William Gannon

GEOG 579: Lab 5 Report

Geographically Weighted Regression

Question 1 (1 points): If you choose to work on the homicide data set, the dependent variable can be either homicide rate or homicide count. Which did you choose? And why? If you work on a data set of your own choice, briefly describe the dependent variable in your analysis.

Answer: I want to use the homicide rate(HR) because then I don't need to account for the population in the explanatory variable side of model.

Question 2 (3 points): There are lots of demographic variables in the homicide data set (or the data set of your own choice). Which of them did you choose as independent variables (for both regular regression and GWR)? Based on what judgement?

Answer: The 3 combination of variables I found and will look into testing for the final map.

39.18 %-FH90 **||** 43.18% RD90, MFIL89 **||** 46.18% RD90, MFIL89, PS90, DV90, BLK90

(FH)% female headed households, (RD) resource deprivation, (MFIL) log of median family income,(PS) population structure, (DV) divorce rate, (BLK) % black

Explanation:

I used a tool available in ArcMap Called 'Exploratory Regression'. From what I read in the documentation it runs a regression analysis on each variable you give and a combination range of the independent variables to return a predicted model that pass diagnostic tests for OLS. Then Runs Global Moran's I on models residuals.

Model Testing for:

- Explanatory variables where all of the coefficients are statistically significant.
- Coefficients reflecting the expected, or at least a justifiable, relationship between each explanatory variable and the dependent variable.
- Explanatory variables that get at different aspects of what you are trying to model (none are redundant; small VIF values less than 7.5)
- Normally distributed residuals indicating your model is free from bias (the Jarque-Bera p-value is not statistically significant)
- Randomly distributed over and under predictions indicating model residuals are normally distributed (the spatial autocorrelation p-value is not statistically significant)

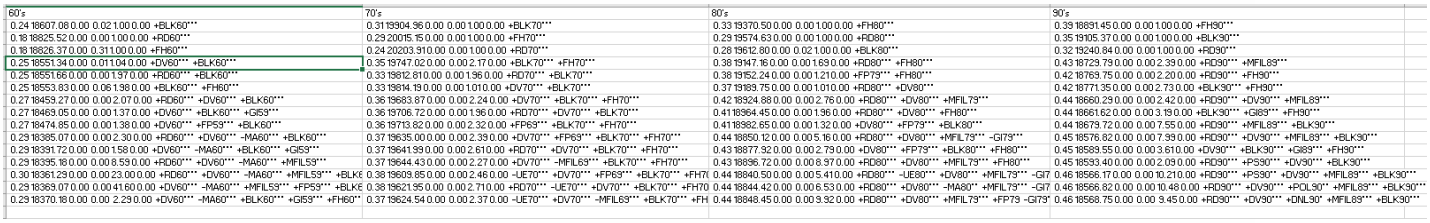
```
    Summary of Variable Significance
Variable % Significant % Negative % Positive
DV90            100.00       0.00     100.00
BLK90            99.82       0.18      99.82
FH90             98.37       0.18      99.82
RD90             97.65       1.08      98.92
GI89             88.97      23.33      76.67
FP89             85.53      32.91      67.09
MFIL89           81.19      30.38      69.62
UE90             68.35      64.92      35.08
POL90            59.30      25.58      74.42
PS90             57.56      23.26      76.74
DNL90            54.84      31.40      68.60
MA90             54.25      54.97      45.03

---------------------------------------------
```

Tool Inputs:

```
Maximum_Number_of_Explanatory_Variables="5"|| Minimum_Number_of_Explanatory_Variables="1",
Minimum_Acceptable_Adj_R_Squared="0.5"     || Maximum_Coefficient_p_value_Cutoff="0.05",
Maximum_VIF_Value_Cutoff="7.5"             || Minimum_Acceptable_Jarque_Bera_p_value="0.1",
Minimum_Acceptable_Spatial_Autocorrelation_p_value="0.1"
```

Now basing this on the out of box settings of the tool listed above I ran it on each of these variables on each year. I did this just by copying the python and changing the years. Here are the results for each year listed as checking with 1 variable up to 5.



