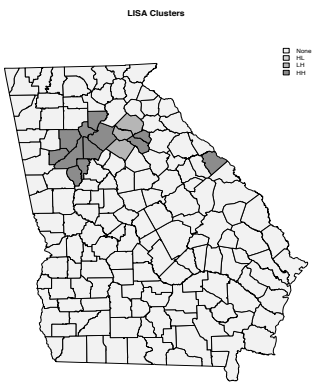
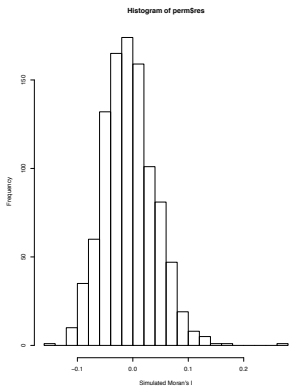
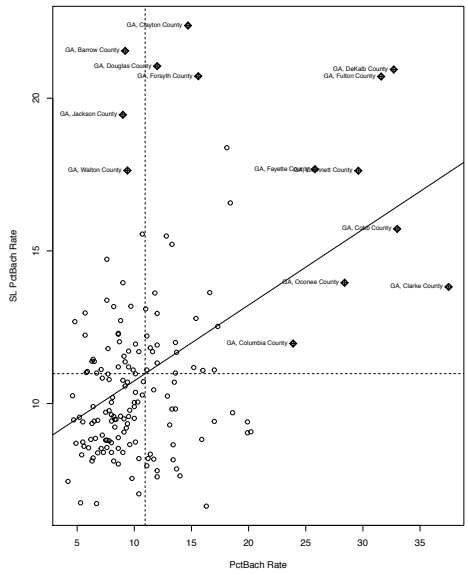
Alexandra Plassaras (amp2261)
G4071(Advanced GIS) Final Exam (Spring 2016)
Due May 13th (Midnight)

1) We are interested in the county level correlates of educational attainment. Use the data concerning educational attainment in GA to test for potential spatial dependence in the DV PctBach (Percent with a Bachelor's degree)….. include any tables, maps, graphs, or screen shots of information referenced in the responses to the below questions.

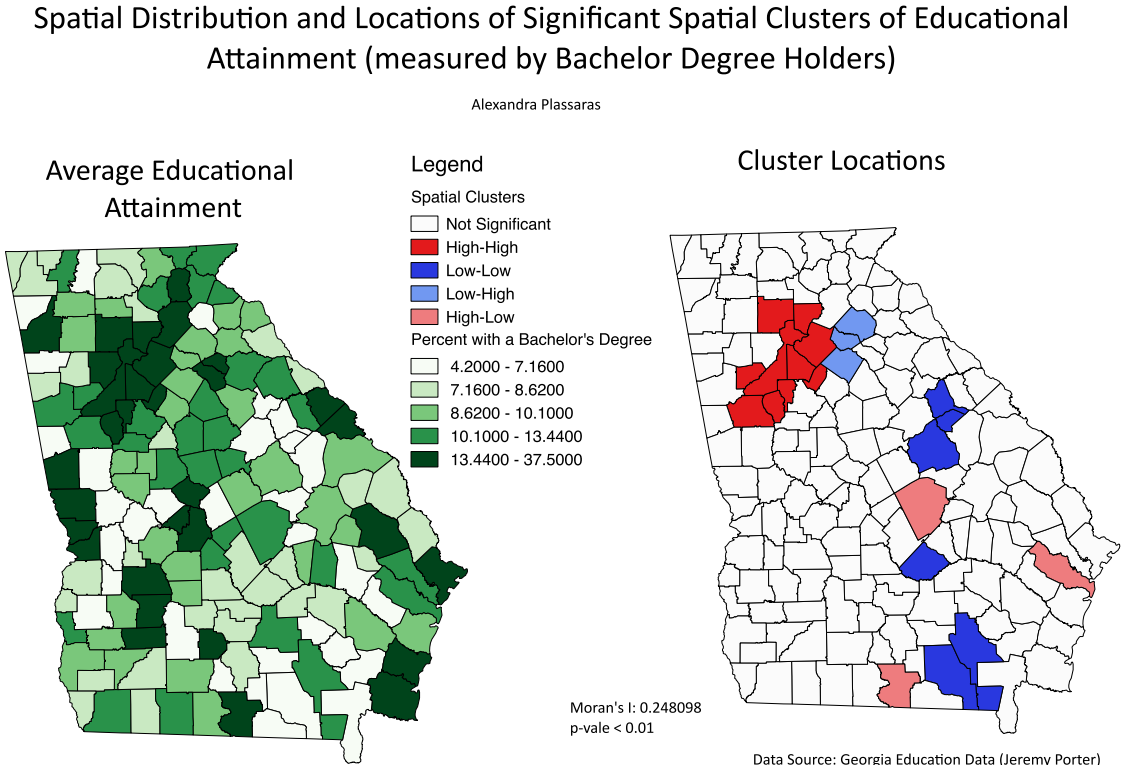
- **Is there spatial dependence in the data?**
  From the images below we can see visually that there does appear to be spatial dependence in out data. Counties such as Clayton County, Fulton County and Cobb County all appear to have high values of PctBach surrounded by high values.

