Code

Homework 2

Akira Di Sandro, Sofia Fasullo, Amy Solano

2024-11-02

Introduction

This report iterates on our earlier work that sought to understand whether certain demographic and housing factors could be used to reasonably predict median house values using Philadelphia census block group data. Previously, we carried out OLS regression to examine the relationship between the dependent variable, median house value (log transformed); and the predictors percentage of detached single-family homes, percentage of residents who hold a bachelor's degree or above, percentage of vacant housing units, and number of households living in poverty (log transformed). This model indicated that each predictor had a statistically significant relationship with the dependent variable and an R-squared value of 0.66, meaning that our model captures approximately 66% of the variance in median house value.

In addition to our OLS regression, we created choropleth maps of our variables and model residuals which indicated the presence of spatial autocorrelation. OLS analysis is often inappropriate when dealing with spatial data, in part due to the assumption that observations in the data are independent of one another. This assumption is violated in cases like ours where observations are spatially autocorrelated with similar values clustering together in space. In this report, we will use spatial lag, spatial error, and geographically weighted regression to improve upon our OLS model through the use of spatial methods.

Show

Methods

A Description on the Concept of Spatial Autocorrelation

The 1st Law of Geography is that "everything is related to everything else, but near things are more related than distant things" (Waldo Tobler, 1970). When applied to the field of statistics, this means that data variables have relationships not only to each other but also to space, and that spatial variables can be used to help predict non-spatial variables. In essence, this describes the field of spatial statistics.

The Moran's I statistic is a measurement of the spatial clustering of data. Moran's I values range between 1 (perfect spatial clustering), and -1 (perfect spatial repelling), with 0 showing no sign of spatial correlation. The formula for Moran's I is as such:

$$\text{Moran's I} = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Where

- N: The total number of spatial units (observations).
- x_i: The value of the variable at location i.
- x_j: The value of the variable at location j.
- $\bar{x}$: The mean value of the variable across all locations.
- w_{ij}: The spatial weight between locations i and j (typically set to 1 if i and j are neighbors, and 0 otherwise).

This formula is an expansion of the correlation coefficient statistic to accommodate relations between variables that are near each other as accounted for by the weights.

The spatial weight between locations can take many forms depending how the relationship of proximity is defined. Some spatial weight formulas use chess analogies, such as rook and queen. These refer to how spatially adjacent features would be weighted by where a chess piece such as a rook or queen would move to the adjacent tiles on a chess board.

In our analysis we will be using a queen weights matrix, which considers the entire square of adjacent grid tiles- including side and corner tiles - as the "near" variables to be considered when calculating the Moran's I. The queen matrix will be used exclusively and consistently throughout this report. However, many statisticians choose to use multiple different weights matrices throughout their study, because it shows that their results are not dependent on one particular matrix. We will be using only the queen matrix because it is a broad, catch-all consideration of spatial relationships, with all spatially contiguous data weighed in.

Calculating the Moran's I value of our data produces one number we can observe to be close to 1, -1 or 0. To add more context to this number, we can run a significance test to understand whether this Moran's I value is especially far from 0 in comparison to random permutations of the data itself. To do this, we set our null hypothesis to Moran's I = 0 (no spatial pattern) and our alternate hypothesis to Moran's I != 0 (some spatial pattern). To simulate a distribution under the alternate hypothesis, we run 999 permutations of random spatial shuffling of the data and create 999 Moran's I values from this and display their distribution. According to the Central Limit Theorem, this distribution mathematically computes to a normal distribution, and a z-test can be used, so long as there are over 30 observations in the data. We compared our original Moran's I value and calculated its p-value, or probability that it falls within this distribution. If the p-value is very low (below 5%), we can reject the null hypothesis that there is no spatial pattern within the data.

While Moran's I calculates whether there exists any spatial pattern of data across an entire dataset, the Local Indices of Spatial Autocorrelation (LISA) statistic measures whether there is any immediate clustering of data values for every individual observation in space. LISA is also known as Local Moran's I, and the Moran's I test described above is often called Global Moran's I.

The significance test for LISA is very similar to that of Global Moran's I, with the null hypothesis being that the LISA value for each observation is 0 (no local clustering). This hypothesis is tested by simulating a random spatial distribution with a mathematical normality by randomly shuffling the data 999 times, but keeping the observation in question constantly pinned to its original position. From there, the calculation of the p-value of the observed LISA value is the same as Global Moran's I.

A Review of OLS Regressions and Assumptions

Ordinary Least Squares (OLS) regression is a statistical method used to estimate the relationship between a dependent variable and one or more independent variables by minimizing the sum of the squares of the residuals. By regressing the dependent variable on another variable or set of variables, it uses the relationship between these variables to estimate the dependent variable for each observed data point. The key assumptions of OLS include linearity, independence of errors, homoscedasticity (constant variance of errors), normality of errors, and no multicollinearity among independent variables. For a more detailed discussion of OLS and its assumptions, please refer to Homework 1.

When data has a spatial component, the assumption that errors are random and independent often does not hold - they may be spatially correlated. This can be tested by examining the spatial autocorrelation of the residuals using Moran's I statistic. Another way to assess OLS residuals for spatial autocorrelation is to regress them on nearby residuals, specifically the residuals from neighboring block groups as defined by the Queen contiguity matrix. In this context, the slope (b) at the bottom of the scatterplot of OLS residuals (OLS_RESIDU) against weighted residuals (WT_RESIDU) indicates the degree of spatial dependence in the residuals and is calculated as the change in OLS residuals per unit change in nearby weighted residuals.

GeoDa or R, the tools used for OLS regression, also provide methods to test other regression assumptions.

1. **Homoscedasticity:** This assumption is tied to the independence of errors. In GeoDa/R, tests such as the Breusch-Pagan test and the White test are commonly used to examine data for heteroscedasticity. The null hypothesis (H_0) for these tests states that there is homoscedasticity (constant variance of errors), while the alternative hypothesis (H_1) states that heteroscedasticity is present (non-constant variance of errors).
2. **Normality of Errors:** The assumption of normality of errors can be tested using the Shapiro-Wilk test in GeoDa/R. The null hypothesis (H_0) for this test states that the errors are normally distributed, while the alternative hypothesis (H_1) posits that the errors are not normally distributed.

Spatial Lag and Spatial Error Regression

Both GeoDa and R will be utilized for running spatial lag and spatial error regressions in this analysis.

Spatial lag regression incorporates space into the model by using the dependent variable's values nearby (or spatially lagged) to the observed point as a predictor variable in the model. The model equation can be expressed as:

$$y = \beta_0 + \rho W y + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

Where

- y is the dependent variable.
- $beta_0$ is the intercept.
- ρ is the spatial autoregressive coefficient, representing the degree of spatial dependence.
- $W y$ is the spatially lagged dependent variable, which captures the influence of neighboring values of y.
- $beta_1, \ldots, beta_n$ are the coefficients for the independent variables.
- $epsilon$ is the error term, capturing unobserved factors.

Spatial error regression addresses spatial autocorrelation in the error term rather than the dependent variable itself. It uses a spatial lag of the error term as a predictor in the model. The model equation for the spatial error model is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \lambda W \epsilon + u$$

Where

- y is the dependent variable.
- $beta_0$ is the intercept.
- $\lambda W \epsilon$ is the spatially lagged error, which captures the influence of neighboring values of ϵ.
- $beta_1, \ldots, beta_n$ are the coefficients for the independent variables.
- u is the random noise term, capturing unobserved factors.

The assumptions required for OLS regression remain applicable to both spatial lag and spatial error regression models, with the exception of the assumption regarding the spatial independence of observations.

The goal of employing spatial lag and spatial error regression is to achieve reduced spatial autocorrelation in the regression residuals, enhancing the reliability of the model estimates.

The results of the spatial lag regression will be compared to the OLS regression results, as will the spatial error regression results. The evaluation of model performance will be based on several criteria:

1. **Akaike Information Criterion (AIC)/Schwarz Criterion (BIC):** These criteria measure the relative quality of statistical models for a given dataset. A lower AIC or BIC indicates a better-fitting model. The null hypothesis states that there is no difference in this criteria between the OLS and spatial models, while the alternative hypothesis suggests that the spatial model fits better than the OLS model.
2. **Log Likelihood:** This criterion assesses how well the model explains the observed data, with higher values indicating better fit. The null hypothesis asserts that there is no difference in model fit.
3. **Likelihood Ratio Test:** This test compares the goodness-of-fit between two models, with the null hypothesis stating that the simpler model (e.g., OLS) is sufficient, while the alternative hypothesis posits that the more complex model (e.g., spatial model) provides a significantly better fit.

Additionally, another way to compare OLS results with spatial lag and spatial error results is by examining the Moran's I statistic of the regression residuals. A significant Moran's I indicates the presence of spatial autocorrelation in the residuals, suggesting that the spatial models may be more appropriate. If the spatial models exhibit reduced Moran's I values, this would indicate improved performance over OLS.

Geographically Weighted Regression (GWR)

Simpson's paradox says that what is true for the whole might not be true for subgroups, and vice versa. Disaggregating data can give us a better glance at what's going on in different geographic areas. GWR helps us do this by running local regressions and estimating beta parameters around each observation. We run a model to describe the relationship between the dependent variable and predictors around a specific location (every observation in the data set). When running a regression for location i, GWR assigns greater weights to observations that are closer to location i. (defining proximity to say which neighbors have a stronger weight is up to the researcher – various methods: include neighbors whose centroids are within one mile of location i; include 50 nearest neighbors). Weight of observation varies with location i. GWR requires a large number of observations (at least 300). Otherwise, you're basically running a global regression for each "local" regression.

We can conceptualize the bandwidth to be the “distance” from each location for which we assign greater weights to the observations that lie within this bandwidth. For fixed Bandwidth, the number of observations will vary around each observation, but the bandwidth distance (h) will stay the same. In adaptive bandwidth, the number of observations will remain constant, and bandwidth will change according to the degree of clustering. If there is significant clustering of observations around a location, the bandwidth will be smaller compared to a location where there are not as many neighboring observations.

Fixed bandwidth is more appropriate when the distribution of observations is relatively stable across space (polygons similar in size, points that are relatively uniformly distributed). Adaptive bandwidth is more appropriate when distribution of observations varies in space. In our case, we prefer adaptive bandwidth because the size and shapes of Philadelphia block groups vary greatly. This way, we can ensure that each local regression will use a similar amount of neighboring observations.

Even if two variables might not show any multicollinearity globally, when working with spatial data, there may be areas in which the variables are highly collinear. This phenomenon is called Local multicollinearity. This is why we need to pay attention to the condition number and not include local regressions which have a condition number > 30, because this means that there are variables that exhibit local multicollinearity around this region. In general, we need to be sure that all of our variables show spatial variability.

Since we run a regression for every observation, we would be conducting thousands of tests (4 predictors = 4 coefficients + 1 intercept per local regression for 1720 observations = 8600 tests) to see whether each parameter is statistically significantly different from zero. We would expect 43 of these tests to return as significant just by chance (type I error).