60's	70's	80's	90's
0.24 18607.08 0.00 0.02 1.00 0.00 +BLK60***	0.31 13904.96 0.00 0.00 1.00 0.00 +BLK70***	0.33 19370.50 0.00 0.00 1.00 0.00 +FH80***	0.39 18891.45 0.00 0.00 1.00 0.00 +FH90***
0.18 18825.52 0.00 0.00 1.00 0.00 +RD60***	0.29 20015.15 0.00 0.00 1.00 0.00 +FH70***	0.29 19574.63 0.00 0.00 1.00 0.00 +RD80***	0.35 19105.37 0.00 0.00 1.00 0.00 +BLK90***
0.18 18826.37 0.00 0.31 1.00 0.00 +FH60***	0.24 20203.91 0.00 0.00 1.00 0.00 +RD70***	0.28 19612.80 0.00 0.02 1.00 0.00 +BLK80***	0.32 19240.84 0.00 0.00 1.00 0.00 +RD90***
0.25 18551.34 0.00 0.01 1.04 0.00 +DV60*** +BLK60***	0.35 19747.02 0.00 0.00 2.17 0.00 +BLK70*** +FH70***	0.38 19147.16 0.00 0.00 1.69 0.00 +RD80*** +FH80***	0.43 18729.79 0.00 0.00 2.39 0.00 +RD90*** +MFIL89***
0.25 18551.66 0.00 0.00 1.37 0.00 +RD60*** +BLK60***	0.33 19812.81 0.00 0.00 1.96 0.00 +RD70*** +BLK70***	0.38 19152.24 0.00 0.00 1.21 0.00 +FP79*** +FH80***	0.42 18769.75 0.00 0.00 2.20 0.00 +RD90*** +FH90***
0.25 18553.83 0.00 0.06 1.36 0.00 +BLK60*** +FH60***	0.33 19814.19 0.00 0.00 1.01 0.00 +DV70*** +BLK70***	0.37 19183.75 0.00 0.00 1.01 0.00 +RD80*** +DV80***	0.42 18771.35 0.00 0.00 2.73 0.00 +BLK90*** +FH90***
0.27 18459.27 0.00 0.00 2.07 0.00 +RD60*** +DV60*** +BLK60***	0.36 19683.87 0.00 0.00 2.24 0.00 +DV70*** +BLK70*** +FH70***	0.42 18924.88 0.00 0.00 2.76 0.00 +RD80*** +DV80*** +MFIL79***	0.44 18660.29 0.00 0.00 2.42 0.00 +RD90*** +DV90*** +MFIL89***
0.27 18469.05 0.00 0.00 1.37 0.00 +DV60*** +BLK60*** +GI59***	0.36 19706.72 0.00 0.00 1.96 0.00 +RD70*** +DV70*** +BLK70***	0.41 18964.45 0.00 0.00 1.96 0.00 +RD80*** +DV80*** +FH80***	0.44 18661.62 0.00 0.00 3.19 0.00 +BLK90*** +GI89*** +FH90***
0.27 18474.85 0.00 0.00 1.38 0.00 +DV60*** +FP59*** +BLK60***	0.36 19713.82 0.00 0.00 2.32 0.00 +FP69*** +BLK70*** +FH70***	0.41 18982.65 0.00 0.00 1.32 0.00 +DV80*** +FP79*** +BLK80***	0.44 18679.72 0.00 0.00 7.55 0.00 +RD90*** +MFIL89*** +BLK90***