<table>
<tr><td colspan="2">Spatial Dependence Results:</td></tr>
<tr><td>Geographic Distribution of Significant Spatial Clusters</td><td>Simulated Moran's I Distribution</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td colspan="2">Scatterplot of PctBach rates by Spatially Lagged PctBach Rates</td></tr>
<tr><td colspan="2"> 
</td></tr>
</table>

- **What is the magnitude of any potential relationship and is it significant? How do you know?**
  The magnitude of any potential relationship can be confirmed using Moran's I which is used to test for the presence of spatial dependence. From the R output below we can see that the Moran's I value is 0.248098122, which is statistically significant at the 1% level. This means that our Moran's I value tells us that there is positive spatial autocorrelation and that clusters are occurring that are not due by sheer randomness.

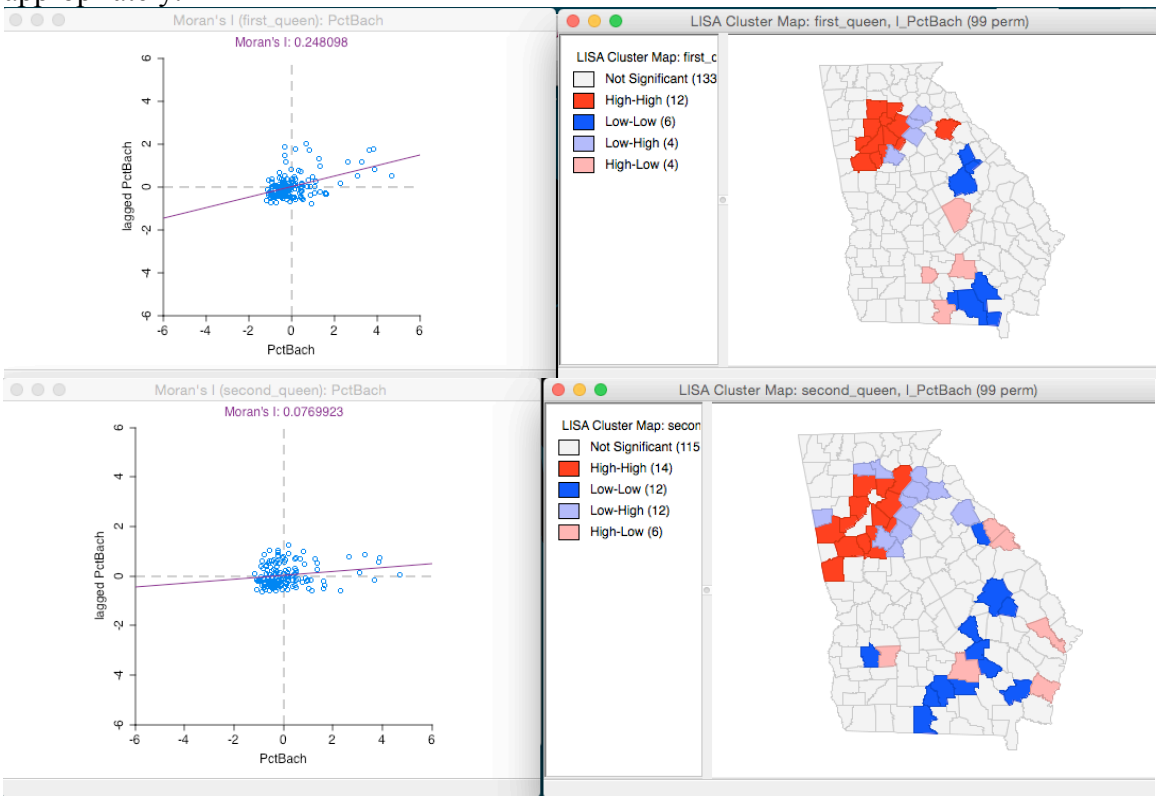
```
> moran.test(ga_edu$PctBach, listw=ga_edu_nbq_w)

        Moran's I test under randomisation

data:  ga_edu$PctBach
weights: ga_edu_nbq_w

Moran I statistic standard deviate = 5.2975, p-value = 5.869e-08
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.248098122      -0.006329114       0.002306653
```

- **Create a Map in QGIS presenting the locations of the significant clusters and using the standard color identifiers for each of the cluster codes listed.**



2) Using the same data above (Georgia Education data)…. examine the effects the percent rural (pctrural), the percent in poverty (pctpov), and the percent black (pctblack). Be sure to identify the proper spatially weighted regression model and give support for that decision. Also, interpret all results including the spatial parameter.