Results

Spatial Autocorrelation

Show

```
## [1] "The Global Moran's I value for the log of median house value is about  0.79
, with a p-value of about 0 (it is extremely small)."
```

Show

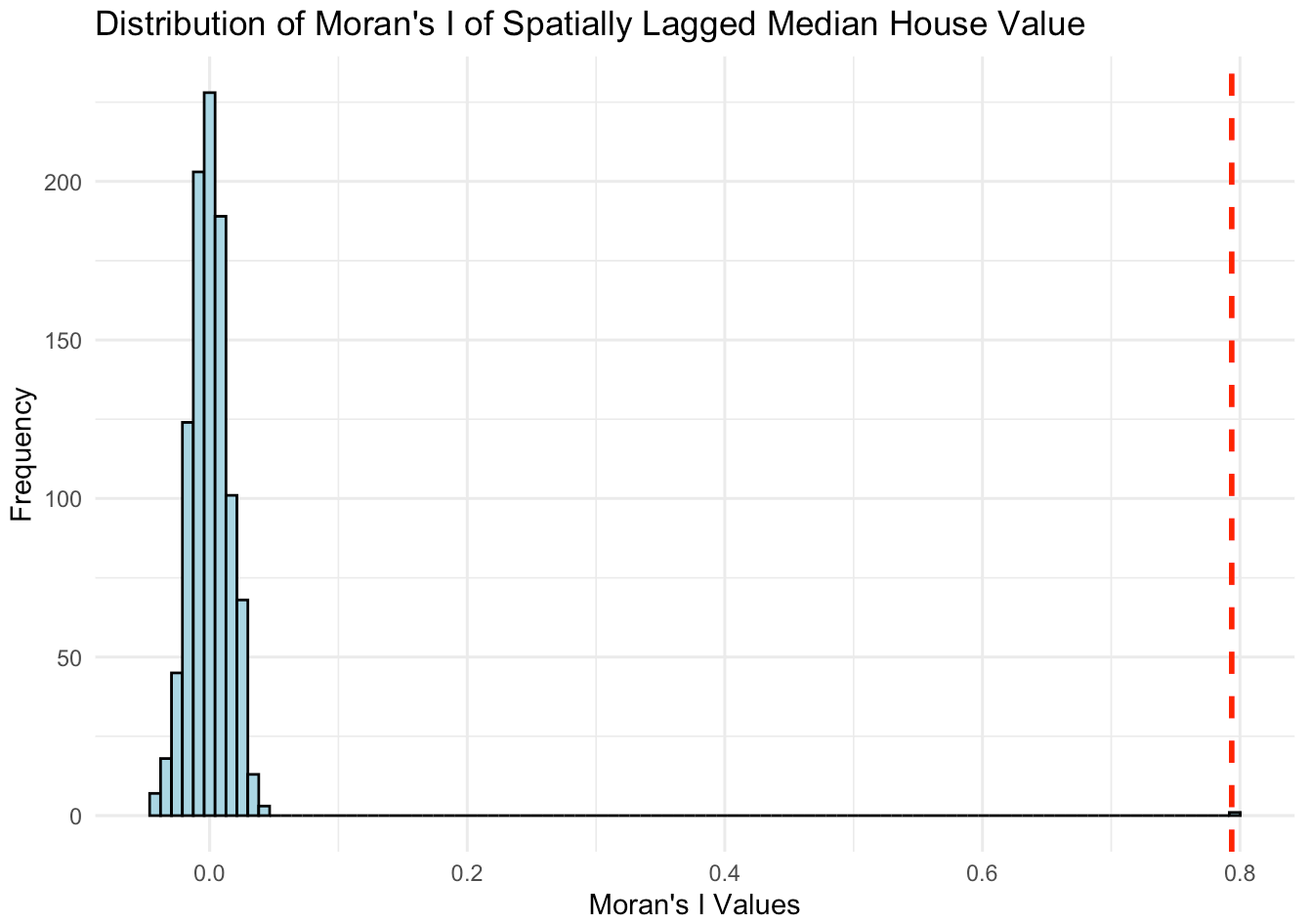

Figure 1.

Show

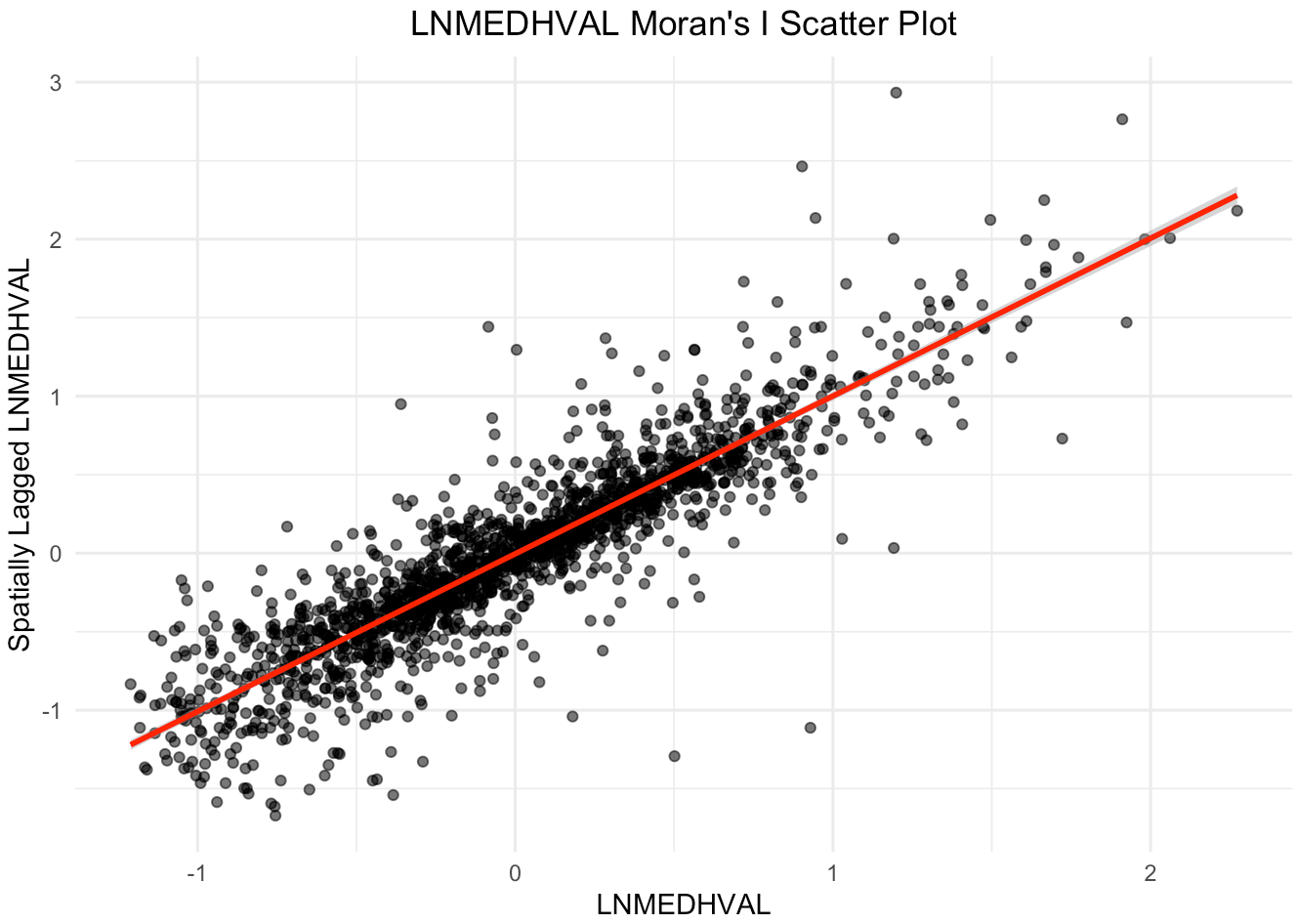

Figure 2.

We observe a global Moran's I value of 0.794 for our dependent variable, median house value (log-transformed) LNMEDHVAL which is statistically significant (pseudo $p < .001$). From Figure 1. it can be seen that the Moran's I value falls far outside the distribution. This means that we see statistically significant positive spatial autocorrelation of our dependent variable, meaning similar values of median house value are clustered across Philadelphia.

Show

Show

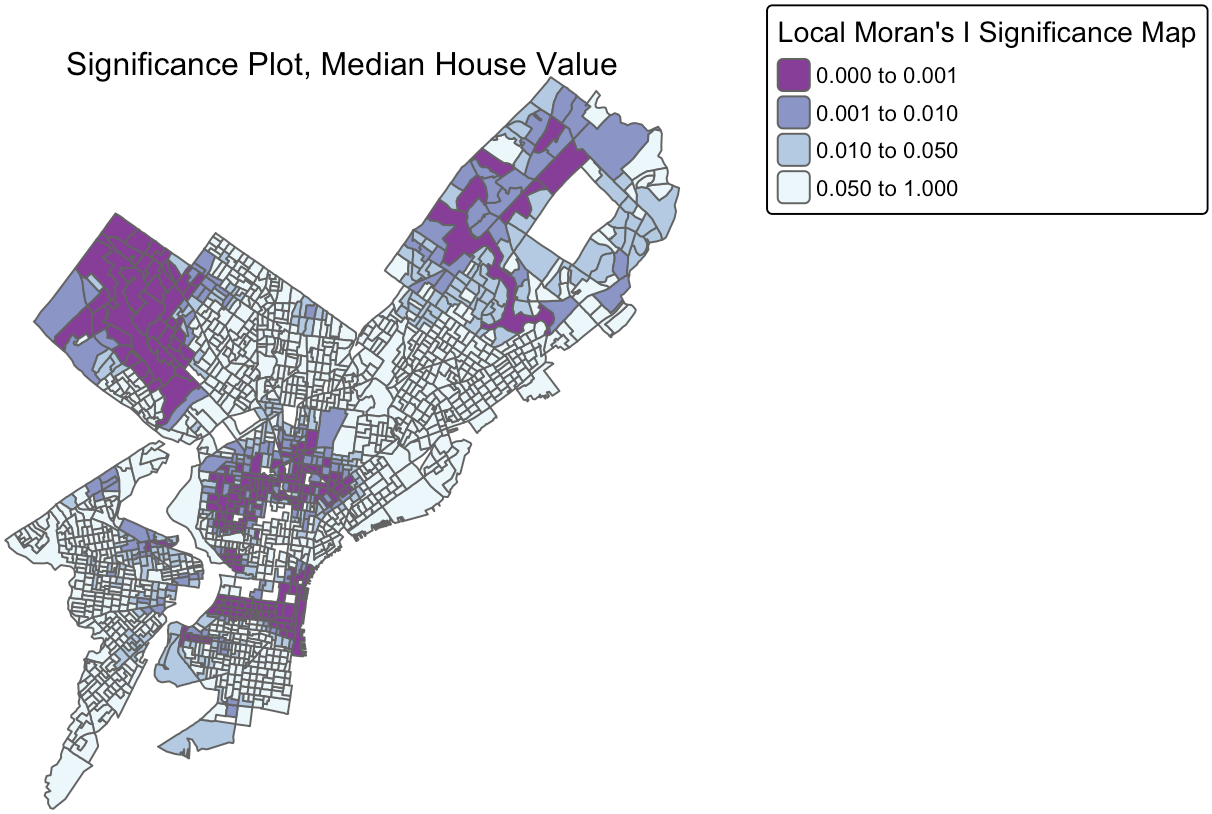

Figure 3.