0.29 18385.07 0.00 0.00 2.30 0.00 +RD60*** +DV60*** -MA60*** +BLK60***	0.37 19635.00 0.00 0.00 2.39 0.00 +DV70*** +FP69*** +BLK70*** +FH70***	0.44 18850.12 0.00 0.00 5.16 0.00 +RD80*** +DV80*** +MFIL79*** -GI79***	0.45 18576.82 0.00 0.00 7.39 0.00 +RD90*** +DV90*** +MFIL89*** +BLK90***
0.29 18391.72 0.00 0.00 1.58 0.00 +DV60*** -MA60*** +BLK60*** +GI59***	0.37 19641.99 0.00 0.00 2.61 0.00 +RD70*** +DV70*** +BLK70*** +FH70***	0.43 18877.92 0.00 0.00 2.79 0.00 +DV80*** +FP79*** +BLK80*** +FH80***	0.45 18589.55 0.00 0.00 3.61 0.00 +DV90*** +BLK90*** +GI89*** +FH90***
0.29 18395.18 0.00 0.00 8.59 0.00 +RD60*** +DV60*** -MA60*** +MFIL59***	0.37 19644.43 0.00 0.00 2.27 0.00 +DV70*** -MFIL69*** +BLK70*** +FH70***	0.43 18896.72 0.00 0.00 8.97 0.00 +RD80*** +DV80*** +MFIL79*** +FH80***	0.45 18593.40 0.00 0.00 2.09 0.00 +RD90*** +PS90*** +DV90*** +BLK90***
0.30 18361.29 0.00 0.00 23.00 0.00 +RD60*** +DV60*** -MA60*** +MFIL59*** +BLK6	0.38 19609.85 0.00 0.00 2.46 0.00 -UE70*** +DV70*** +FP69*** +BLK70*** +FH7(	0.44 18840.50 0.00 0.00 5.41 0.00 +RD80*** -UE80*** +DV80*** +MFIL79*** -GI7	0.46 18566.17 0.00 0.00 10.21 0.00 +RD90*** +PS90*** +DV90*** +MFIL89*** +BLK90***
0.29 18369.07 0.00 0.00 41.60 0.00 +DV60*** -MA60*** +MFIL59*** +FP59*** +BLK6	0.38 19621.95 0.00 0.00 2.71 0.00 +RD70*** -UE70*** +DV70*** +BLK70*** +FH7(	0.44 18844.42 0.00 0.00 6.53 0.00 +RD80*** +DV80*** -MA80*** +MFIL79*** -GI7	0.46 18566.82 0.00 0.00 10.48 0.00 +RD90*** +DV90*** +POL90** +MFIL89*** +BLK90***
0.29 18370.18 0.00 0.00 2.29 0.00 +DV60*** -MA60*** +BLK60*** +GI59*** +FH60***	0.37 19624.54 0.00 0.00 2.37 0.00 -UE70*** +DV70*** -MFIL69*** +BLK70*** +FH	0.44 18848.45 0.00 0.00 9.32 0.00 +RD80*** +DV80*** +MFIL79*** +FP79 -GI79	0.46 18568.75 0.00 0.00 9.45 0.00 +RD90*** +DV90*** +DNL90* +MFIL89*** +BLK90***