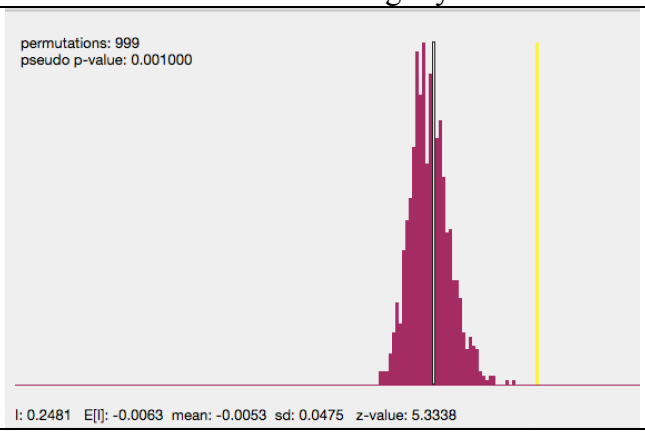
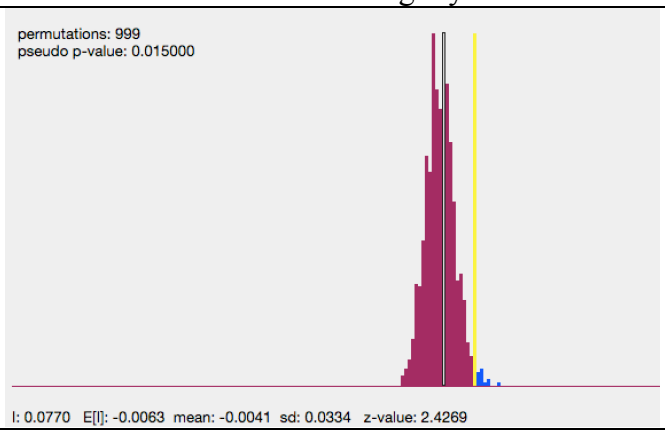
To examine the effects pctrural, pctpov, and pctblack have on pctbach we first create our 1st and 2nd order contiguities to ensure that we've captured the spatial process appropriately.



From the results we can see that there are differences in the ability to detect spatial dependence given the assumed process and associated definition used. The 1st order results who well defined clusters of high and low percentages of bachelor degree holders across the state while the 2nd order map is much less definitive. The Moran's I which is the most useful piece of information to compare the two contiguities shows us that the 1st order is a much better at showing a positive spatial autocorrelation and that clusters are occurring that are not due by sheer randomness (0.248098 compared to the 2nd order's Moran's I of 0.0769923).

By testing the significance of the statistics we can see that the Moran's I associated with the detection spatial dependence at the 1st order is statistically significant given a probability of associate with making a mistake when reporting spatial dependence being 0.1% (p-value = 0.001000). On the other hand, the p-value associated with the 2nd order

weight increases that probability to 1.5% (p-value = 0.0150). Given these results we will no longer use the 2nd order weight as it does not capture the spatial connectivity of the process being examined as well as the 1st order weight.

1st Order Contiguity	2nd Order Contiguity
	

Now that we have identified the proper spatially weighted regression model we implement a simple bivariate regression to see if pctrural, pctpov and pctblack are related to pctbach in both OLS (with spatial diagnostics) and Spatial Regression form.

