   Adjusted R-squared:  0.7691
F-statistic: 585.1 on 3 and 523 DF,  p-value: < 2.2e-16
```

Since the data we are modeling is the spatial data, by using OLS model specification we will be violating the assumption that the observations are independent. Out of two spatial models, I select the tracts.SAR model specified using errorsarlm() function. That is because lesser the AIC, parsimonious the model is. Tracts.SAR model has less AIC than tracts.LAG model. Also, the likelihood ratio test value for tracts.SAR model is greater than tracts.LAG model. That means the SAR model fits the data better.

Code Output:

```
> moran.mc(residuals(tracts.SAR), tracts.linkW, nsim=9999)

        Monte-Carlo simulation of Moran I

data:  residuals(tracts.SAR)
weights: tracts.linkW
number of simulations + 1: 10000

statistic = -0.037214, observed rank = 986, p-value = 0.9014
alternative hypothesis: greater
```

Here the p-value is greater than 0.05, therefore we fail to reject the null hypothesis that the residuals are randomly distributed. This confirms that there is no spatial autocorrelation in the residuals and hence the tracts.SAR model accounts for the spatial autocorrelation in the data.

The residuals are expected to be spatial uncorrelated for properly specified model because either the independent variables considered in the model or the proper spatial model specification accounts for the spatial variation present in the dependent variable and hence spatial autocorrelation in the dependent variable no longer reflects in the residuals.

Part III: Local autocorrelation (4 points)

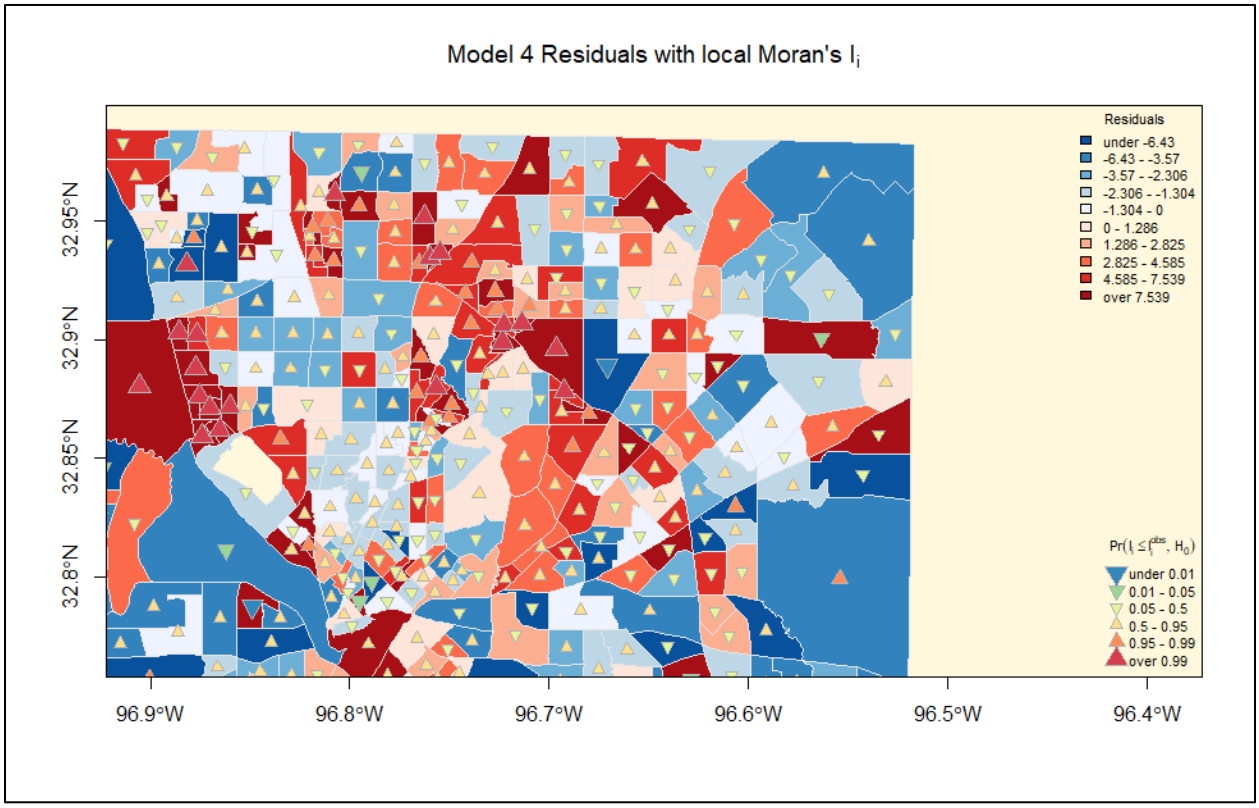
Task 5 (2 points): For the model with **PCTNOHINS~PCTHISPAN+PCTUNIVDEG+PCTFAMPOV** under the assumption of spatial independence calculated the cumulative probabilities $\Pr(I_i \leq I_i^{observed} | \mathbf{X}, \sigma^2 \cdot \mathbf{I})$ of local Moran's I_i for all areas $i \in \{1, \ldots, n\}$ with the saddlepoint approximation. Use the neighbor's adjacency link structure in the *W*-coding scheme.

Map the residuals with a bipolar map theme for the sub-region of the ***north-east quadrant*** of Dallas county and superimpose map symbols for the cumulative probabilities of local Moran's I_i.

Interpret your map by looking for significant clusters and/or hot spots.