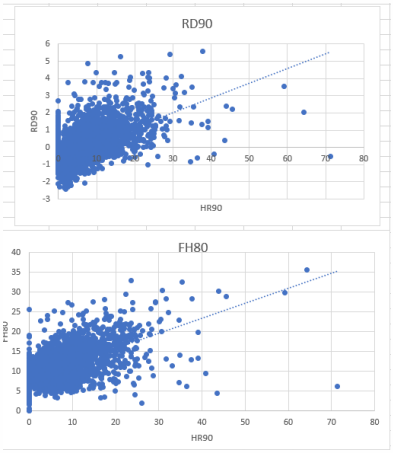
The trend I am seeing is that the more recent and the higher combination of data is yielding to a higher r^2 value. I then decided to up the number of variables used to see if I can get that number higher because now it is getting us between 39%-46% from our models. The result of the testing for higher combinations resulted in no success as any more combinations give out an error due to there being data redundancy from more attributes 46.18% is the highest. The tool also gives out warnings of redundant data and excludes those results I included PS, POL, and DNL which all have to do with population and would correlate with each other closely so it's good to know that portion is working.

ALTERNATE: Correlation Coefficient

I also wanted to double check and ran a correlation coefficient in excel and found similar results. Also running one on the variables from the decade prior to see if maybe there was a correlation because in some cases there could be a lag in cause in effect. Some did and some didn't. Here is the average CC for all decades to HR90:

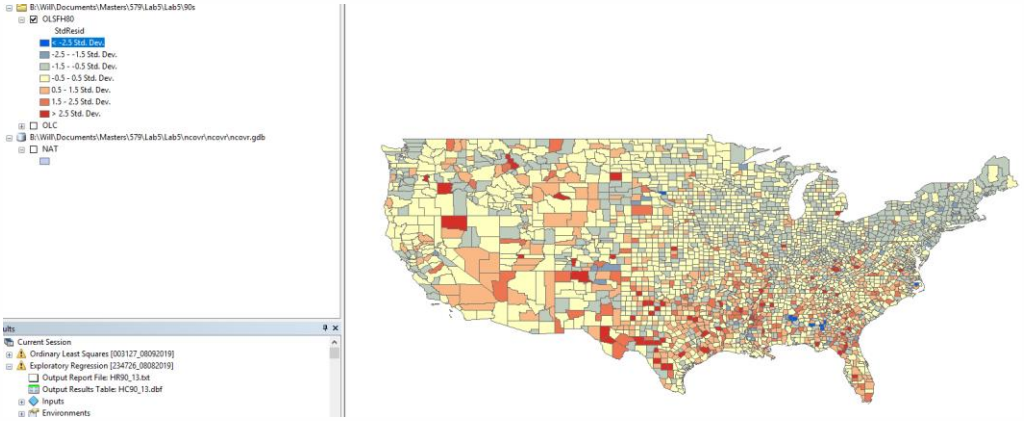
RD	FP	BLK	GI	FH	
Positive	Positive	Positive	Positive	Positive	
0.503384688	0.410878279	0.55874016	0.367032018	0.569178556	
PO	PS	UE	DV	POL	DNL
None-Positive	None-Positive	None-positive	None-positive	None-positive	none-positive
0.133735043	0.104556103	0.100691453	0.175766012	0.10237982	0.09843031
MA	MFIL				
Slightly negative	negative				
-0.215275576	-0.286942069				

Charted Results of chosen attributes. As you can see both positive have a positive correlation coefficient.