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  GeorgiaEduc
Dependent Variable  :     PctBach  Number of Observations:  159
Mean dependent var  :     10.9472  Number of Variables   :    4
S.D. dependent var  :     5.67909  Degrees of Freedom    :  155

R-squared           :    0.485273  F-statistic           :     48.7102
Adjusted R-squared  :    0.475311  Prob(F-statistic)     :  3.1068e-22
Sum squared residual:     2639.56  Log likelihood        :    -448.964
Sigma-square        :     17.0294  Akaike info criterion :     905.927
S.E. of regression  :     4.12667  Schwarz criterion     :     918.203
Sigma-square ML     :      16.601
S.E of regression ML:     4.07443

------------------------------------------------------------------------
    Variable   Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------
    CONSTANT      23.85462       1.173043       20.33566      0.00000
    PctRural    -0.1113945     0.01287848      -8.649661      0.00000
      PctPov    -0.3457784     0.07086291       -4.87954      0.00000
    PctBlack    0.05833108     0.02918745       1.998499      0.04741
------------------------------------------------------------------------
```

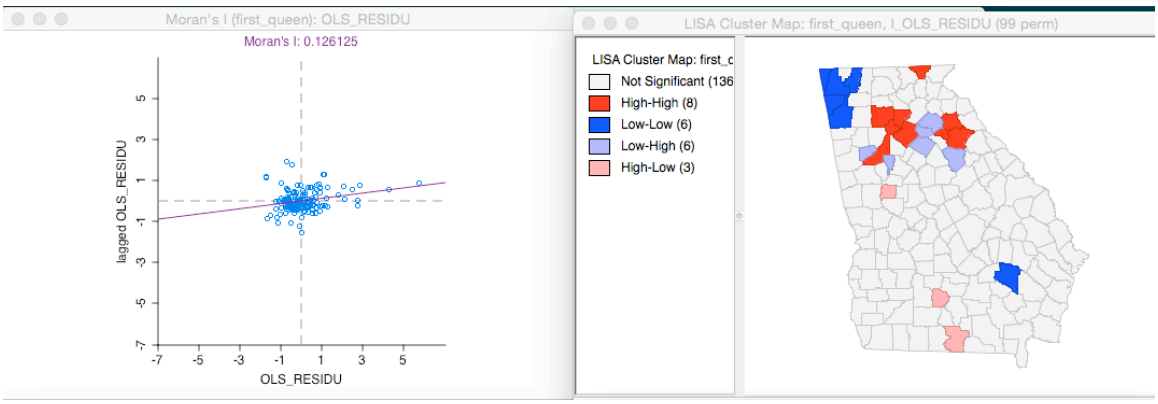
From our regression model we see that the percent rural and percent in poverty are both negatively associated (significantly, p-value < 0.01) while percent black is positively associated (significantly, p-value < 0.05). Thus, for every additional percent a county is in poverty, we can expect a decrease of 34.58 percentage points for people holding a bachelors degree. For every additional percent a county is rural, we can expect a decrease of 11.14 percentage points for people holding a bachelors degree. Lastly, for every additional percent a county is black, we can expect an increase of 5.83 percentage points of people holding a bachelors degree. Since these are uncontrolled OLS results there are other models that might help explain the relationship better, but this is a good baseline for understanding the original relationship. We also see that this model can help explain our model by 48.52% (R-squared) or by 47.53% (Adjusted R-squared).

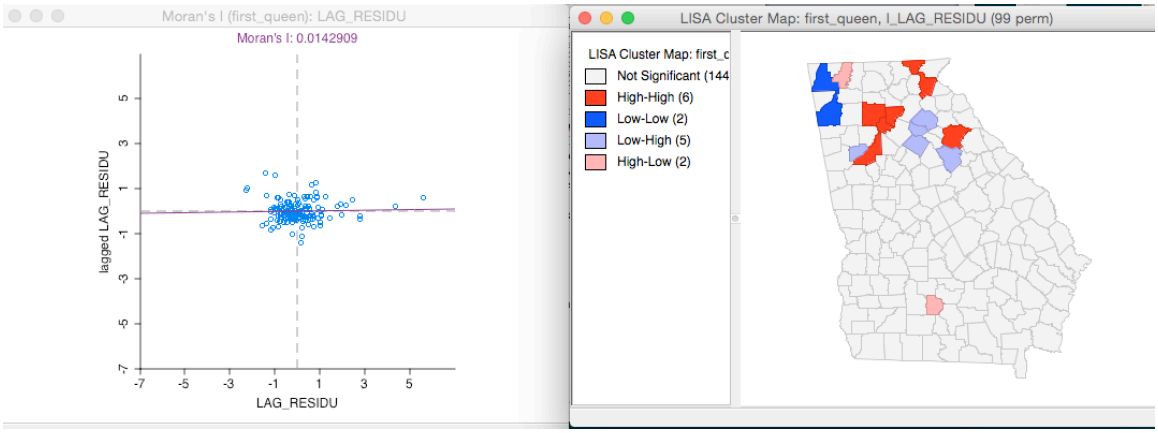
```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : first_queen.gal
   (row-standardized weights)
TEST                          MI/DF        VALUE          PROB
Moran's I (error)             0.1261        2.9274        0.00342
Lagrange Multiplier (lag)        1          6.9517        0.00837
Robust LM (lag)                  1          0.9420        0.33176
Lagrange Multiplier (error)      1          6.4024        0.01140
Robust LM (error)                1          0.3927        0.53087
Lagrange Multiplier (SARMA)      2          7.3444        0.02542
```

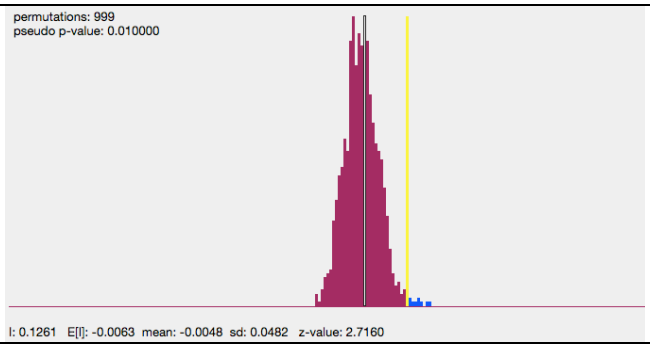
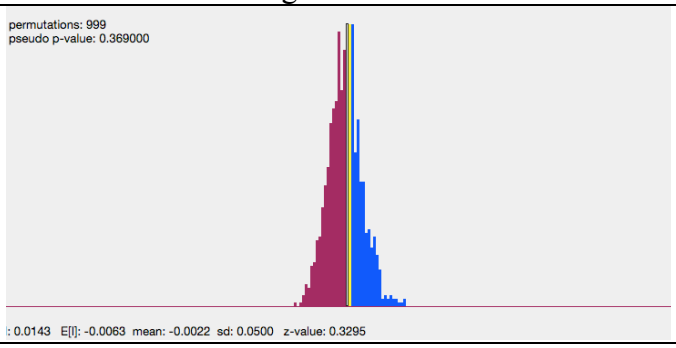
In the results we see above we see that by including the pctblack, pctpov and pctrural in the county we have reduced the 1st order Moran's I value from 0.248098 (taken from the scatterplot above) to 0.1261. So some of the spatial dependence in the variable is actually a product of the fact that county level rural, poverty and black rates are clustered in space. We also see that the Lagrange Multiplier lag and error tests indicate that both lag and error dependence is prevalent (p-values for both are < 0.05). However when controlling for error and lag dependence they become insignificant (p-values: 0.33176 and 0.53087 respectively). These results indicate that we should run a spatial lag regression model with the 1st order weight to appropriately account for the existing spatial dependence.

