Advanced Statistics
Assignment 2
Claude M. Schrader

Research Question:

In our OLS regression model, we explored the effect that college graduation rate, Population density, and broadband availability had on the broadband adoption rate across counties in Kentucky.

In this analysis, we perform a Geographically Weighted Regression to visualize the spatial distribution of our data, and help us understand how strongly the global model represents what's happening at the county level.

Measures:

adopt05 is the percentage of Kentucky households subscribing to broadband from a 2005 survey. This is the dependant variable.
college tracks the percent of college graduates in the population in 2004.
hhpct_0401 is the percentage of Kentucky households with broadband available.
popden2000 is the Population Density as of the 2000 census.

	adopt05	college	hhpct_0401	popden2000
min	0.071000000	0.041000000	0.00400000	16.100000
max	0.548000000	0.472000000	1.00000000	1336.700000
mean	0.238783333	0.208175000	0.55985000	83.035833
var	0.010237717	0.006586062	0.08080245	23495.153075
std.dev	0.101181606	0.081154554	0.28425771	153.281287

Methods:

First, we will create a few plots which assist in assessing the spatial distribution in the linear model.

We will create a weighted neighbor plot, which will show us the level of connectivity between counties, which is especially important around the border, when edge effects could come into play.

Using this neighbor information, we will create a lagged means choropleth map, which shows us the level of the dependent variable (broadband adoption) in a county's neighbors. This should give us a statewide visualization of any possible clustering that might be happening.

Next, we will use this neighbor connectivity table to calculate Moran's Index of autocorrelation using two methods, one approximating it using the normal distribution, and the other is calculating the index using 10,000 trials in a Monte Carlo simulation. We will also create

a scatterplot of the index. This will provide insight into the extent to which nearby counties are correlated, related to their distance.

We will then run a Geographically Weighted Regression, again regressing adopt05 on college, popden2000, and hhpct_0401. For the first step, we start with a bandwidth calculation for these counties. This calculation will decide how many counties are involved in the GWR' definition of "nearby counties". The second step of this regression is running the actual GWR function, assigning to it the bandwidth calculated in the proceeding step. This function will perform a regression for each county, generating individual parameter estimates and standard errors for each one, and the local R^2. We will also manually calculate the t and p values for each county, using arithmetic and the pt() function.

Geographically Weighted Regression provides a lot of data, so to make sense of it all, we will create choropleth maps.

Results:

Linear Model:

	Estimate	Standardized	Std. Error	t value	Pr(>|t|)
(Intercept)	0.08623234	0.00000000	0.02123268	4.061	0.0000890
college	0.42934669	0.34436534	0.10008186	4.290	0.0000372
hhpct_0401	0.08673592	0.24367427	0.02800457	3.097	0.002450
popden2000	0.00017598	0.26659431	0.00005173	3.402	0.000918

F-statistic: 31.12 on 3 and 116 DF, p-value: 0.000000000000007787

Figure 1: Is a basic histogram displaying the dependent variable, Broadband Adoption

Figure 2: Is a choropleth map visualizing the varying levels of broadband adoption across Kentucky.

Figure 3: Is a choropleth map of the residuals from our linear model, which could indicate to us whether or not our OLS model has left any important independent variables out. Generally speaking, we have a decent number of counties in the lower residual classes, and there isn't much strong clustering, which might indicate that our linear model is accounting for a lot of the variance.

Figure 4: is our weighted neighbor plot, visualizing the connectivity between a county and its neighbors, based on a Queen's Matrix. Because of the geography of the western edge of the state, there are a handful of counties with very few in-state neighbors, and one county with only one. This could be a source of weakness in our model, as out-of-state neighbors wouldn't be accounted for.

Figure 5: This is a lagged means plot of Broadband adoption. The darker counties have neighbors with relatively high adoption, so it gives us a picture into clusters where many states have similar adoption figures.

Figure 6: This is Moran's plot, providing an important visualization of Moran's index of Autocorrelation. The fit line here has a fairly steep slope, which would indicate a relatively strong effect. Since it has a strong positive slope, that tells us that counties with high adoption tend to be close to other counties with high adoption, and counties with low adoption also tend to be clustered.

Based on a range from -0.773 to 1.018, our I value is 0.25478, with a p-value of less than .01, which is statistically significant at the 99% level.

Geographically Weighted Model:

```
gwr(formula = adopt05 ~ college + hhpct_0401 + popden2000, data = kentucky,
bandwidth = kentucky.bw, gweight = gwr.Gauss, hatmatrix = T)
Fixed bandwidth: 232541.9
Coefficient Estimate at Median:
X.Intercept       0.079871972
college           0.408763825
hhpct_0401        0.093995408
popden2000        0.000179620

Residual sum of squares: 0.4353187
Quasi-global R2: 0.6426801
```

This GWR function does not provide t-value and p-values, so we must calculate them ourselves for all 120 counties for each of the three independent variables. To get the t value, we divide the parameter estimate by the standard error. Once we have that figure, we use the pt() function to look up the p value. I added both of these sets of data to the same data frame storing our GWR output.