Show

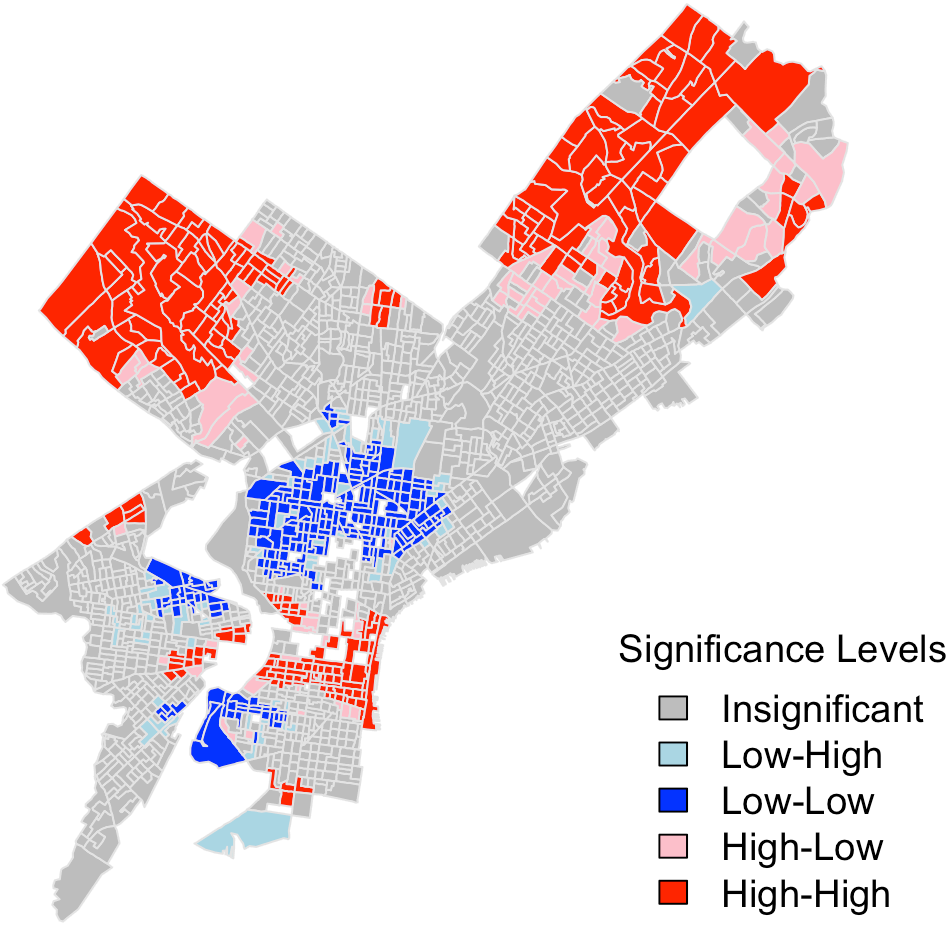

Figure 4.

According to the Cluster and Significance Maps seen in Figure 3. and Figure 4. the Northeast, Northwest, Center City (from the Schuylkill to the Delaware river) block groups of Philadelphia, and University City show local clustering of high median house values. Meanwhile, Southwest Philadelphia, the Grays Ferry neighborhood, North Philadelphia, and the area around Fairmount Park West show local clustering of low median house values. There are also a handful of block groups with low median house value that are surrounded by high median house value block groups and vice versa. The areas for which we see non-significant local Moran's I include Fairmount, Chinatown North, Kensington, Fishtown, Southeast Philadelphia, West Philadelphia, and Fern Rock. These neighborhoods are much more diverse in median house value.

A Review of OLS Regression and Assumptions: Results

Show

```
##
## Call:
## lm(formula = LNMEDHVAL ~ LNNBELPOV100 + PCTBACHMOR + PCTVACANT +
##     PCTSINGLES, data = regdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25817 -0.20391  0.03822  0.21743  2.24345
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  11.1137781  0.0465318 238.843 < 0.0000000000000002 ***
## LNNBELPOV100 -0.0789035  0.0084567  -9.330 < 0.0000000000000002 ***
## PCTBACHMOR    0.0209095  0.0005432  38.494 < 0.0000000000000002 ***
## PCTVACANT    -0.0191563  0.0009779 -19.590 < 0.0000000000000002 ***
## PCTSINGLES    0.0029770  0.0007032   4.234            0.0000242 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3665 on 1715 degrees of freedom
## Multiple R-squared:  0.6623, Adjusted R-squared:  0.6615
## F-statistic: 840.9 on 4 and 1715 DF,  p-value: < 0.00000000000000022
```

According to the summary of OLS results, the model is significant ($p < .001$ for the F-Ratio) with all four predictors being statistically significant. Roughly 66.2% of the variance in LNMEDHVAL is explained by this model.

Show

```
## 'log Lik.' -711.4933 (df=6)
```

Show

```
##
##  Breusch-Pagan test
##
## data:  ols
## BP = 113.19, df = 4, p-value < 0.00000000000000022
```

Show

```
##
##  studentized Breusch-Pagan test
##
## data:  ols
## BP = 42.868, df = 4, p-value = 0.00000001102
```

Show

```
## White's test results
##
## Null hypothesis: Homoskedasticity of the residuals
## Alternative hypothesis: Heteroskedasticity of the residuals
## Test Statistic: 43.94
## P-value: 0
```

Show

```
##
##  Jarque Bera Test
##
## data:  ols$residuals
## X-squared = 778.96, df = 2, p-value < 0.00000000000000022
```

Both of the tests for heteroscedasticity (Breusch-Pagan and Koenker-Bassett tests) are non-zero (113.2 and 42.9, respectively) with statistically significant p-values ($p < 0.001$). A note is that the GeoDa tests created different results for these tests, (162.9 and 61.7, respectively). This means that we observe heteroscedasticity, where the variance in our OLS residuals is not constant and depends on the predictors.

The plot of standardized residuals by predicted values from our first homework assignment, however, makes it seem like there is homoscedasticity with the exception of a few outliers. Though, we do see from this plot that there is less variance in the standardized residuals when the log-transformed median house value is between 11 to 11.5 compared to the rest of the log-transformed median house values, potentially warning us of heteroscedasticity.

We observe a value of 779.0 for the Jarque-Bera test with a p-value of < 0.001, meaning we have non-normality of errors – another clue that tells us that an OLS model may not be the best way to look at our data.

However, reviewing our histogram of residuals from homework 1, it seems like we do see a normal distribution of OLS residuals. There are, again, a few outliers at the both tails of the histogram that hint at possible non-normality of residuals.

Show

Show

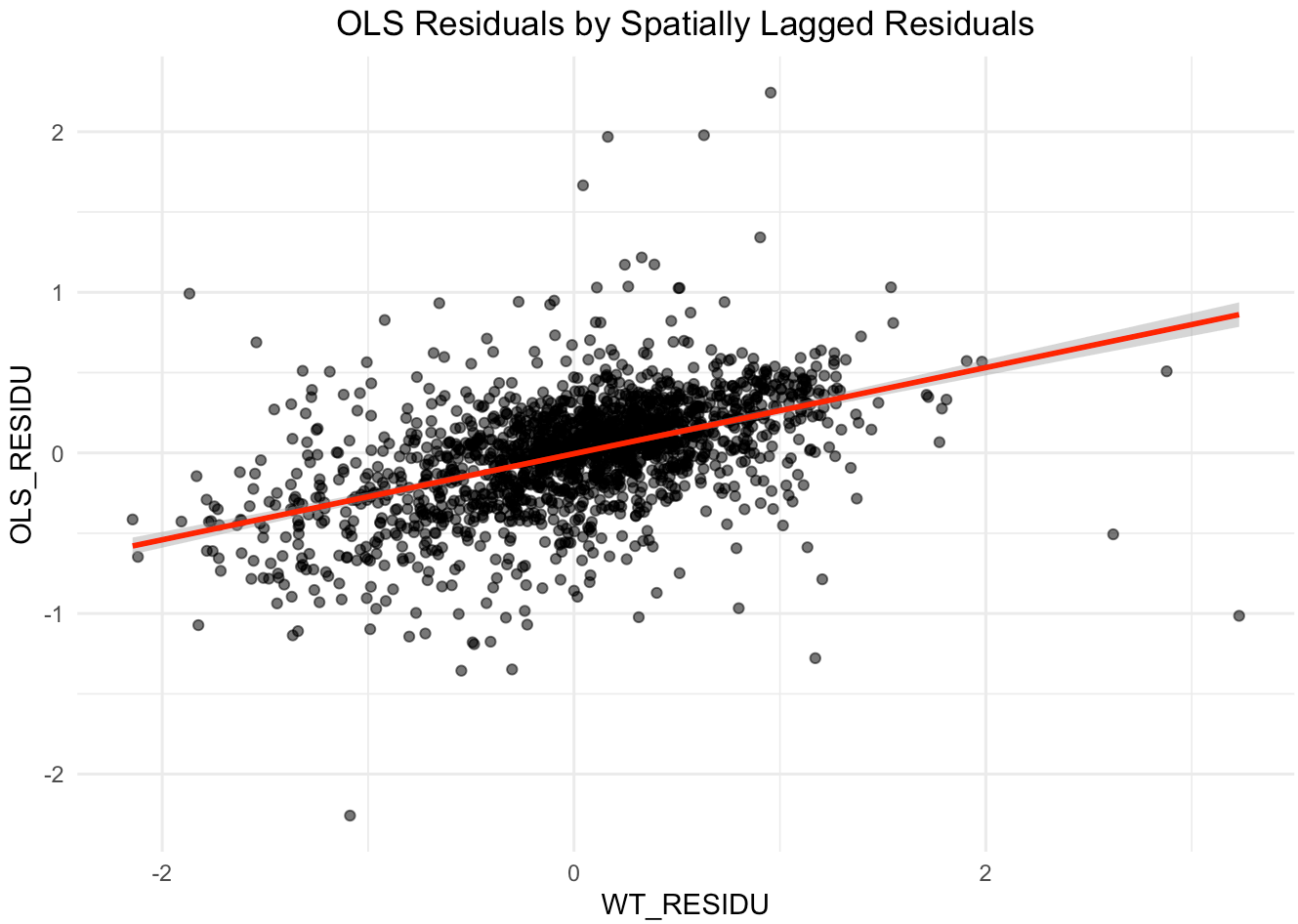

Figure 5.

In this scatterplot of the OLS_RESIDU by WT_RESIDU (Figure 5), we observe a clear positive correlation between the two variables.

Show

```
##
## Call:
## lm(formula = OLS_RESIDU ~ WT_RESIDU, data = regdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96184 -0.16301  0.02138  0.16914  1.99217
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -0.004725   0.007754  -0.609               0.542
## WT_RESIDU    0.268051   0.011860  22.601 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3215 on 1718 degrees of freedom
## Multiple R-squared:  0.2292, Adjusted R-squared:  0.2287
## F-statistic: 510.8 on 1 and 1718 DF,  p-value: < 0.00000000000000022
```

In regressing WT_RESIDU on OLS_RESIDU, we observe an R2 of 0.229. There is a clear positive correlation between the two variables. The slope b is 0.268 with $p < 0.001$, meaning that there is significant spatial autocorrelation that is slightly positive. This makes sense looking at Figure 5.