```
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : GeorgiaEduc
Spatial Weight      : first_queen.gal
Dependent Variable  :     PctBach  Number of Observations:  159
Mean dependent var  :     10.9472  Number of Variables   :    5
S.D. dependent var  :     5.67909  Degrees of Freedom    :  154
Lag coeff.   (Rho)  :    0.264441

R-squared           :    0.514097  Log likelihood        :   -445.502
Sq. Correlation     : -            Akaike info criterion :    901.005
Sigma-square        :     15.6714  Schwarz criterion     :    916.349
S.E of regression   :     3.95871

------------------------------------------------------------------------
    Variable    Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------
   W_PctBach     0.2644408     0.09283466       2.848514     0.00439
    CONSTANT      19.57696       1.851361       10.57436     0.00000
    PctRural    -0.1085315     0.01241265      -8.743619     0.00000
      PctPov    -0.2688337     0.07248608      -3.708763     0.00021
    PctBlack    0.04683344     0.02830775       1.654439     0.09804
------------------------------------------------------------------------
```

After calculating our lag regression model we can see that we have increased our R-square from about 49% to 51% variance explained. More interesting, pctbalck is now only mildly significant as a predictor of pctbach at the 10% level (p-value = 0.09804) and the spatial lab of pctbach (W_PctBach) is positively associated with the local value of pctbach. In interpretation, each additional percentage point of people who have earned a bachelor's degree on average for counties that are considered first order neighbors results in a 0.2644408 percentage point increase in the local unit's percentage of bachelor degree holders. Pctrural and pctpov still have a significantly negative associate with the percentage of bachelor degree holders per county.

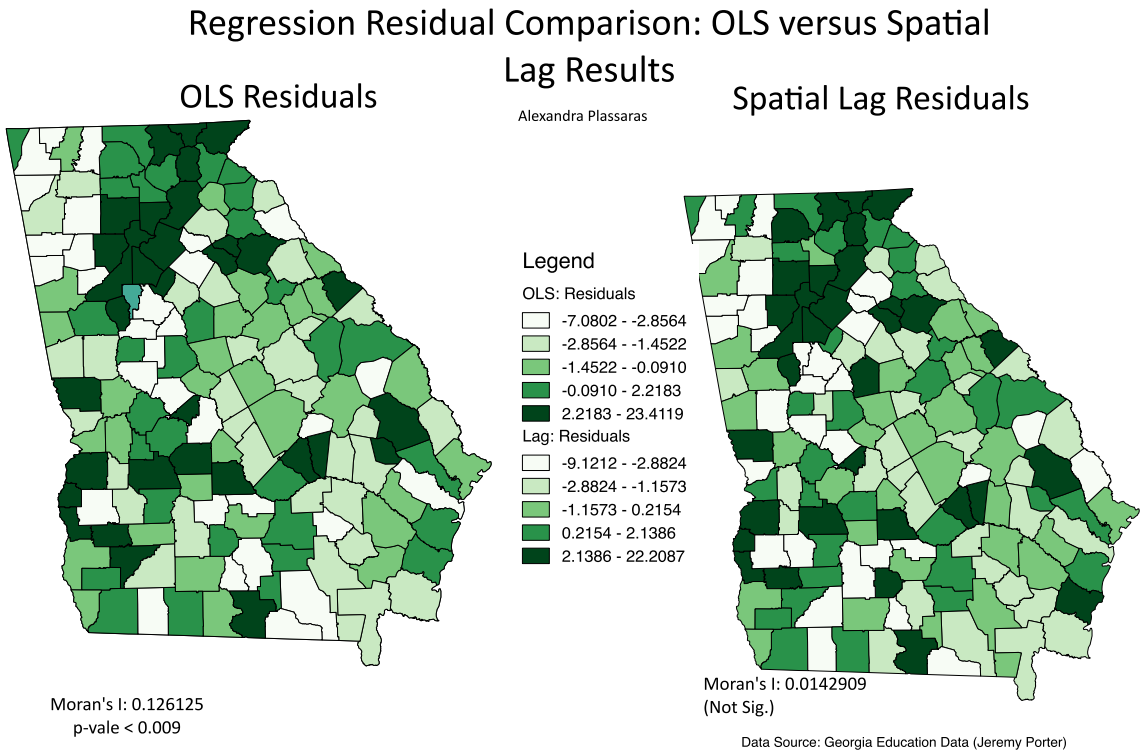




OLS Residuals	Lag Residuals
	