Figure 7 – 12: Figures seven through 12 provide two choropleth maps visualizing selected output from the GWR analysis. The first map for each variable visualizes the parameter estimate on a county basis across Kentucky. These three maps show that out three independent variables do have strong spatial clustering, but generally the areas with the strongest effect on the dependent variable don't overlap too much, which would indicate that the global model provides a decent fit when looking at the local level.

The second map for each independent variable visualizes the p-value, to display the statistical significance of this regression at a county level. For the most part, the counties with a strongest effect on Broadband Adoption also are statistically significant at least at the 95% level. The one exception is the handful of counties at the extreme western edge of the state. This part of the state does not have much statistical significance from any of the independent variables. Because this part of the state is very long and narrow, these counties do not have many in-state neighbors, and in the case of one county, it is only directly connected to one other county in Figure 4.

Figure 13: this is a choropleth map representing the local R^2 values, which is a measure of correlation among all variables. There is a bit of a split to this graph – the eastern half of the state generally has much stronger values than the west. Given that our model is looking at

college graduation rates and population density, if the western side of Kentucky is more rural, that could explain this regional relative weakness in our model.

Conclusion:

Generally speaking, this model provides a pretty strong fit based on this data. The parameter estimates of our three variables in total provide a decent statistically significant coverage of the state. This would indicate that although different variables have stronger effects on different parts of the state, as long as they're all accounted for, the global model can provide a pretty good fit statewide.

For further analysis, the western counties could use more research. Rerunning this analysis including nearby counties in neighboring states might provide a stronger analysis of this region. Further research into the demographics and geography of this area could also indicate whether there are missing independent variables.

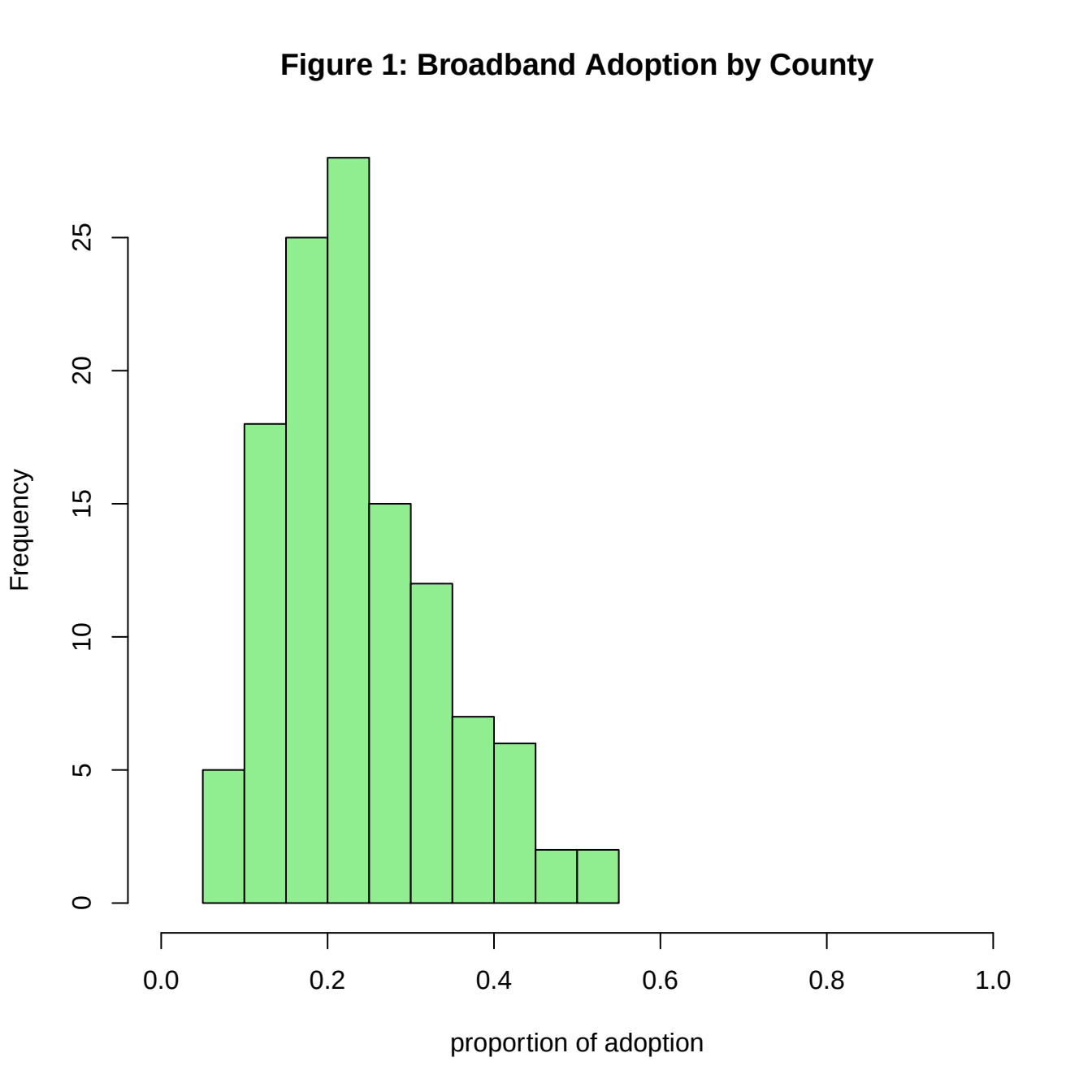
Figure 1: Broadband Adoption by County
Frequency
0
5
10
15
20
25
0.0
0.2
0.4
0.6
0.8
1.0
proportion of adoption