Show

```
## [1] "The Global Moran's I value for OLS residuals is 0.31, with a p-value of abo
ut 0."
```

Show

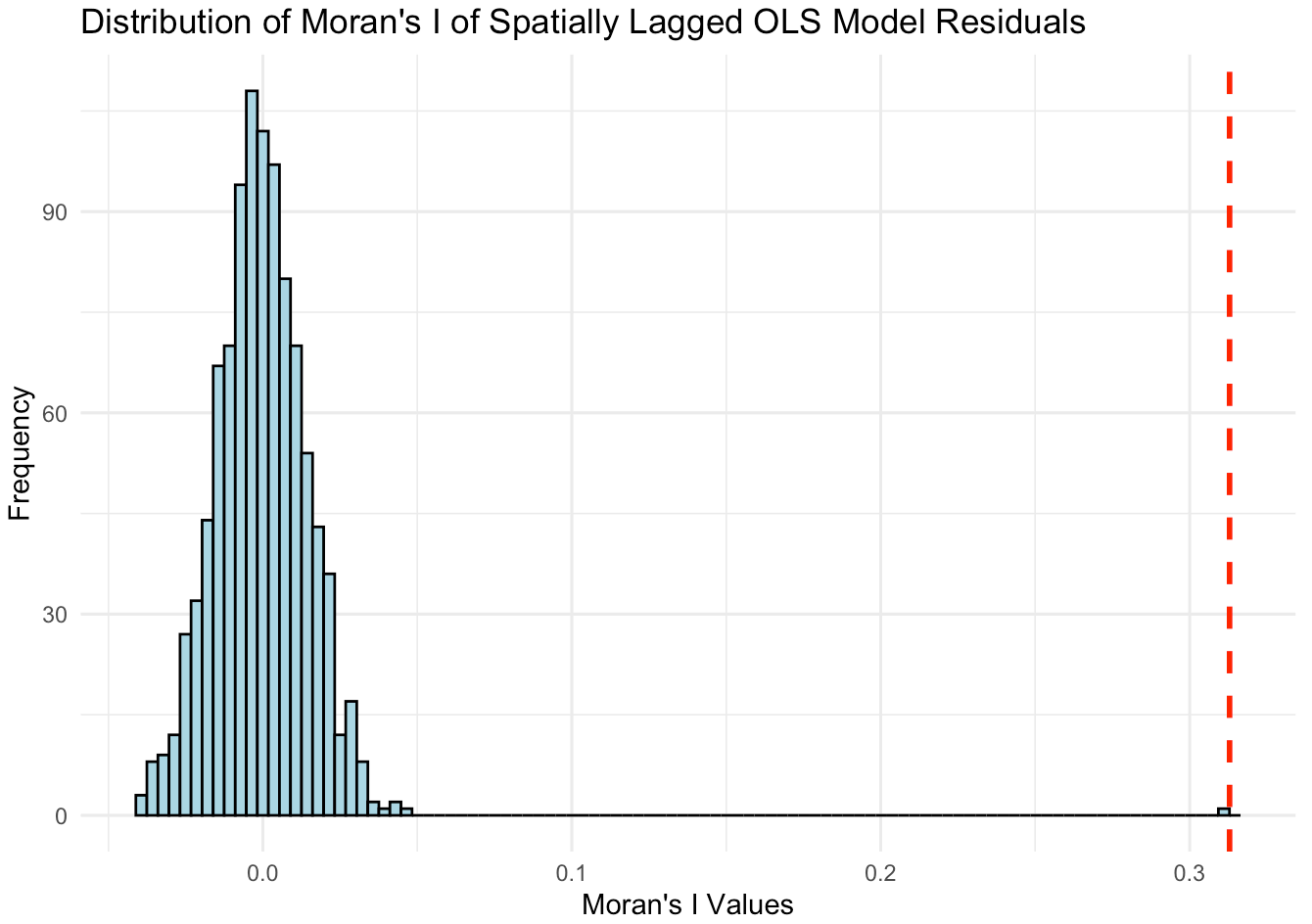

Figure 6.

Show

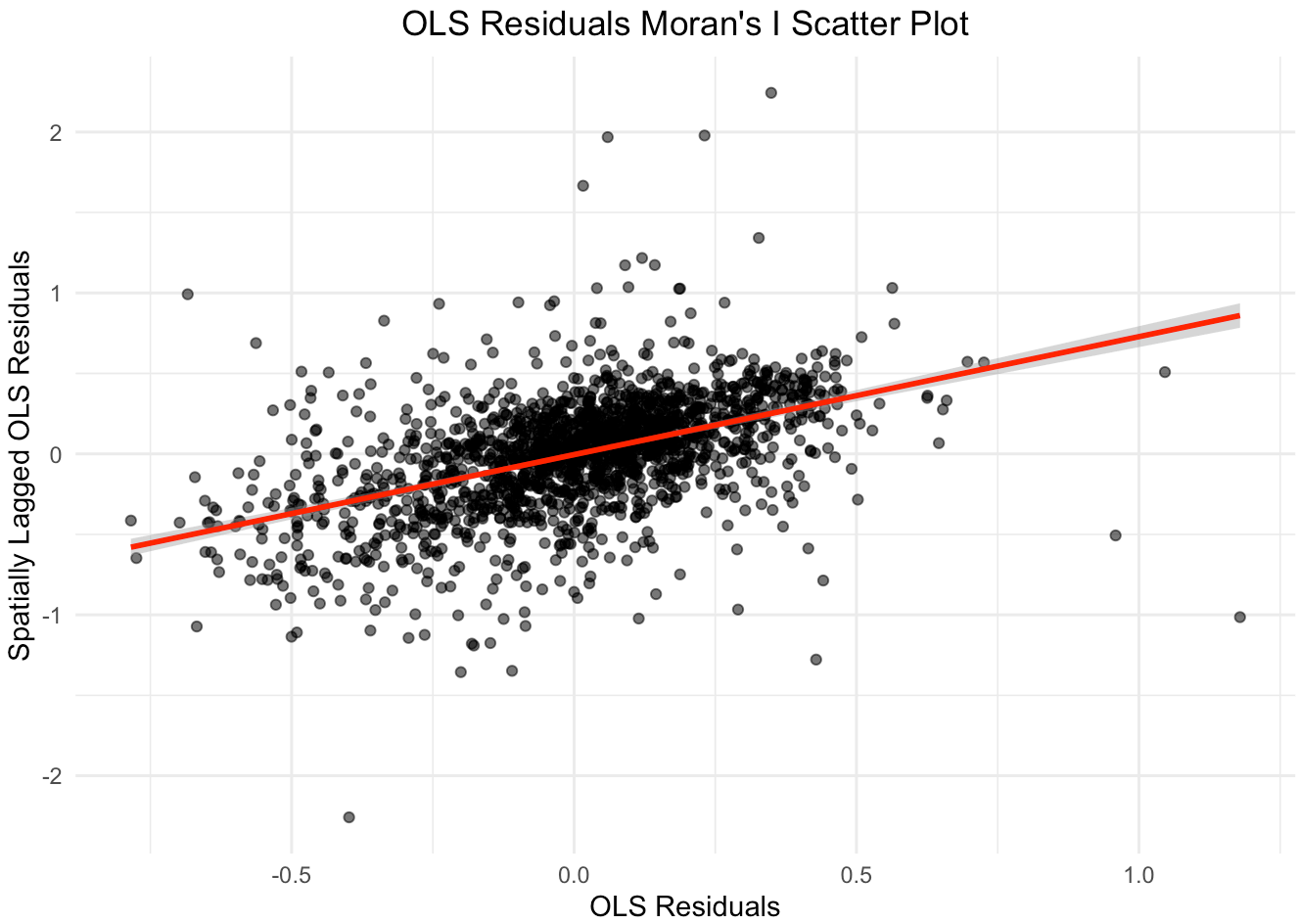

Figure 7.

The Moran's I for OLS_RESIDU is 0.31 with $p < 0.001$ (see Figure 6 for distribution), indicating that there is statistically significant, positive spatial autocorrelation in the OLS residuals. This is problematic because it means that there is a spatial pattern our model was not able to capture using the predictors that we chose.

The Moran's I for OLS_RESIDU and Beta coefficients of weighted residuals tell us a similar story – our OLS model is not capturing spatial patterns in our data. This prompts us to use other models such as the Spatial Lag and Spatial Error Regressions in order to better capture patterns in our data.

Spatial Lag and Spatial Error Results

Show

```
##
## Call:lagsarlm(formula = LNMEDHVAL ~ LNNBELPOV100 + PCTBACHMOR + PCTVACANT +
##     PCTSINGLES, data = regdata, listw = queenlist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.655421 -0.117248  0.018654  0.133126  1.726436
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##                 Estimate  Std. Error  z value             Pr(>|z|)
## (Intercept)   3.89845505  0.20111357  19.3843 < 0.00000000000000022
## LNNBELPOV100 -0.03405466  0.00629287  -5.4116         0.00000006246
## PCTBACHMOR    0.00851381  0.00052193  16.3120 < 0.00000000000000022
## PCTVACANT    -0.00852940  0.00074367 -11.4694 < 0.00000000000000022
## PCTSINGLES    0.00203342  0.00051577   3.9425         0.00008063502
##
## Rho: 0.6511, LR test value: 911.51, p-value: < 0.000000000000000222
## Asymptotic standard error: 0.01805
##     z-value: 36.072, p-value: < 0.00000000000000222
## Wald statistic: 1301.2, p-value: < 0.000000000000000222
##
## Log likelihood: -255.74 for lag model
## ML residual variance (sigma squared): 0.071948, (sigma: 0.26823)
## Number of observations: 1720
## Number of parameters estimated: 7
## AIC: 525.48, (AIC for lm: 1435)
## LM test for residual autocorrelation
## test value: 67.737, p-value: 0.00000000000000022204
```

Show

```
##
##  Likelihood ratio for spatial linear models
##
## data:
## Likelihood ratio = 911.51, df = 1, p-value < 0.00000000000000022
## sample estimates:
## Log likelihood of lagreg    Log likelihood of ols
##                -255.7400                -711.4933
```

Show

```
##
##  Breusch-Pagan test
##
## data:
## BP = 210.76, df = 4, p-value < 0.00000000000000022
```

Show

```
##
##  studentized Breusch-Pagan test
##
## data:
## BP = 51.411, df = 4, p-value = 0.0000000001832
```

Show

```
##
##  Jarque Bera Test
##
## data:  lagreg$residuals
## X-squared = 2756.9, df = 2, p-value < 0.00000000000000022
```

We present the results of the Spatial Lag regression. The W_LNMEDHVAL term in this output represents the spatial lag of the dependent variable, LNMEDHVAL. For our Spatial Lag Model, this value is non-zero, and specifically 0.65 with $p < 0.001$, meaning it is statistically significant. This means that the value of median house value is associated with the median house values nearby.

The remaining predictors (LNNBELPOV, PCTSINGLES, PCTBACHMOR, PCTVACANT) are also statistically significant ($p < 0.001$) and non-zero. The coefficients on these predictors are similar to those in the OLS results where PCTBACHMOR and PCTSINGLES are positively correlated with LNMEDHVAL, while PCTVACANT and LNNBELPOV are negatively correlated with LNMEDHVAL, through the actual values of these coefficients themselves are much smaller in the Spatial Lag model, since the spatial lag of LNMEDHVAL is able to capture much of the variance in LNMEDHVAL.

The Breusch-Pagan test results in a value of 210.76 (R) and 220.39 (GeoDa) with $p < 0.001$, signifying that the spatial lag regression residuals are still heteroscedastic.

The Spatial Lag regression resulted in an AIC of 523.48 and a Schwarz Criterion of 556.18, while these values for the OLS regression were 1432.99 and 1460.24, respectively, meaning the former regression model has a much better fit than the latter as there is less information lost. The Log Likelihood was -255.74 for the Spatial Lag regression and -711.49 for the OLS regression, again, showing that the former model has a better fit (as it has a less negative log likelihood). We observe a value of 911.51 (with $p < 0.001$) in the Likelihood Ratio Test for the Spatial Lag Regression, which tells us that this model is a better specification than OLS.