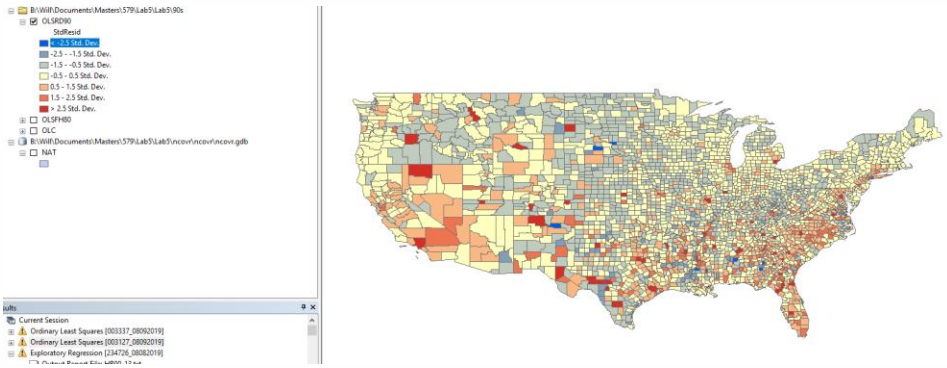


OLS Maps of Each

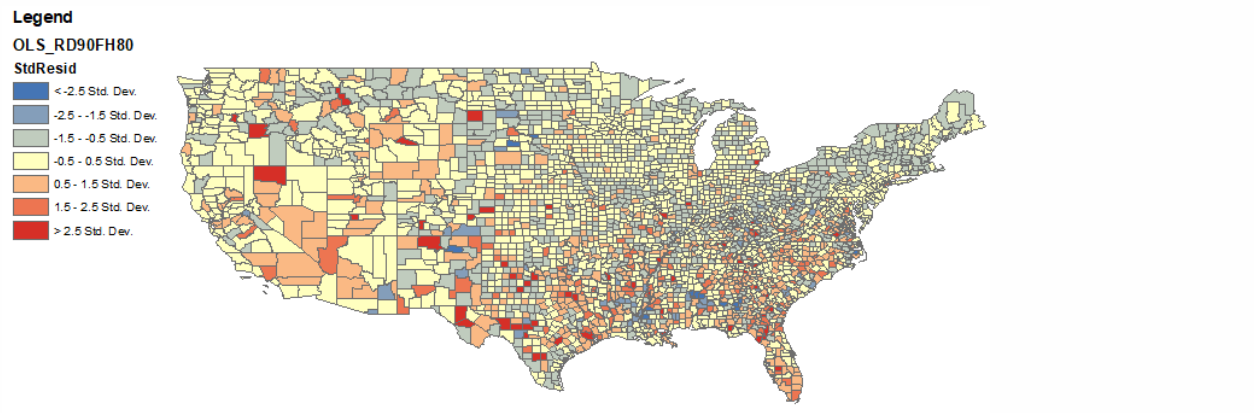
FH80



RD90



Both



Question 3 (3 points): What does each of the parameters mean? How would different settings of the parameters affect the results of GWR? How did you set the parameters? And why?

Answer:

What does each Mean?

In the GWR parameters menu there are several options to take into consideration before running. There is the Dependent and Explanatory variables, of which there are many, which should have been figured out beforehand to avoid some data to be repetitive or just not relevant. There is the Kernel type which decides the distance or number and neighbors to search for value with “Fixed” or “Adaptive”. Finally, there is the bandwidth which also helps choose distance or number of neighbors.

How different setting effect the results?

So after playing around with the GWR tool with our data to fully understand the results, along with some searching on the best ways other people have approached using county data, I came to the conclusion that the setting it to Fixed made the data a bit skewed in places where the data was constrained to small counties vs large counties. This along with the variance of the size of the counties made me investigate what Adaptive had to offer. Using Adaptive meant to automatically choose the number of neighbors based on the bandwidth method which I think is better when using county data to prevent distortion from the size of each county.

The different bandwidth methods to use Akaike Information Criterion, cross validation, and Bandwidth Parameter(choose your own). The ACC and CV choose the number of neighbors automatically based on the respective equation.

How did you set the parameters? And why?

My final choice in choosing Parameters was:

Dependent Variable: HR90

-Because the rate is population adjusted

Explanatory Variable: R90,FH80

-From looking at the Correlation Coefficients and Using the Exploratory Regression Tool.

Kernel Type: Adaptive

-To choose based on neighborhood size not distance

Bandwidth Method: AICc

-To automatically choose neighborhood size since it is easier than testing various sizes myself to observe. CV had a r2=.5396 and 161 Neighborhood Size to AICc r2=.5564 and neighborhood size of 121

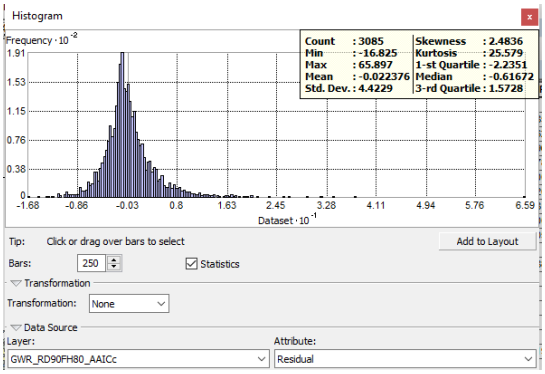
Question 4 (3 points): Plot two histograms showing the frequency distribution of the residuals: one for regular regression and the other for GWR. Describe each histogram in terms of its center, spread, and shape. Is the mean residual close to zero? Does the distribution resemble a normal (Gaussian) distribution? Is it narrow and tall, or wide and flat?

Answer:

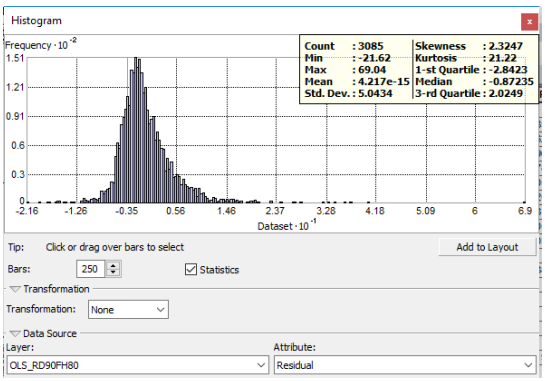
Both using 250 bars.