Figure 2: Kentucky Broadband Adoption

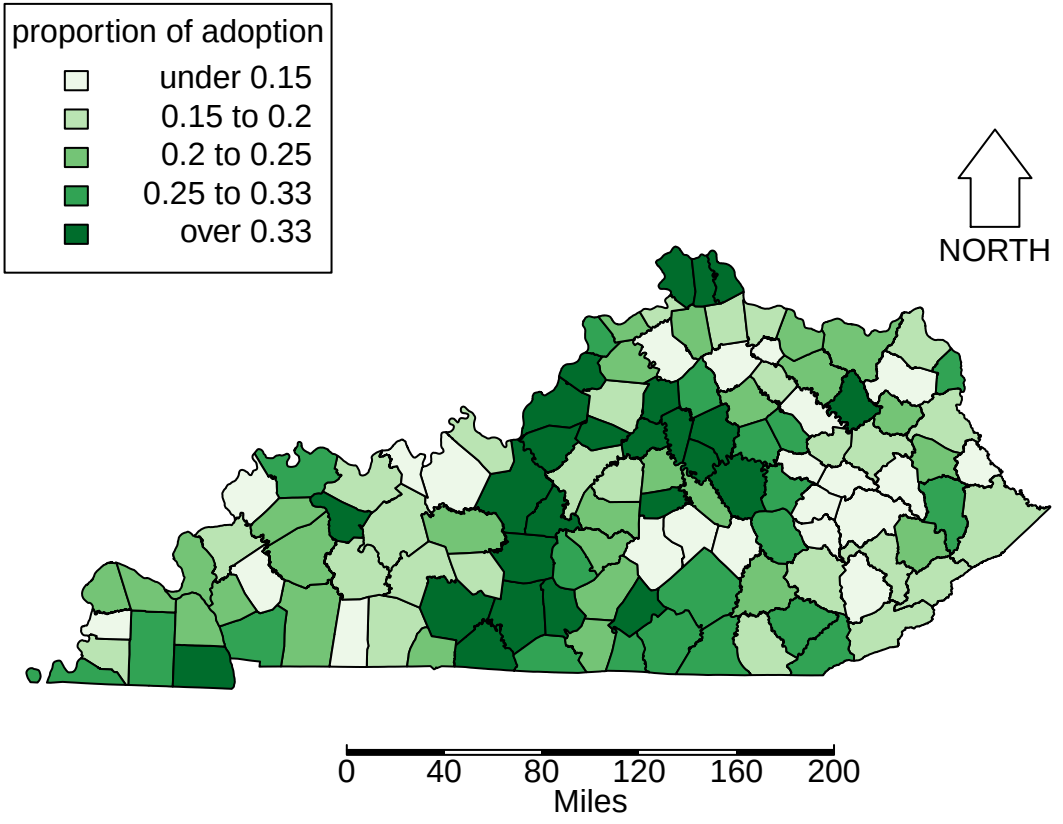

Figure 3: Residuals from linear model lm01

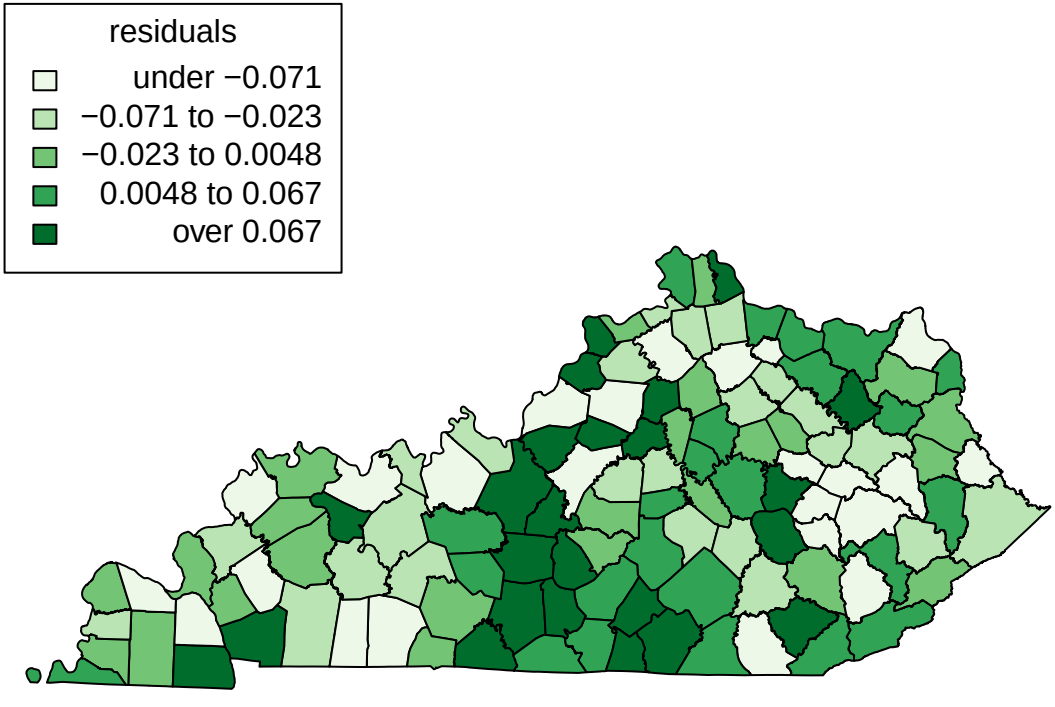