Lastly, after examining the effect of the spatial weight on improving the fit of the model through residual analysis we can see that the OLS residuals are significantly dependent on space (p-value = 0.01000) and produce clusters of areas with high and low residuals which contribute directly to the production of unreliable regression results 'on average'. Once the weight is included in the spatial lag model, the residuals that are produced are now uncorrelated with space and meet the assumptions necessary to produce reliable regression results in relation to identical and independently distributed error terms.

3) In addition to the results above, map the residuals from the OLS (classic) model and the spatially weighted regression model chosen in #2. Briefly interpret the results as a comparison of the two maps. Are the residuals from the OLS clustered? Are the residuals from the spatial model clustered? Provide statistical proof of your response.

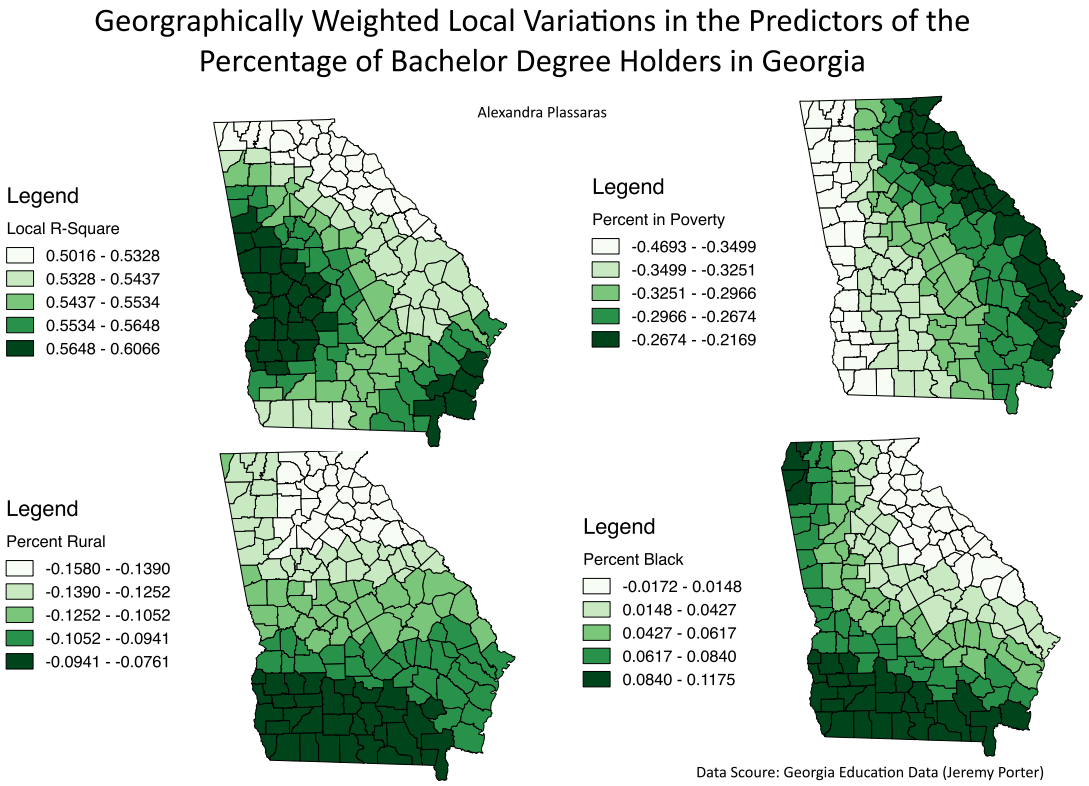


Above we can see both the OLS Residuals and the Spatial Lag Residuals mapped. Here we can see that the OLS and Spatial residuals appear very similar. From the analysis conducted in the question above we can see that the OLS Residual has a higher Moran's I (0.126125) that is statistically significant at the 1% level than the Moran's I for the Lag Residual (0.0142909) which is not statistically significant. Thus we can say that the OLS results are significantly clustered whereas the residuals from the spatial model are not. The statistical proof is provided in question 2 above.

4) Using the same Georgia Education data, examine the data for potential non-stationarity in predicting the percent with a bachelors degree by the percent rural, the percent in poverty, and the percent black. As your deliverable for this question, map the locally varying R-Square coefficients and briefly interpret the strength of the model across the state. Where does the model fit well? Where does the model fit poorly? (and in relative terms).