Show

Show

```
## [1] "The Global Moran's I value for the residuals for the spatial lag model is a
bout -0.08, with a p-value of about 0.002."
```

Show

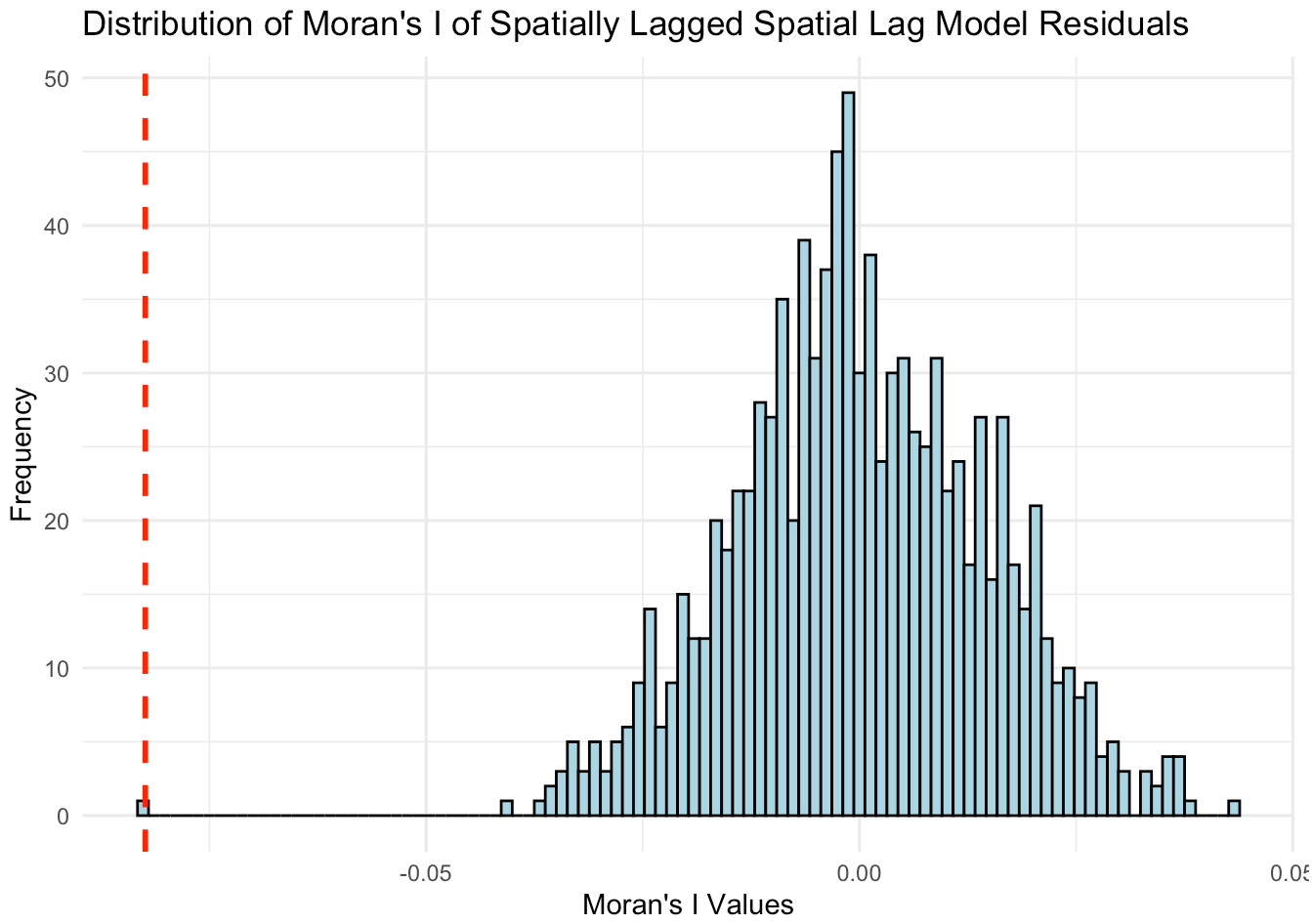

Figure 8.

Show

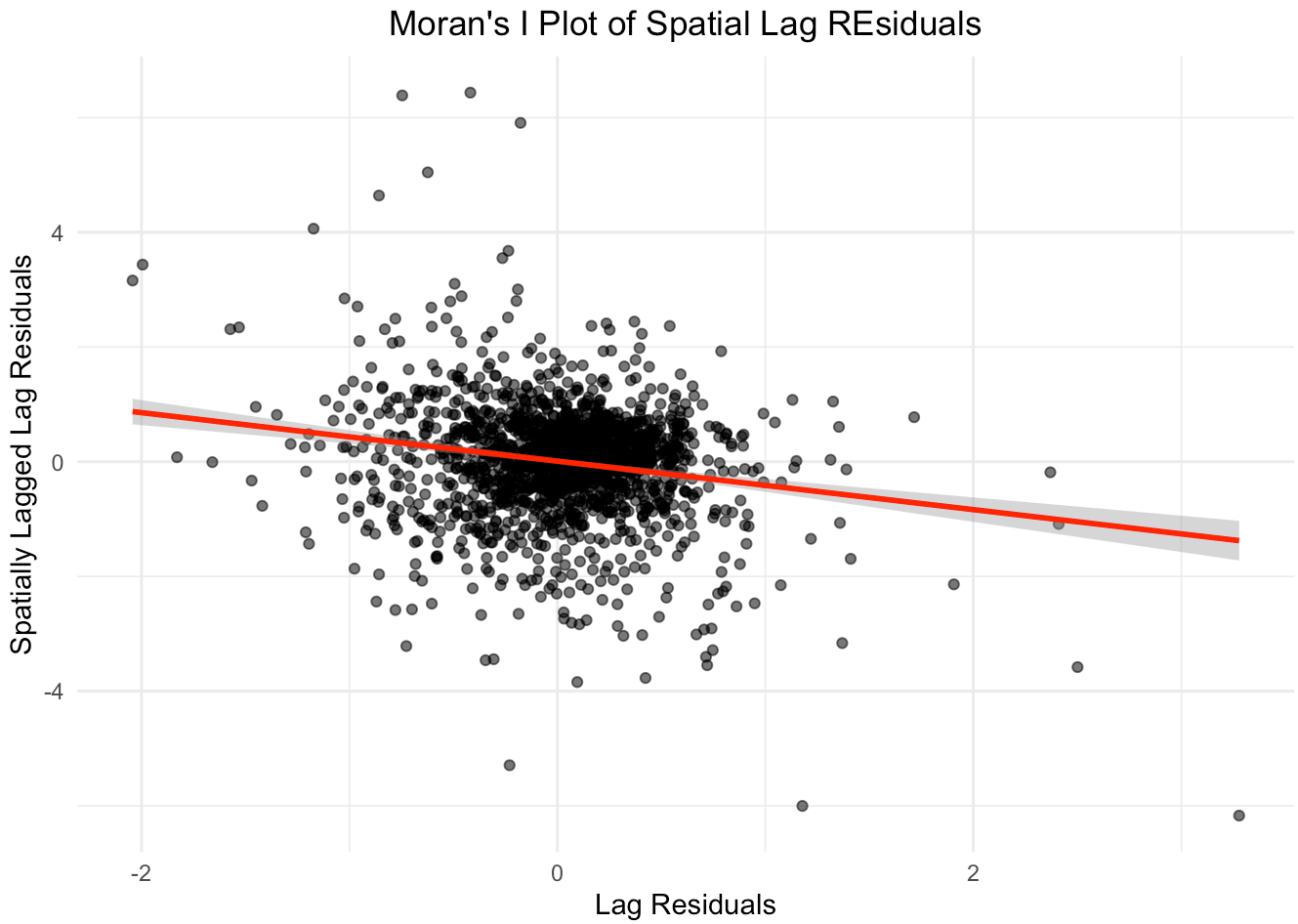

Figure 9.

The Moran's I scatter plot (Figure 9.) of spatial lag regression residuals (LAG_RESIDU) shows us that there is a slight negative correlation between LAG_RESIDU and the lagged LAG_RESIDU, with a Moran's I value of -0.082 ($p < 0.001$). Though the Moran's I histogram shows us that this value is significant, it is a very weak correlation. There seems to be much less spatial autocorrelation in these residuals than in OLS residuals.

In summary, the Spatial Lag Model outperformed the OLS model in all respects.

Show

```
##
## Call:errorsarlm(formula = LNMEDHVAL ~ LNNBELPOV100 + PCTBACHMOR +
##     PCTVACANT + PCTSINGLES, data = regdata, listw = queenlist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.926477 -0.115408  0.014889  0.133852  1.948663
##
## Type: error
## Coefficients: (asymptotic standard errors)
##                 Estimate  Std. Error  z value             Pr(>|z|)
## (Intercept)  10.90643419  0.05346781 203.9813 < 0.00000000000000022
## LNNBELPOV100 -0.03453407  0.00708933  -4.8713      0.00000110882641
## PCTBACHMOR    0.00981293  0.00072896  13.4615 < 0.00000000000000022
## PCTVACANT    -0.00578308  0.00088670  -6.5220      0.00000000006937
## PCTSINGLES    0.00267792  0.00062083   4.3134      0.00001607387709
##
## Lambda: 0.81492, LR test value: 677.61, p-value: < 0.000000000000000222
## Asymptotic standard error: 0.016373
##     z-value: 49.772, p-value: < 0.00000000000000222
## Wald statistic: 2477.2, p-value: < 0.00000000000000222
##
## Log likelihood: -372.6904 for error model
## ML residual variance (sigma squared): 0.076551, (sigma: 0.27668)
## Number of observations: 1720
## Number of parameters estimated: 7
## AIC: 759.38, (AIC for lm: 1435)
```

Show

```
##
##  Likelihood ratio for spatial linear models
##
## data:
## Likelihood ratio = 677.61, df = 1, p-value < 0.00000000000000022
## sample estimates:
## Log likelihood of errreg    Log likelihood of ols
##                -372.6904                -711.4933
```

Show

```
##
##  Breusch-Pagan test
##
## data:
## BP = 23.213, df = 4, p-value = 0.0001148
```

Show

```
##
##  studentized Breusch-Pagan test
##
## data:
## BP = 5.1627, df = 4, p-value = 0.271
```

Show

```
##
##  Jarque Bera Test
##
## data:  errreg$residuals
## X-squared = 3507, df = 2, p-value < 0.00000000000000022
```

These are the results of the Spatial Error regression. The lambda term in this output refers to the spatial lag of the OLS regression residuals. This value is non-zero, lambda = 0.81, with $p < 0.001$, meaning it is statistically significant. Once again, this is more evidence that LNMEDHVAL of one observation is very strongly positively correlated with the LNMEDHVAL of nearby observations.

Similar to the results of the Spatial Lag regression, the remaining predictors (LNNBELPOV, PCTSINGLES, PCTBACHMOR, PCTVACANT) are also statistically significant ($p < 0.001$) and non-zero in our Spatial Error regression. Again, the coefficients on these predictors are similar to those in the OLS results where PCTBACHMOR and PCTSINGLES are positively correlated with LNMEDHVAL, while PCTVACANT and LNNBELPOV are negatively correlated with LNMEDHVAL, through the actual values of these coefficients themselves are much smaller in the Spatial Error model, since the the lambda coefficient is able to capture much of the variance in LNMEDHVAL.

The Breusch-Pagan test results in a value of 23.213 (R) 210.99 (GeoDa) with $p < 0.001$, meaning the spatial error regression residuals are still heteroscedastic. The difference between these two tests is very dramatic and needs to be investigated by looking into the formulas both platforms use.

The Spatial Error regression resulted in an AIC of 755.38 and a Schwarz Criterion of 782.63, while these values for the OLS regression were 1432.99 and 1460.24, respectively, meaning the former regression model has a much better fit than the latter as there is less information lost. The Log Likelihood was -372.69 for the Spatial Error regression and -711.49 for the OLS regression, again, showing that the former model has a better fit (as it has a less negative log likelihood). We observe a value of 677.61 (with $p < 0.001$) in the Likelihood Ratio Test for the Spatial Error Regression which tells us that this model is a better specification than OLS.