Figure 4: Weighted Neighbor plot of Kentucky Broadband Adoption

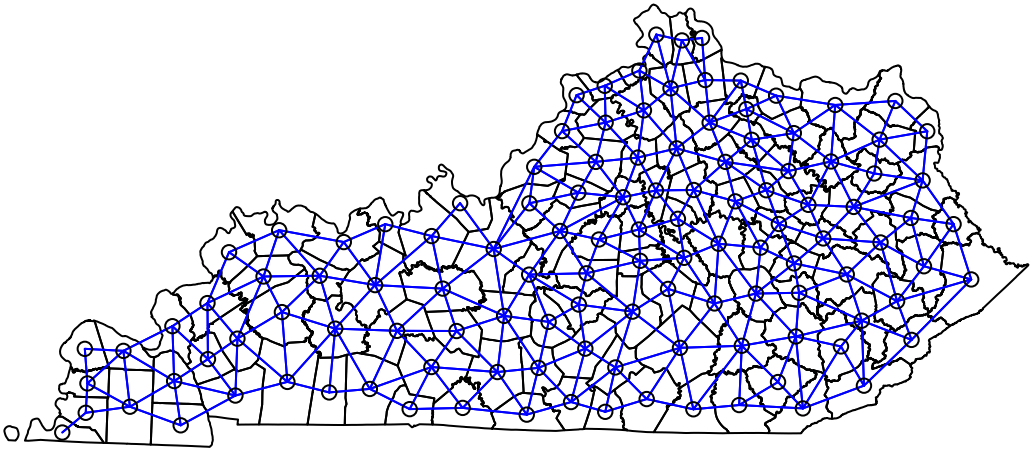

Figure 5: Lagged Means plot of Kentucky Broadband Adoption

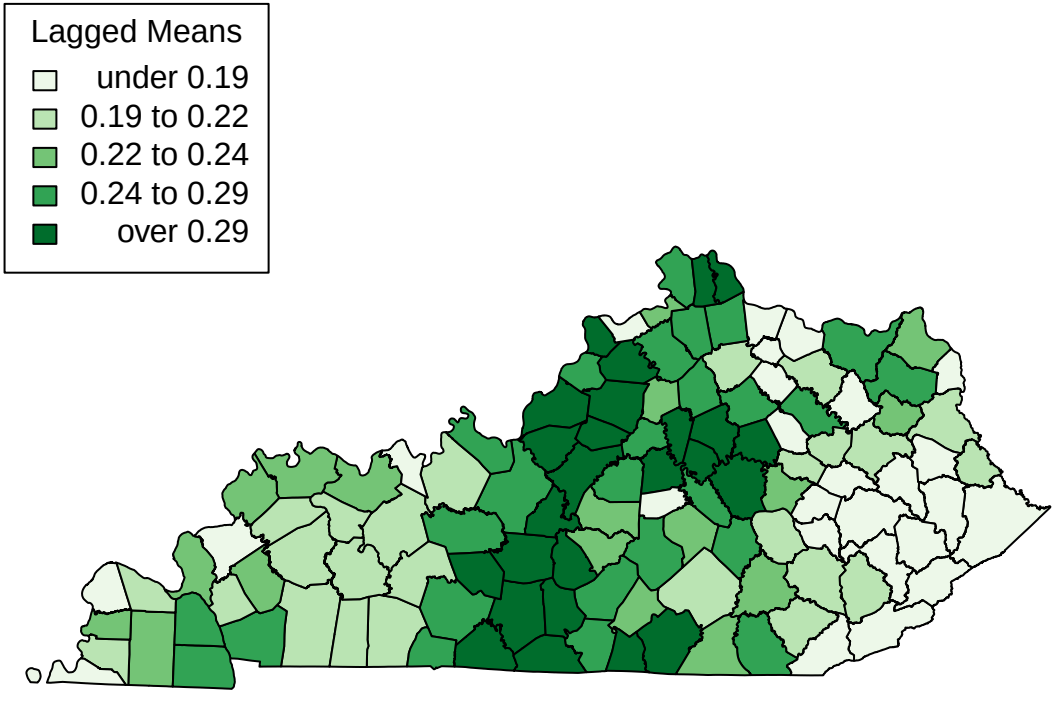