```
load("E:\\Courses\\Spring2020\\TexMix_0.5.1\\TexMix\\data\\bndShp.RData")

zoom <- bbox(bndShp)

zoom[1,1] <- -96.77763

zoom[2,1] <- 32.76727

plot(bndShp, col="cornsilk", axes=T, xlim=zoom[1,], ylim=zoom[2,], bg="cornsilk")


High.locM <- summary(localmoran.sad(lm4, tracts.link, style="W",

                    resfun=residuals, alternative="less", select=seq_along(mod_tractShp)))


mapBiPolar(residuals(lm4), mod_tractShp,neg.breaks=5, pos.breaks=5, break.value=0,

  map.title=expression(paste("Model 4 Residuals with local Moran's ", I[i])),

  legend.title="Residuals", legend.cex= 0.7, legend.pos="topright", add.to.map=T)
```

```
mapProbsAdd(High.locM[["Pr. (Sad)"]], mod_tractShp,
legend.pos="bottomright",
              legend.cex=0.7, symbol.cex=1.5)
```

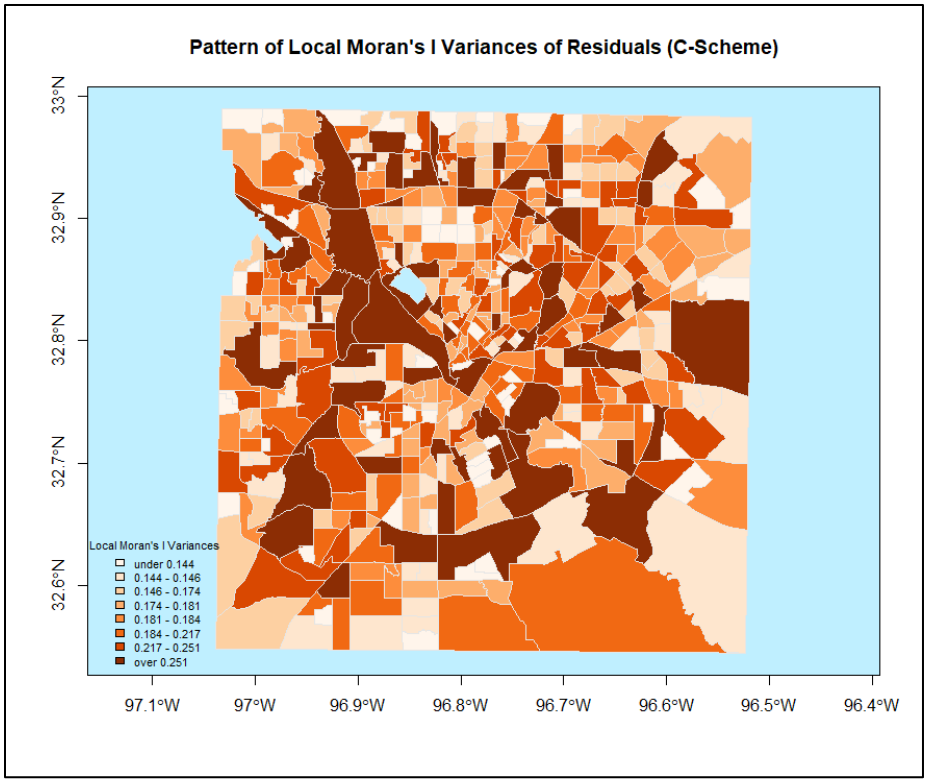


In most part of the maps, significant spatial clustering is observed. For example, autocorrelated residuals with positive value at about middle portion of the left edge of the study area represent significant spatial cluster. There are two polygons in the south-east direction of this cluster which also show spatial clustering of residuals with positive value. Similarly, there is spatial cluster of positive residuals in the middle of the study area and there are several significant spatial clusters in the north-east direction of this cluster. In addition, there are also spatial clusters of negative residuals. In the southeast corner of the study area, there are two polygons which show significant spatial clustering of residuals with negative value. Two polygons are also observed in the near the north-west corner of the study area which are significant spatial clusters.

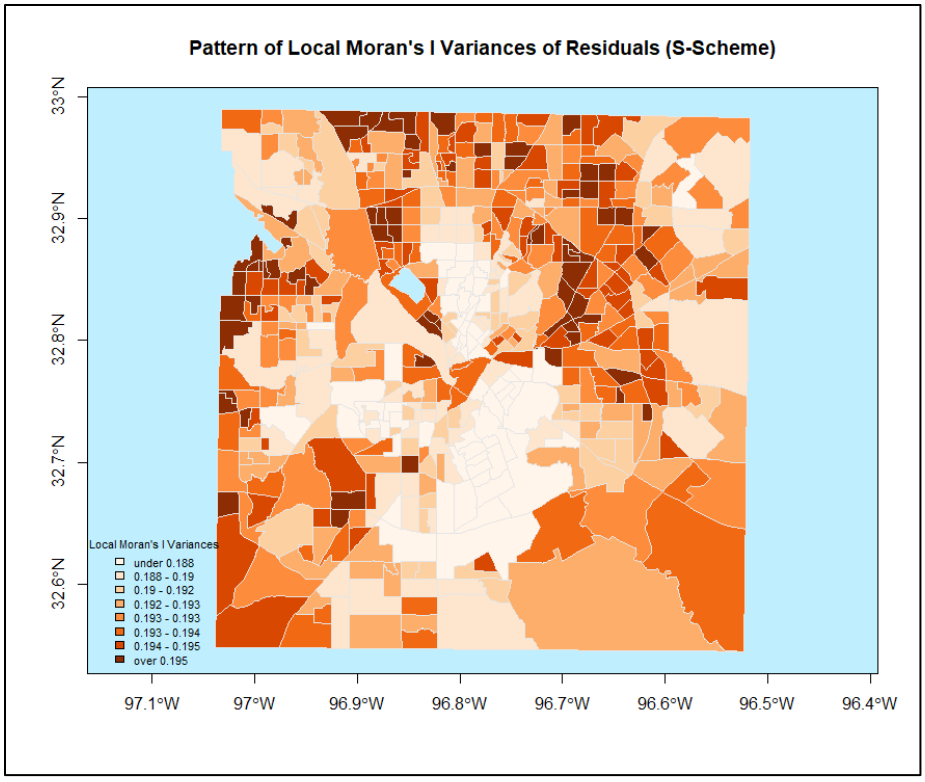