Show

```
## [1] "The Global Moran's I value for the residuals for the spatial error model is
about  -0.09 , with a p-value of about  0.002 ."
```

Show

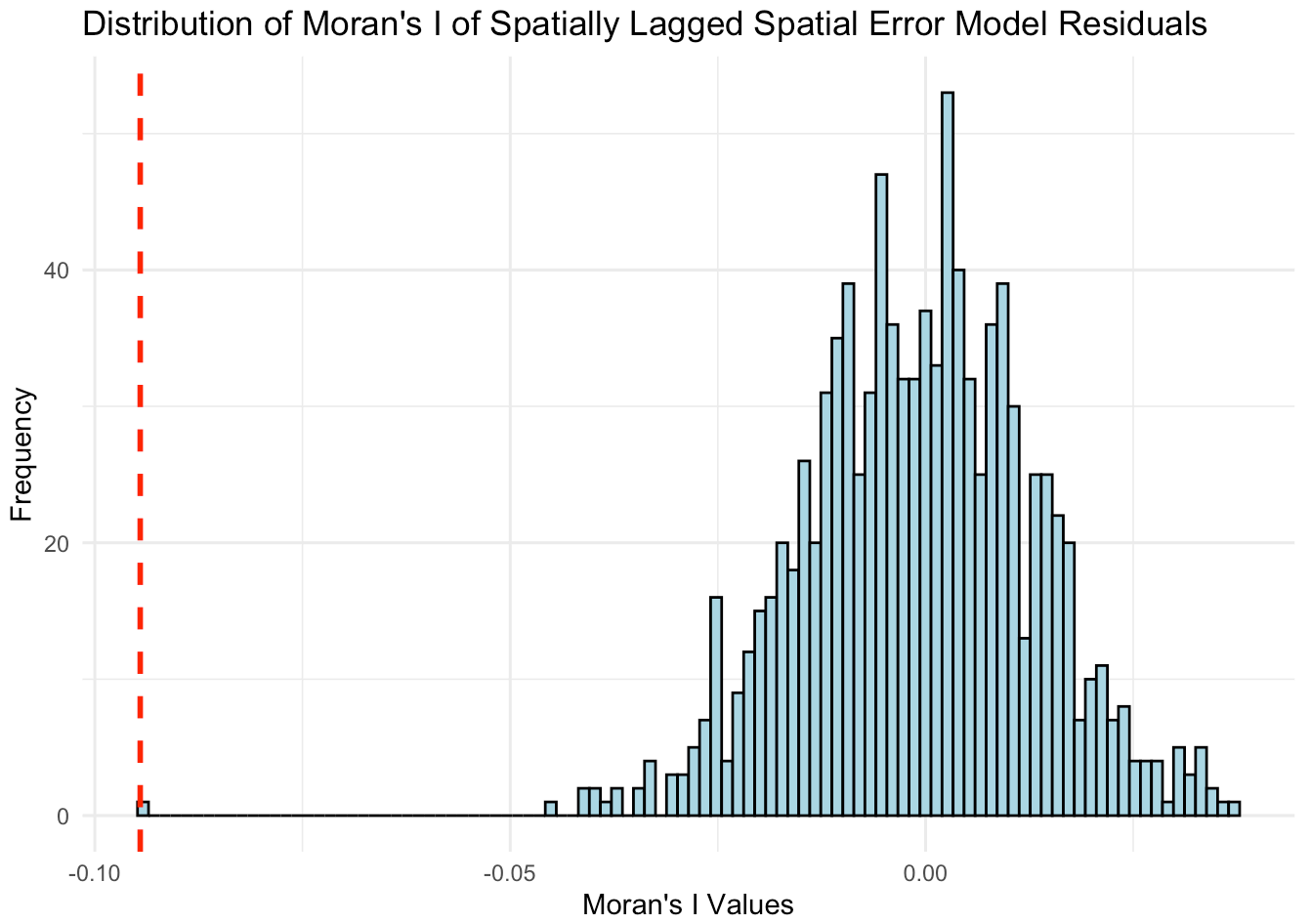

Figure 10.

Show

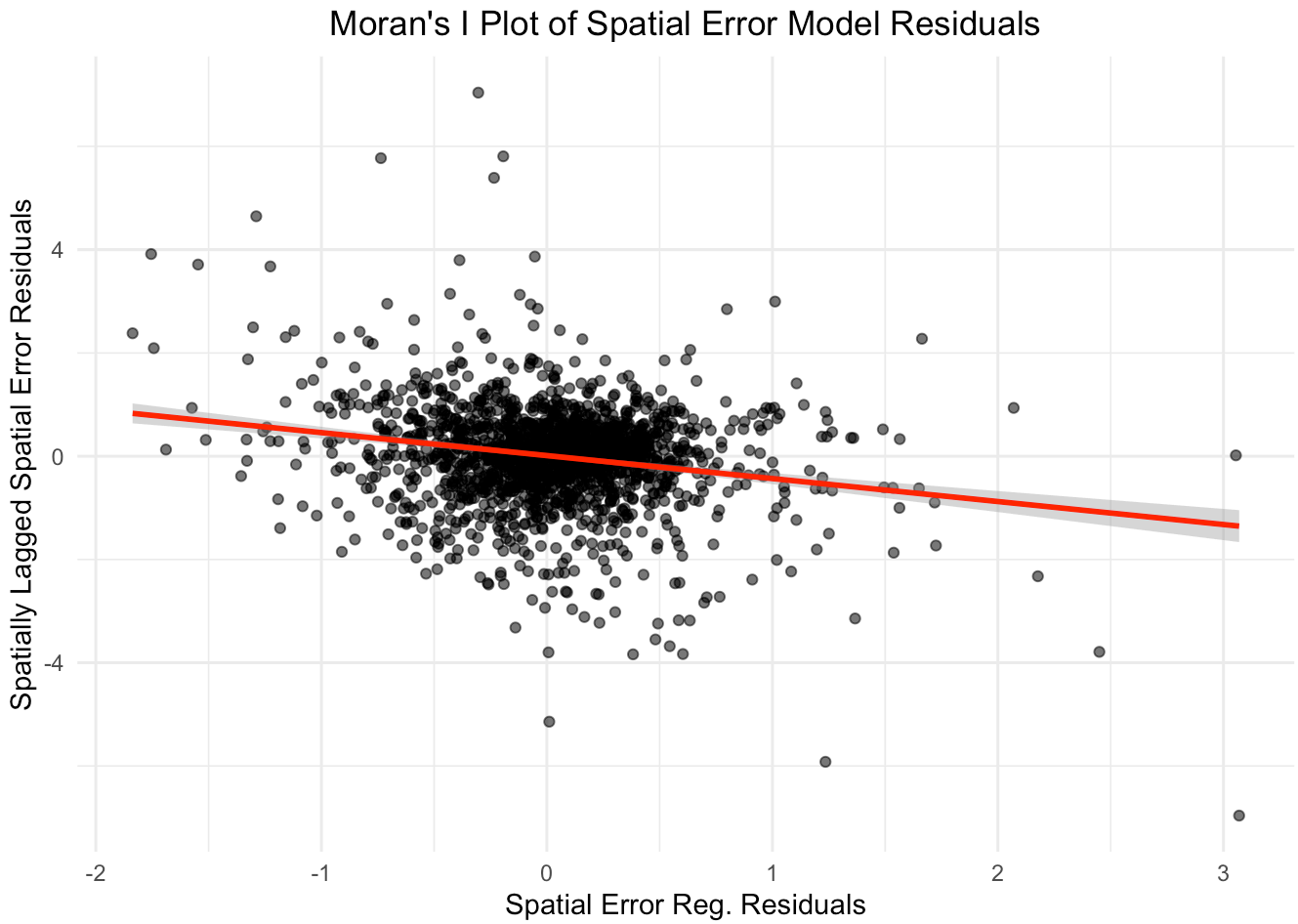

Figure 11.

The Moran's I scatter plot (Figure 11.) of spatial error regression residuals (ERR_RESIDU) shows us that there is a slight negative correlation between ERR_RESIDU and the lagged ERR_RESIDU, with a Moran's I value of -0.095 ($p < 0.001$). Though the Moran's I histogram shows us that this value is significant, it is a very weak correlation. There seems to be much less spatial autocorrelation in these residuals than in OLS residuals.

In sum, though the Spatial Error regression outperformed the OLS regression, the Spatial Lag regression reigns superior when comparing all three models. The Spatial Error model resulted in an AIC and SC of 755.38 and 782.63, respectively, while the Spatial Lag model resulted in an AIC and SC of 523.48 and 556.18, respectively. Recall that we can say a model has a better goodness of fit when it has a lower AIC and SC, thus, we can conclude that the Spatial Lag model outperforms the Spatial Error model in this respect. The Spatial Lag model also results in residuals that are slightly less autocorrelated (the Spatial Lag residuals show a lower absolute value of Moran's I). D

Geographically Weighted Regression Results

Show

Show

```
## Call:
## gwr(formula = LNMEDHVAL ~ LNNBELPOV100 + PCTBACHMOR + PCTVACANT +
##     PCTSINGLES, data = shps, gweight = gwr.Gauss, adapt = bw,
##     hatmatrix = TRUE, se.fit = TRUE)
## Kernel function: gwr.Gauss
## Adaptive quantile: 0.008130619 (about 13 of 1720 data points)
## Summary of GWR coefficient estimates at data points:
##                   Min.    1st Qu.     Median    3rd Qu.       Max.  Global
## X.Intercept.  9.6727618 10.7143173 10.9542384 11.1742009 12.0831381 11.1138
## LNNBELPOV100 -0.2365244 -0.0733572 -0.0401186 -0.0126657  0.0948768 -0.0789
## PCTBACHMOR    0.0010974  0.0101380  0.0149279  0.0202187  0.0347258  0.0209
## PCTVACANT    -0.0317407 -0.0142383 -0.0089599 -0.0035770  0.0167916 -0.0192
## PCTSINGLES   -0.0249706 -0.0075550 -0.0016626  0.0042280  0.0143340  0.0030
## Number of data points: 1720
## Effective number of parameters (residual: 2traceS - traceS'S): 360.5225
## Effective degrees of freedom (residual: 2traceS - traceS'S): 1359.477
## Sigma (residual: 2traceS - traceS'S): 0.2762201
## Effective number of parameters (model: traceS): 257.9061
## Effective degrees of freedom (model: traceS): 1462.094
## Sigma (model: traceS): 0.2663506
## Sigma (ML): 0.245571
## AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 660.7924
## AIC (GWR p. 96, eq. 4.22): 308.7123
## Residual sum of squares: 103.7248
## Quasi-global R2: 0.8479244
```