```
> gwrG
Call:
gwr(formula = PctBach ~ PctRural + PctPov + PctBlack, data = GA_GWR,
    bandwidth = bwG, gweight = gwr.Gauss)
Kernel function: gwr.Gauss
Fixed bandwidth: 127545.7
Summary of GWR coefficient estimates at data points:
                Min.  1st Qu.   Median  3rd Qu.     Max.  Global
X.Intercept. 19.51000 21.32000 24.04000 25.95000 27.51000 23.8546
PctRural     -0.15800 -0.13580 -0.11410 -0.09695 -0.07607 -0.1114
PctPov       -0.46930 -0.34320 -0.31480 -0.27540 -0.21690 -0.3458
PctBlack     -0.01723  0.02405  0.04926  0.07741  0.11750  0.0583
```

From the R output above we can see the summary variations in the model intercept, the β of percent rural, the β of percent in poverty and the β of percent black. The results are visualized below.



As we can see our Georgia education data does show signs of spatial non-stationarity because the same stimuli provokes different responses in different parts of the state. If we compare the GWR residuals form the OLS Residuals above we can see that the GWR residuals are much lower and less spatially dependent. As such our GWR model gives us a better fit to our data. In general we can see that the established relationship between percent of bachelor degree holders and the three other variables, we can see that our local R-Squared model is performing quite well given that the range for our Local R-Squared

values is 0.5016 to 0.6066. Additionally we can see that the Central West region of Georgia has a higher percentage of people holding bachelor degrees. Conversely counties in the north and northeast of the state have lower percentages of people holding bachelor degrees. Looking at our GWR results on pctrural, pctpov and pct black we see that our local regression model is performing poorly in regards to percent rural given that our residual range is from -0.158 to -0.076 as well as for percent in poverty given that the range is -0.4693 to -0.2169. For the residuals looking at percent black however, the model fits decently well given that the range is mostly positive, ranging from -0.0172 to 0.1175.

5) Identify potential space-time clusters in the lung cancer incidence data (lung.geo and lung.cas). Note the data are monthly counts of lung cancer cases ranging 2001 (Jan. 1973) to from 2228 (Dec. 1992). In this case the number 2001 represents the base line month and each increase in that number is an additional month (i.e. 2002 = Feb. 1973; 2004 = March 1973, 2005 = April 1973). The first column in the data is the county name, the second is the actual incidence, and the third is the month code (explained above but I have also given you a data conversion xls file). Once you identify the significant space time clusters, create a map in QGIS from the results by linking the results by county name to the New Mexico shapefile (following along with Lab 9 for guidance). Briefly interpret the space and time locations of any significant clusters in the lung cancer incidence data.

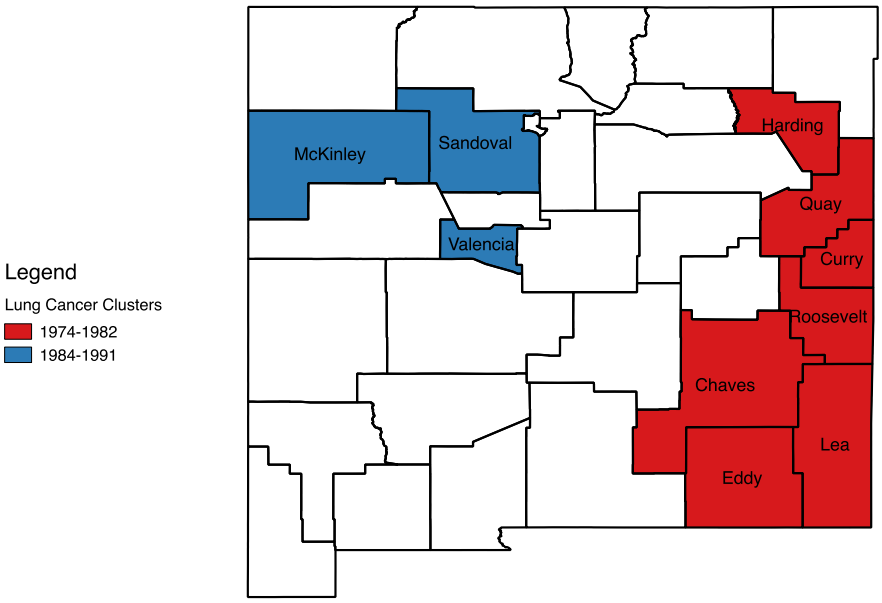
The following space-time clusters were calculated:

```
CLUSTERS DETECTED

1.Location IDs included.: Roosevelt, Curry, DeBaca, Quay, Guadelupe, Chaves, Lea, Harding, Eddy
  Coordinates / radius..: (518,239) / 207.14
  Time frame............: 1974/1/1 to 1982/12/31
  Number of cases.......: 997
  Expected cases........: 880.02
  Observed / expected...: 1.13
  Test statistic........: 8.272132
  P-value...............: 0.012

2.Location IDs included.: SanJuan, RioArriba, McKinley, LosAlamos, Valencia, Sandoval
  Coordinates / radius..: (96,531) / 212.71
  Time frame............: 1984/1/1 to 1991/12/31
  Number of cases.......: 813
  Expected cases........: 716.04
  Observed / expected...: 1.14
  Test statistic........: 6.840192
  P-value...............: 0.094

3.Location IDs included.: Mora, SanMiguel, Taos
  Coordinates / radius..: (335,435) / 52.92
  Time frame............: 1975/1/1 to 1977/12/31
  Number of cases.......: 50
  Expected cases........: 31.77
  Observed / expected...: 1.57
  Test statistic........: 4.461974
  P-value...............: 0.797

4.Location IDs included.: Socorro
  Coordinates / radius..: (199,222) / 0
  Time frame............: 1986/1/1 to 1986/12/31
  Number of cases.......: 13
  Expected cases........: 5.24
  Observed / expected...: 2.48
  Test statistic........: 4.061982
  P-value...............: 0.924
```

Of the four clusters detected, only the first two clusters – Cluster 1 (colored red below) and Cluster 2 (colored blue below) are statistically significant. This first space-time cluster was found in 1974 – 1982 and included Roosevelt, Curry, DeBaca, Quay, Guadelupe, Chave, Lea, Harding and Eddy Counties (p-value < .05). The second space-time cluster was found in 1984-1991 in San Juan, Rio Arriba, McKinley, Los Alamos, Valencia and Sandoval counties (p-value < .10). The other two clusters were not statistically significant enough to be different from random chance. With these results we see that these two time and space areas show significant clustering. Below is a visualization of our statistically significant space-time clusters:

Space-Time Clusters of Lung Cancer in New Mexico

Alexandra Plassaras



Data Source: Jeremy Porter