There are also significant spatial outliers/hot spots in some parts of the study area. In the south-west quadrant of the map, there are at least 4 polygons which are spatial outliers since they are surrounded by residuals with opposite sign of their own residuals. In addition, there are two polygons in the north-east quadrant of the study area which are also significant spatial outliers.

Task 6 (1 points): Calculate for the model in task 5 the variances of local Moran's I_i in the *C*-, *S*-, and *W*-coding schemes and map these variances with a gradient color scheme.

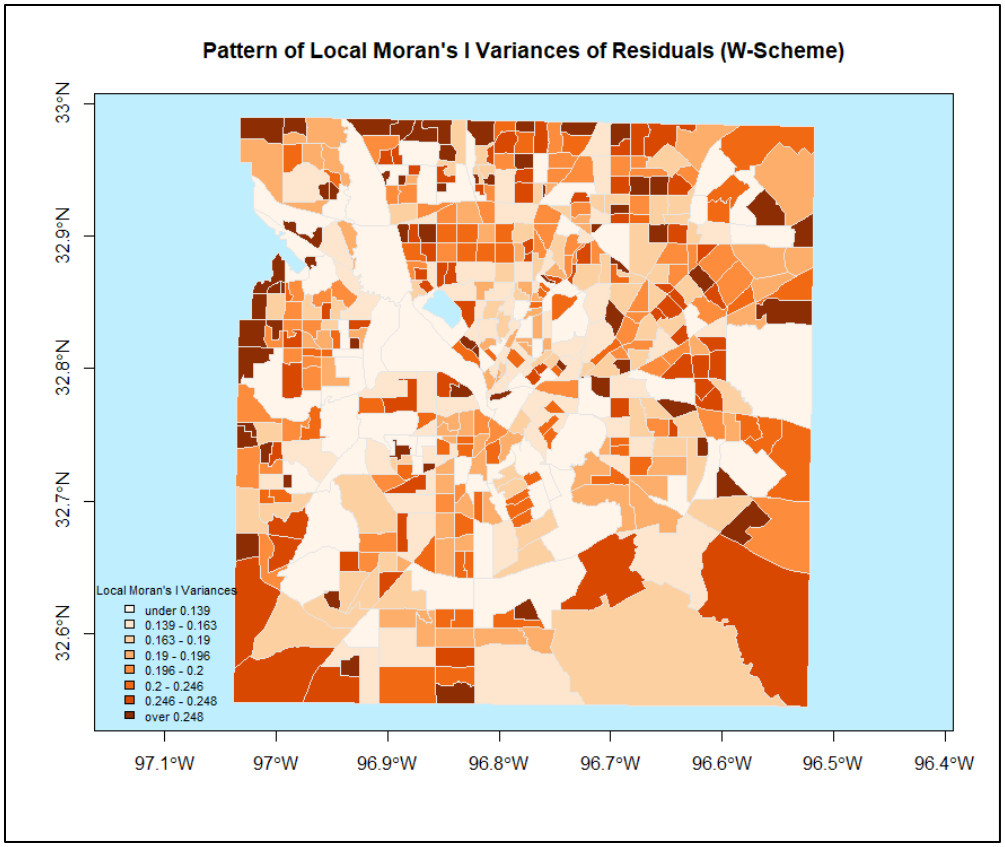
Does a specific pattern emerge?



In a C-coding scheme, for the Dallas county census tracts, larger tracts surrounded by relatively smaller tracts have high local Moran's I variance. Similarly, smaller census tracts surrounded by relatively larger census tracts have less value of local Moran's I variance. Census tracts with few neighbors appears to have less for local Moran's I variance and vice-versa. Most of the census tracts at the edges have received lower value for local Moran's I variance since they are surrounded by few polygons as compared to census tracts in the interior part of the study area. Additionally, overall pattern of local Moran's I variance using C-scheme appears very heterogeneous.



In contrast to C and W coding scheme, this pattern obtained using S-scheme appears homogeneous and seems to show greater degree of spatial autocorrelation. The values of variances vary in a very narrow range of 0.188-0.185. Whereas for C-scheme this range is 0.144-0.251 and for W-coding scheme, it is 0.139-0.248.



The pattern of local Moran's I variances obtained using W-coding scheme appears to be opposite of that obtained using C-coding scheme. That is because, although overall pattern is heterogeneous, here census tracts at the edges of the study are show greater local Moran's I variances as compared to census tracts in the interior part of the study area. In this pattern, census tracts with a smaller number of neighbors appears to have higher value of local Moran's I variances and vice-versa.