They both show a bit of a taller and narrow but a Gaussian Distribution. Both are close to 0 in mean having a median slightly negative between -.6 to -.8. The OLS has more of its distribution to the left side towards the negative residuals. Both are also Tall on the left and very flat and long to the right.

GWR



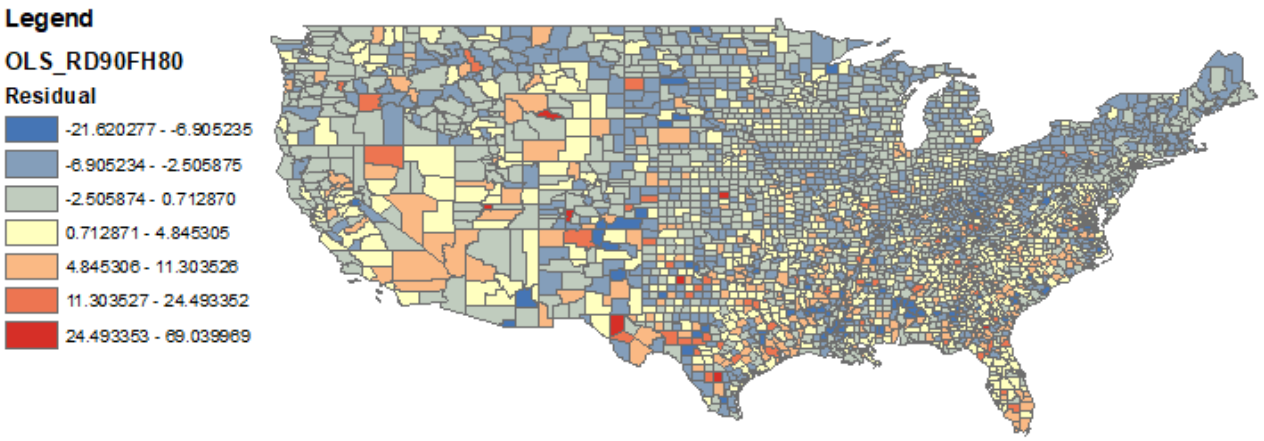
OLS

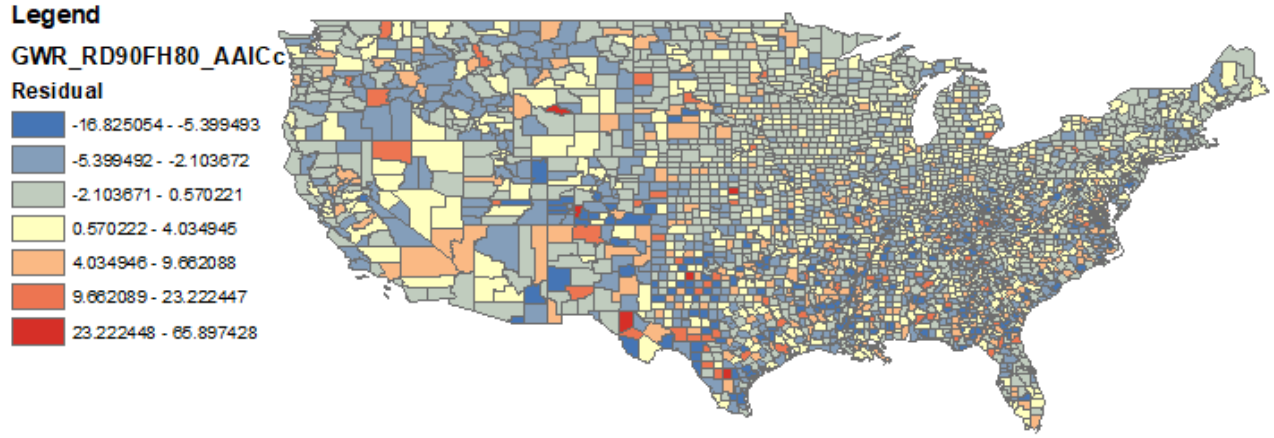


Question 5 (5 points): Create two residual maps (with proper color scheme and legend): one for regular regression and the other for GWR. For each map, describe the spatial pattern of the residual. Does the residual seem to be spatially random? Or there are spatial clusters of high/low residuals (through visual examination or spatial autocorrelation metrics)? Compare the two residual maps and explain the possible reasons that might have resulted in the observed spatial pattern(s) and the differences between them.

Answer:

As we saw from the Histograms visually, we see a slight negative residual expression. Spatially we see that there is a negative residual in the Northern and Midwestern States and a more positive residual in the South and South West. Additionally, it seemed based on our Regressions the South and South West is more random in it's over and under predictions where as in the North and Midwest there is a consistent under prediction. In the GWR there seems to be an overall more accurate prediction. This is also evident in the histogram being more even then the OLS histogram having a heavier, left sided graph.