Figure 6: Moran's Plot

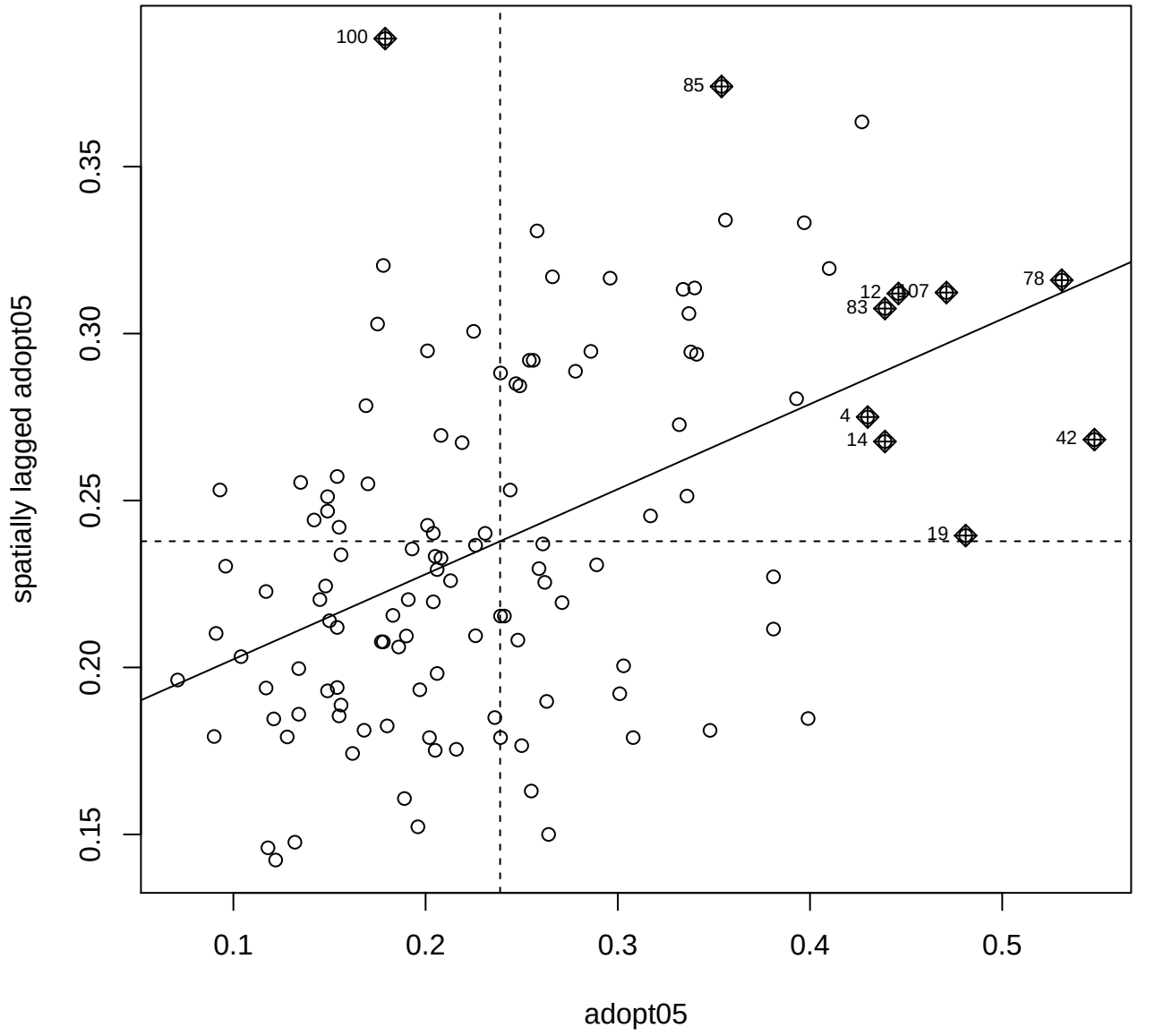

Figure 7: GWR Parameter Estimate for College Graduation Rate

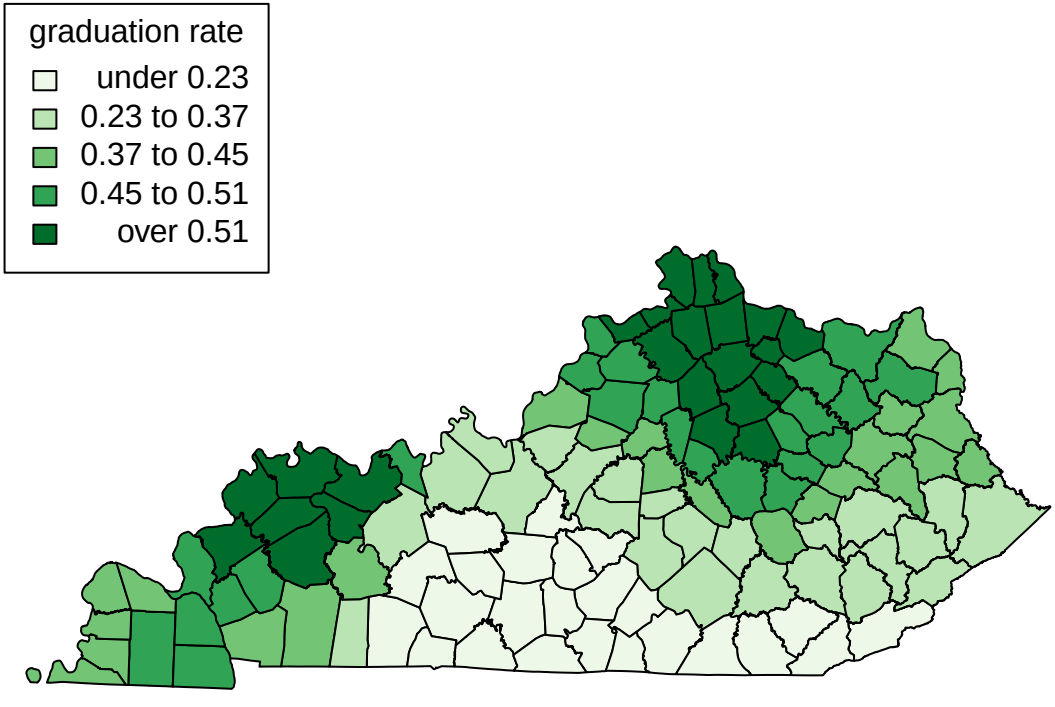