Task 7 (1 point): Discuss how the variance of local Moran's I_i in the different coding schemes influences the contribution of individual areas on global Moran's I. Recall $I = \frac{1}{n}\cdot\sum_{i=1}^{n} I_i$. See also Golgher *et al.* (2016 page 191).

In a C-coding scheme, polygons with greater number of neighbors have greater local Moran's I variance as compared to polygons with a smaller number of neighbors. Reverse is the case for W-coding scheme whereas for S-coding scheme variances remains almost same across all polygons irrespective of the number of neighbors. As a result, in a C-coding scheme polygon with greater number of neighbors influence global Moran's I more than polygons with smaller number of neighbors. Likewise, in a W-coding scheme polygon with a high value of local Moran's I variance influence global Moran's I greatly than those with less variance value. More the variance, greater is the z score and hence larger the value of local Moran's I. The global Moran's I is the nothing but mean of local Moran's I across all polygons. We know that in a simple mean, extreme values contribute more to mean. Therefore, since polygons with greater local Moran's I variances have greater probability of obtaining higher value of Moran's I, those polygons contribute more to global Moran's I. Given the relation between connectivity and local Moran's I variance for W-coding scheme, it means that polygons with smaller number of neighbors influence global Moran's I more than those with larger number of neighbors. Thus, in a C-scheme

polygon at the interior part of study area influence global Moran's I more whereas in a W-coding scheme, polygons at the edge of the study area influence more. In contrast to W and C coding scheme connectivity does not affect S-coding scheme and hence all polygons have nearly same influence on global Moran's I.