Show

```
## Object of class SpatialPolygonsDataFrame
## Coordinates:
##         min       max
## x 2660604.8 2750171.3
## y  207610.6  304858.8
## Is projected: NA
## proj4string : [NA]
## Data attributes:
##      sum.w        X.Intercept.     LNNBELPOV100        PCTBACHMOR
##  Min.   :16.03   Min.   : 9.673   Min.   :-0.23652   Min.   :0.001097
##  1st Qu.:24.47   1st Qu.:10.714   1st Qu.:-0.07336   1st Qu.:0.010138
##  Median :26.64   Median :10.954   Median :-0.04012   Median :0.014928
##  Mean   :27.48   Mean   :10.937   Mean   :-0.04485   Mean   :0.015267
##  3rd Qu.:29.45   3rd Qu.:11.174   3rd Qu.:-0.01267   3rd Qu.:0.020219
##  Max.   :86.70   Max.   :12.083   Max.   : 0.09488   Max.   :0.034726
##    PCTVACANT           PCTSINGLES        X.Intercept._se   LNNBELPOV100_se
##  Min.   :-0.031741   Min.   :-0.024971   Min.   :0.09911   Min.   :0.01707
##  1st Qu.:-0.014238   1st Qu.:-0.007555   1st Qu.:0.19114   1st Qu.:0.03521
##  Median :-0.008960   Median :-0.001663   Median :0.23474   Median :0.04198
##  Mean   :-0.009192   Mean   :-0.002074   Mean   :0.25013   Mean   :0.04413
##  3rd Qu.:-0.003577   3rd Qu.: 0.004228   3rd Qu.:0.29127   3rd Qu.:0.05035
##  Max.   : 0.016792   Max.   : 0.014334   Max.   :0.54791   Max.   :0.09856
##  PCTBACHMOR_se        PCTVACANT_se      PCTSINGLES_se          gwr.e
##  Min.   :0.0007667   Min.   :0.001821   Min.   :0.001177   Min.   :-1.50370
##  1st Qu.:0.0025261   1st Qu.:0.004201   1st Qu.:0.003560   1st Qu.:-0.09867
##  Median :0.0048373   Median :0.005458   Median :0.005214   Median : 0.01654
##  Mean   :0.0049127   Mean   :0.006536   Mean   :0.005118   Mean   : 0.01099
##  3rd Qu.:0.0066118   3rd Qu.:0.007381   3rd Qu.:0.006596   3rd Qu.: 0.12800
##  Max.   :0.0151900   Max.   :0.030192   Max.   :0.010560   Max.   : 1.67766
##       pred           pred.se           localR2       X.Intercept._se_EDF
##  Min.   : 9.578   Min.   :0.02931   Min.   :0.1337   Min.   :0.1028
##  1st Qu.:10.476   1st Qu.:0.05601   1st Qu.:0.5231   1st Qu.:0.1982
##  Median :10.831   Median :0.06773   Median :0.6342   Median :0.2434
##  Mean   :10.871   Mean   :0.07462   Mean   :0.6186   Mean   :0.2594
##  3rd Qu.:11.232   3rd Qu.:0.08449   3rd Qu.:0.7312   3rd Qu.:0.3021
##  Max.   :13.307   Max.   :0.23204   Max.   :0.8863   Max.   :0.5682
##  LNNBELPOV100_se_EDF PCTBACHMOR_se_EDF   PCTVACANT_se_EDF   PCTSINGLES_se_EDF
##  Min.   :0.01770     Min.   :0.0007951   Min.   :0.001889   Min.   :0.001221
##  1st Qu.:0.03651     1st Qu.:0.0026197   1st Qu.:0.004357   1st Qu.:0.003692
##  Median :0.04354     Median :0.0050166   Median :0.005661   Median :0.005407
##  Mean   :0.04576     Mean   :0.0050947   Mean   :0.006778   Mean   :0.005307
##  3rd Qu.:0.05221     3rd Qu.:0.0068568   3rd Qu.:0.007654   3rd Qu.:0.006841
##  Max.   :0.10222     Max.   :0.0157529   Max.   :0.031311   Max.   :0.010951
##    pred.se.1
##  Min.   :0.03040
##  1st Qu.:0.05808
##  Median :0.07024
##  Mean   :0.07739
##  3rd Qu.:0.08762
##  Max.   :0.24063
```

Our GWR yielded an overall R-squared value of 0.85. This is an improvement over the R-squared of our OLS regression 0.66. The GWR does a better job of explaining the variance in median home values, with 85% of the variance in the dependent variable being explained by the model compared to 66% with OLS.

The AIC of GWR was 308.71. For OLS, 41349.6. For the spatial lag model, 523.48. For the spatial error model, 755.381. A lower AIC indicates a better fit, with the GWR having the lowest AIC of all four models. Based on this metric, the GWR is a better fit to our data than the OLS, spatial lag, and spatial error models.

Show

```
## [1] "The Global Moran's I value for the residuals for the spatial gwr model is a
bout 0.03, with a p-value of about 0.016."
```

Show

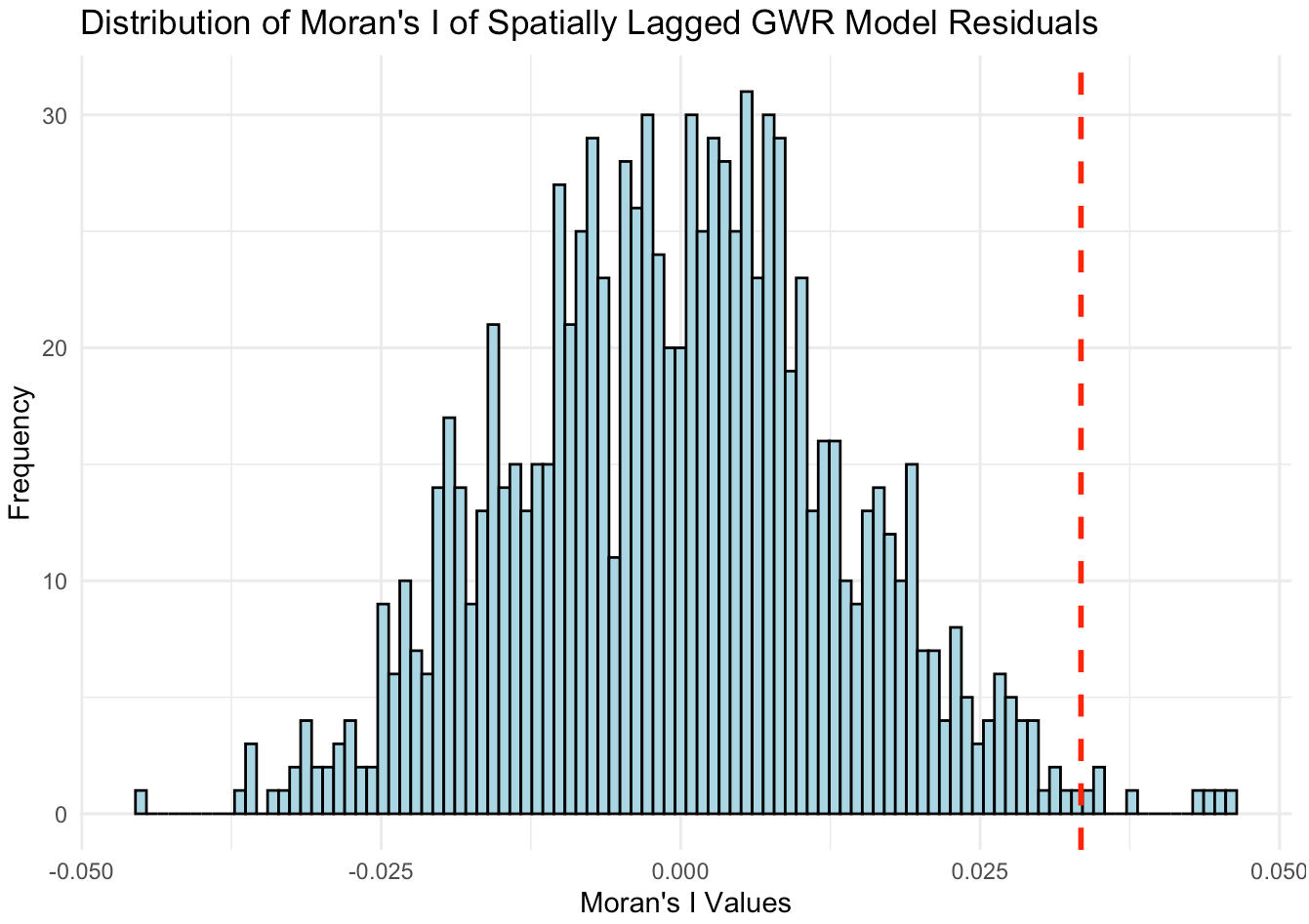

Figure 12.

Show

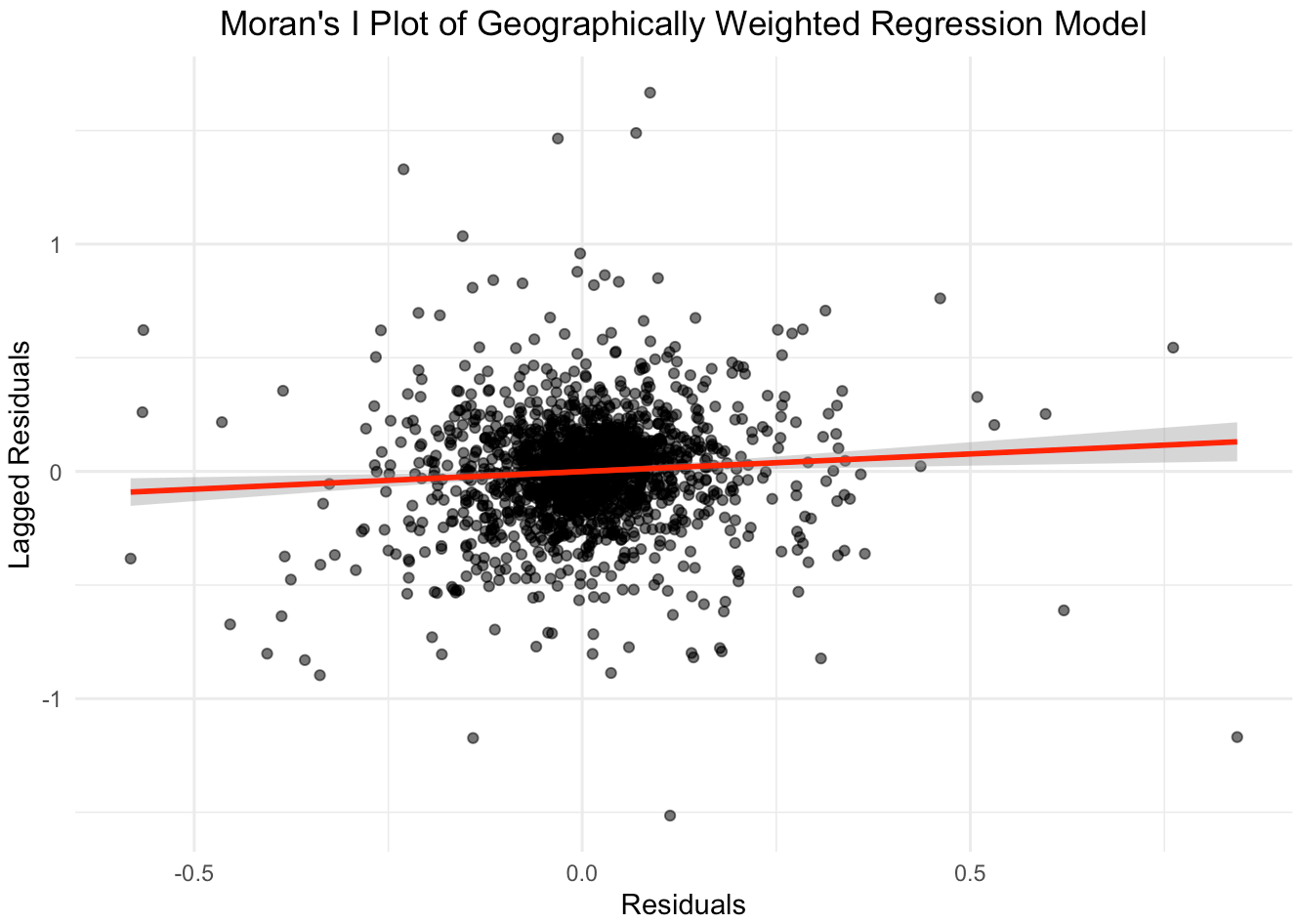

Figure 13.