Figure 8: P−Value for College Graduation Rate

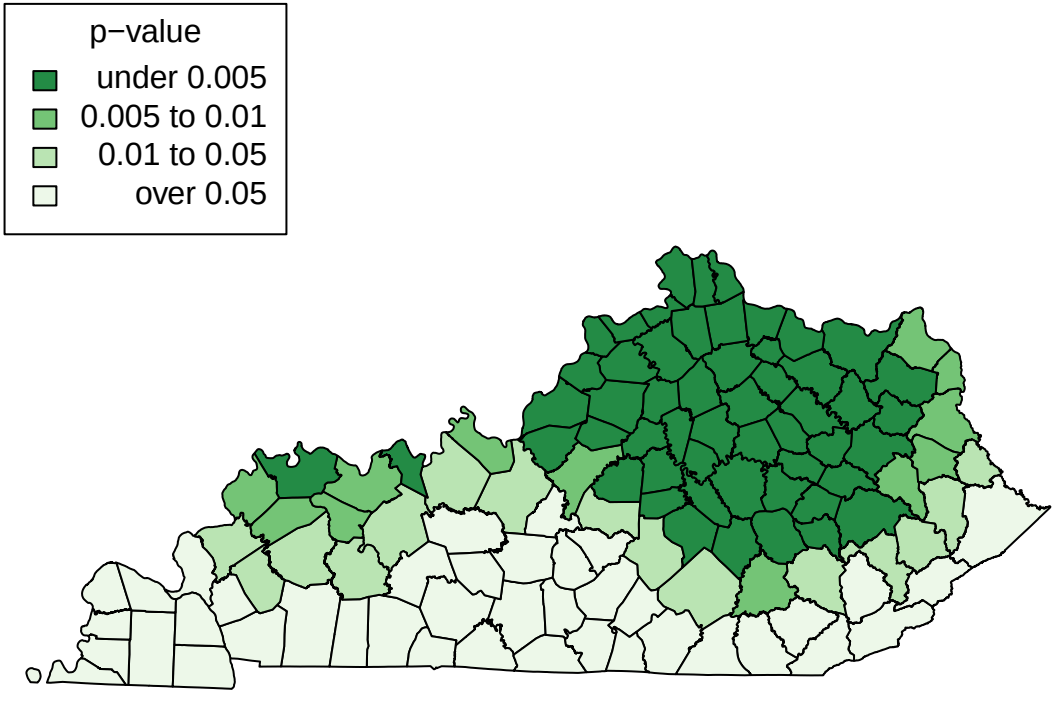

Figure 9: GWR Parameter Estimate for Population Density in 2000