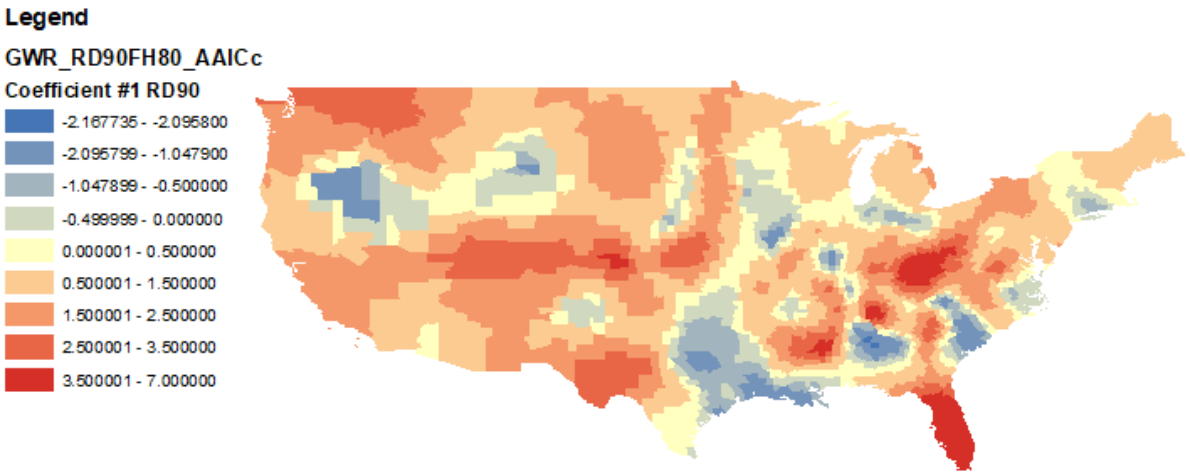




Question 6 (5 points): For GWR, generate maps for the regression coefficients (one map for each variable). Each map shows the spatially varying relationship between the dependent variable and an independent variable. What is the general spatial pattern on each map? How do you interpret the observed spatial pattern?

Answer:

The pattern for FH80 which is the decade prior the recorded Homicide Rate shows varying swings and definite either hotspots and cold spots on where it could be used as a predictor to zoom into areas and preform more analysis. Certain areas have a strong negative and others strong positive with definite middle gaps between the two where there is not a strong relation. In this map Florida, the Rust Belt, the Sun Belt, and the North West appear to have a positive relation while Most of the Eastern Seaboard, West of the Mississippi, and other pockets are closer to zero or a negative relation.



This seems to be a little more generalized in that the range is smaller and the negative relation has a much smaller representation. It shows areas with strong positive correlation but doesn't show a very strong negative relation. From this map you can observe a more positive relation between Female Head of House in the prior decade in the Mid-Atlantic and the Midwest with less of or slightly negative relation everywhere else.

In some places it seems to almost be an inverted map from the RD90.