The Moran's I scatter plot of the GWR residuals (Figure 13.) displays less spatial autocorrelation than the other three models. Our scatter plot of the OLS residuals showed a moderate linear relationship between the residuals and the lagged residuals with a Moran's I statistic of 0.313, suggesting positive spatial autocorrelation. The Spatial Lag model has a lower Moran's I statistic at -0.082 and shows a much weaker, negative relationship between the residuals and the lagged residuals. The Spatial Error model is similar; the Moran's I statistic is -0.095 and the plot shows a similar negative relationship between the residuals, where there is still some spatial autocorrelation. The GWR residuals have no clear relationship with the lagged residuals and the Moran's I statistic is smaller than the rest at 0.03. This indicates that the GWR is better suited to address the autocorrelation in our data.

As we'll see later, local results can vary. Some models may have R squared values near zero (see Figure 13.), which can mean that a model where the predictor is regressed on only nearby observations does not have high predictive power in that area. In this case, it is good to look at individual variables for their beta values in these individual models. We will look closer at the beta coefficients for each variable accross space in Figure 12. to examine this more.

Show

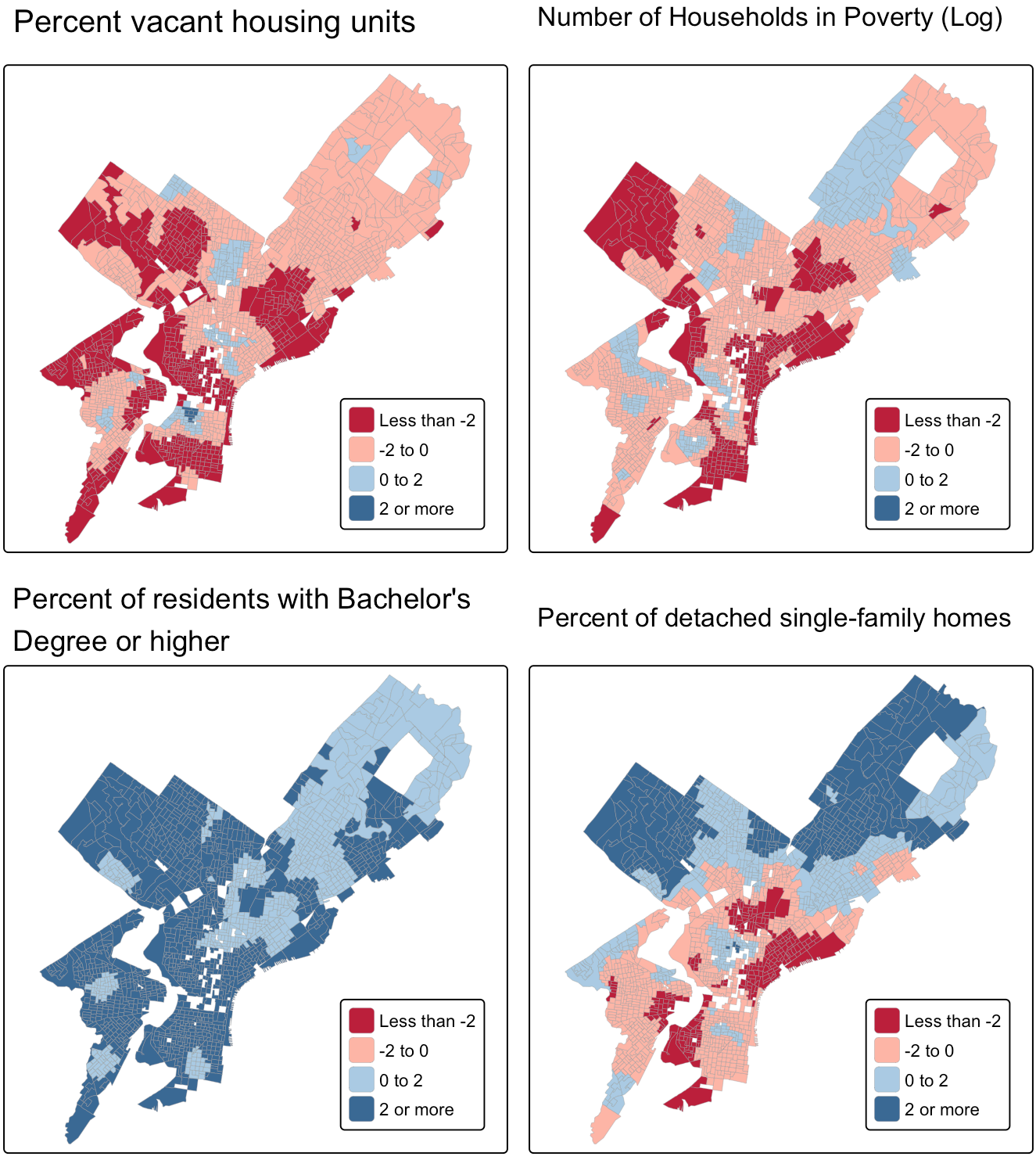

Figure 14

Dividing the coefficients of the local regressions by their standard errors allows us to identify which predictors have a possibly significant relationship with the dependent variable. A coefficient/standard error value of -2 or less indicates a negative relationship with the dependent variable that is possibly significant. A coefficient/standard error value of 2 or greater indicates a positive relationship with the dependent variable that is possibly significant. Coefficient/standard error values falling between -2 and 0 or 0 and 2 represent negative and positive relationships with the dependent variable, respectively, but it is unlikely that these relationships are significant. When mapping the spatial distribution of these values, we specified our breaks and color scale to make negative, possibly significant relationships dark red; and positive, possibly significant relationships dark blue.

There are many parts of the city where the percent of residents with a bachelor's degree or higher has a positive, possibly significant relationship with median house value. Areas like Fairmount, West Philadelphia, and Center City have these significant relationships. Most of the coefficient/standard error values for this predictor have some sort of positive relationship, but in certain parts of South Philadelphia and Northeast Philadelphia, it is unlikely that these relationships are significant. The log-transformed number of residents living in poverty has a negative, possibly significant relationship with the dependent variable in certain areas of the city.

The percentage of vacant housing units shows possibly significant relationships, both negative and positive. In a small portion of Center City, this variable has a positive relationship with the dependent variable. In other areas where the relationship is possibly significant, it is negative. The percentage of detached single-family homes also has areas with both positive and negative possibly significant relationships. In some of the northern parts of the city that are near Montgomery County, the percentage of detached single-family homes has a positive relationship with the dependent variable. In areas like University City and Kensington, the relationship is negative.

Observing the spatial distribution of these values is one way for us to identify regional variation in our model.

Show

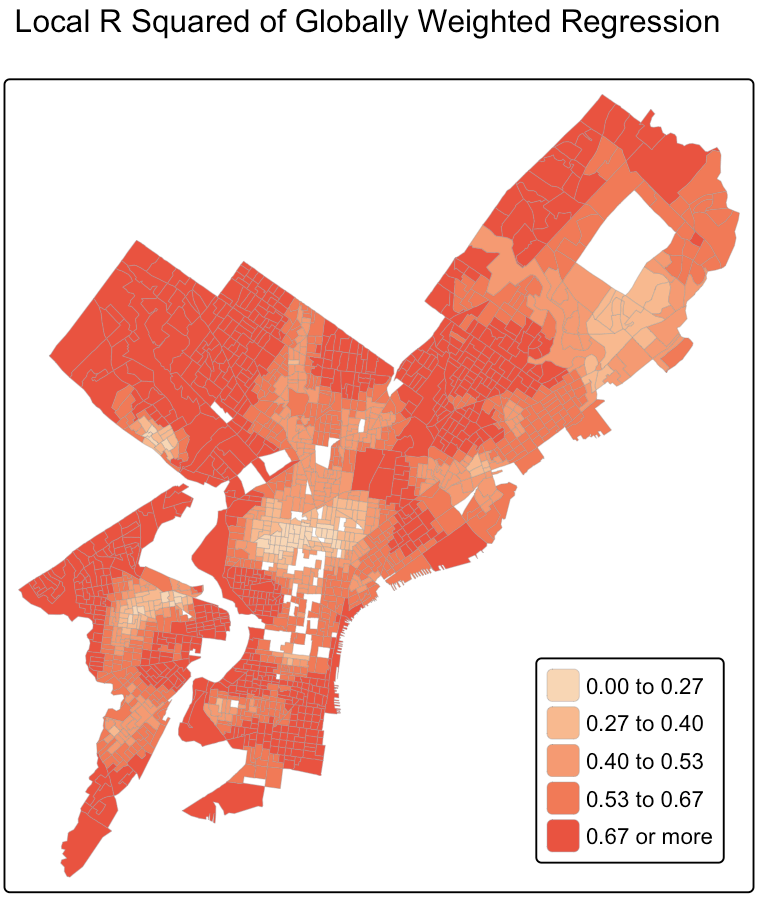

Figure 15.

This choropleth map visualizes the local R-squared values for each point in the GWR (Figure 15), which represent how much of the variance in the dependent variable is explained by the local regression model. Local R-squared values differ across each local regression, meaning that our model explains more variation

in some areas than others. These differences highlight the areas where our selected characteristics are not good predictors of median house value.

Discussion

In this report, we examined four different models to predict Median House Value (specifically, log-transformed median house value) using four predictors (percentage of detached single-family homes, percentage of residents who hold a bachelor's degree or above, percentage of vacant housing units, and log-transformed number of households living in poverty). The models we used were Ordinary Least Squares (OLS), Spatial Lag, Spatial Error, and Geographically Weighted Regressions.

Out of all four models, the GWR performed the best, as it has the lowest AIC, meaning it has the best goodness of fit. In addition, we observed the lowest Moran's I value for GWR residuals compared to the Moran's I values of residuals from the other models, meaning that this model was able to capture spatial characteristics that help predict LNMEDHVAL the best.

In homework 1 we addressed how our data did not meet the OLS assumption of independence of observations since the data are spatially autocorrelated. We created a spatial lag model and a spatial error model to address the spatial autocorrelation and saw some success, but both of these models assume spatial stationarity, or that modeled relationships are constant across space. This assumption does not hold for our data. Geographically weighted regression was an effective method that allowed us to assess how our model performed across space and was the most favorable in terms of AIC, Moran's I, and Global R-squared. Through these observations we came to a better understanding of where our model performs well and the significance of certain predictors across space. However, GWR is largely considered to be an exploratory tool, not a substitute for other spatial methods.

Spatially lagged residuals can be used to identify spatial dependencies in a model by seeing if residuals are similar for nearby observations. For each observation, the spatially lagged residual is the value or average value of its neighbor's residuals. In this report, we used queen weights, meaning that any intersecting block is considered a neighbor. The lagged residuals are variables created using the queen neighbor weights and the residuals from an OLS regression.

Spatial lag model residuals can also help identify spatial dependencies in our model, but are distinct from the weighted residuals of the OLS regression. In a spatial lag model, we specify a spatial weights matrix to calculate the lagged values of the dependent variable, using these new values as a predictor in an entirely new model. The residuals from a spatial lag model reflect some amount of spatial dependency in the data, unlike the weighted residuals from the OLS model that does not account for space. The spatial lag model residuals can also be weighted to further test for spatial autocorrelation.

Both weighted OLS residuals and spatial lag model residuals are helpful, but serve different purposes. After creating an OLS model, we can observe whether there is a relationship between the residuals and weighted residuals and determine whether further analyses are necessary to address spatial autocorrelation. One such analysis would be a spatial lag model, and observing the spatial lag model residuals and the lagged spatial lag model residuals would indicate the effectiveness of the model in accounting for spatial autocorrelation.

ArcGIS has a GWR button, but it is problematic for a couple of reasons. First, ArcGIS uses their "Golden Search" algorithm that is not always guaranteed to find the lowest AICc when looking for the optimum bandwidth to use for the regression. Second, ArcGIS only reports AICc, and not AIC, making it difficult to compare the output of the GWR to other model outputs.