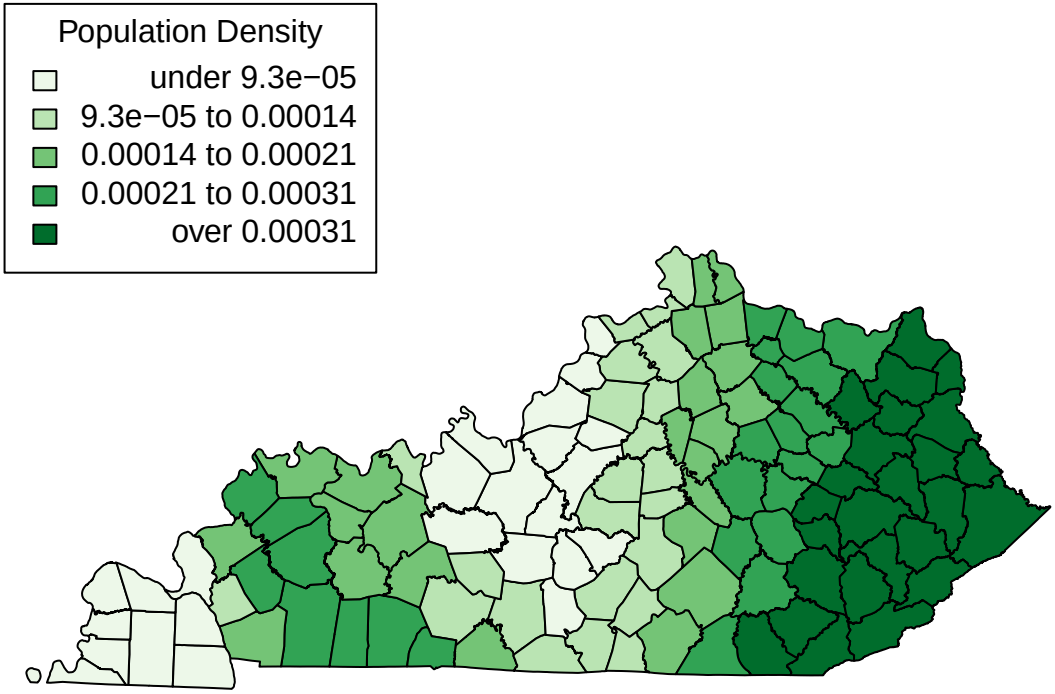

Figure 10: P−Value for Population Density in 2000

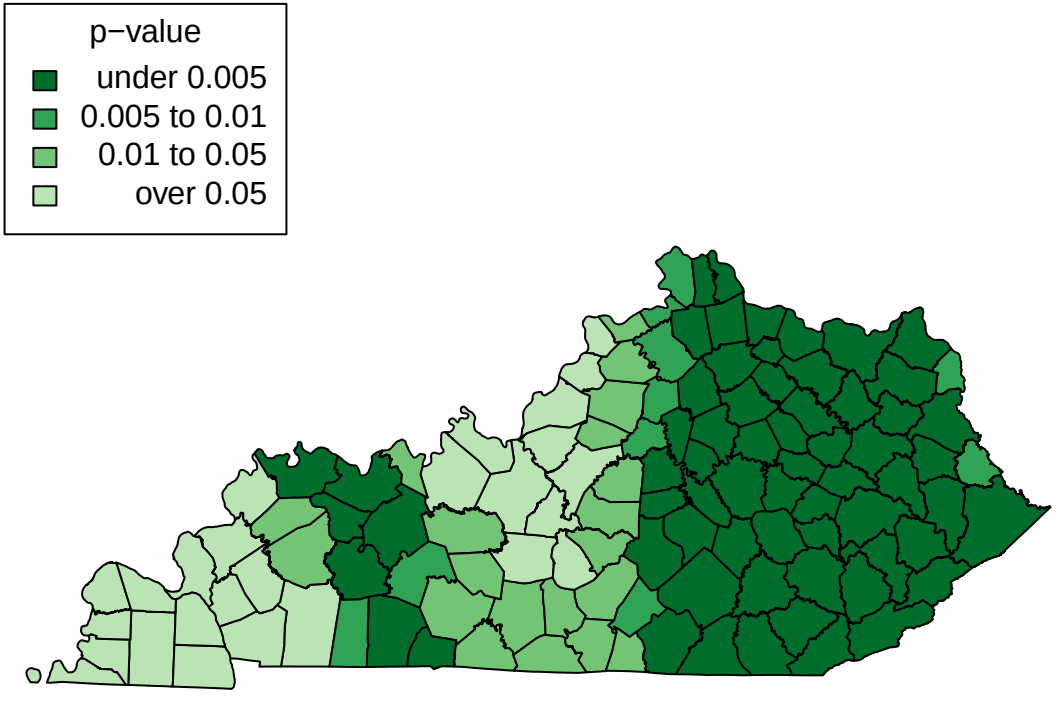

Figure 11: GWR Parameter Estimate for Broadband Availability Rate

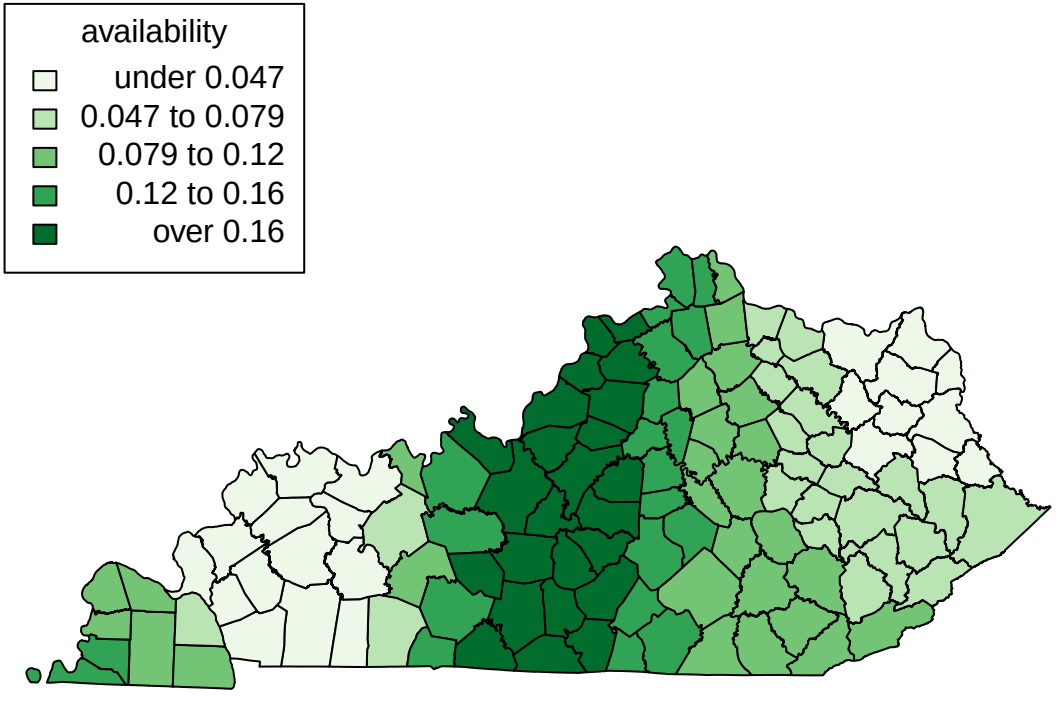

Figure 12: P−Value for Broadband Availability Rate

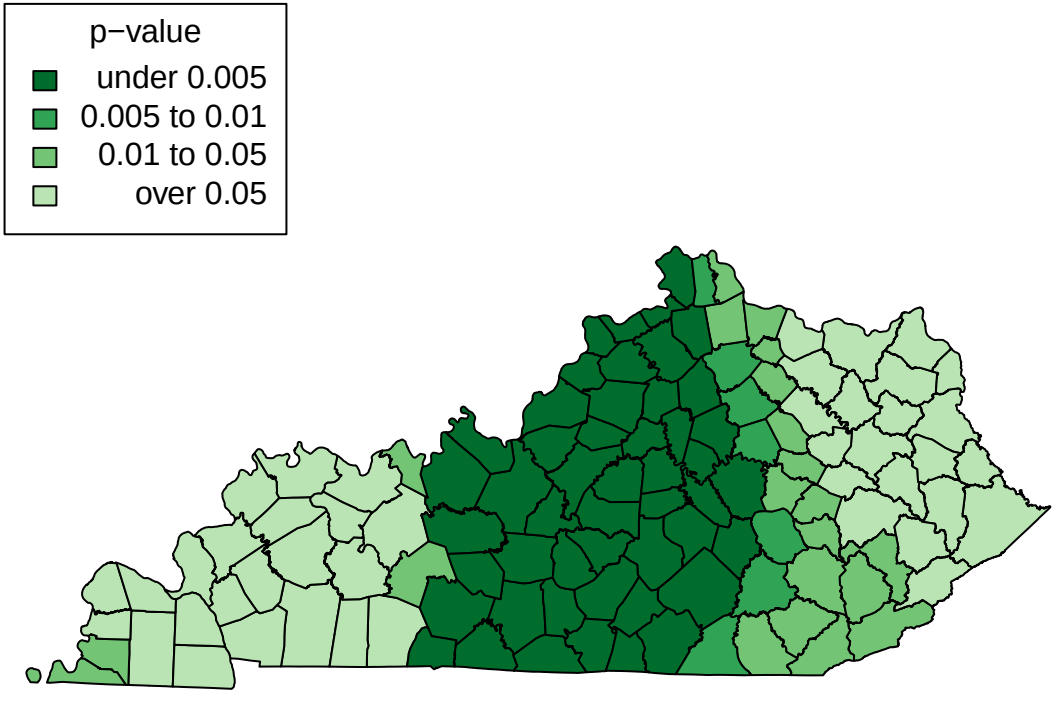

Figure 13: GWR Local R^2 values