Assignment 2 - Spatial Regression

Richard Barad, Dave Drennan, and Jarred Randall

2023-11-14

Contents

Introduction

In this analysis we use a range of different regression models to examine the relationship between the median home sales value by census block for homes in Philadelphia and four predictors. The predictors are:

- **PCTBACHMOR:** proportion of residents in a block group with at least a bachelor's degree
- **PCTVACANT**: proportion of housing units that are vacant
- **PCTSINGLE:** percent of housing units that are detached single family houses
- **LNNBELPOV:** natural log of the number of households living in poverty

In our previous report, we used Ordinary Least Squares (OLS) regression to regress the median home sale values against our four predictors. OLS regression methods do not always perform well when the predictors and/or the dependent variable are spatially clustered and not randomly distributed in space. In this report, we examine the results of regression methods which incorporate spatial clustering. Specifically, we use the spatial error regression, spatial lag regression, and geographic weighted regression (GWR) to assess if these methods perform better than OLS. We will also assess how the spatial models compare to each other.

Methods

A Description of the Concept of Spatial Autocorrelation

Spatial autocorrelation, stemming from temporal autocorrelation (values of a variable at points close in time will be related), describes the relationship for a single variable between the value for the variable at a specific location and at nearby locations. This concept is closely related to Waldo Tobler's first law of geography, which states that "Everything is related to everything else, but near things are more related than distant things." When spatial autocorrelation is present, values of a variable in nearby areas are related to each other and are not independent. Positive spatial correlation is present if observations that are closer to each other in space have similar values. Conversely, negative spatial autocorrelation can be observed if observations that are closer to each other have noticeably different values.

Spatial autocorrelation emphasizes the importance of spatial proximity in understanding the relationships and interactions between geographical features or phenomena. Spatial proximity, delineated through aerial data using polygons, aims to determine the spatial associations between each polygon and all other polygons. There are two main approaches for measuring proximity: distance-based measures of proximity and contiguity-based measures of proximity. Distance-based measures check whether polygon centroids (the centers of gravity of the polygons) are within a certain distance of each other (1=yes, 0=no), whereas contiguity-based measures check whether polygons share a boundary with each other (1=yes, 0=no).

Moran's I

Moran's I is a widely used method of testing for spatial autocorrelation in a dataset. A Moran's I value close to 1 indicates strong positive autocorrelation, signifying that similar values tend to cluster together. Large negative values close to -1 suggest the presence of strong negative spatial autocorrelation, indicating that areas with similar values of a variable tend to repel each other (i.e., dispersion). Values around 0 indicate that there is no spatial autocorrelation present (i.e., random pattern). The formula for Moran's I is stated as:

$$I = \frac{\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(X_i-\bar{X})(X_j-\bar{X})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}\right)}{\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}\right)}$$

The components of the Moran's I formula are defined as where:

- $\bar{X}$ is the mean of the variable X
- X_i is the variable value at a particular location i
- X_j is the variable value at another location j
- w_ij is a weight indexing location of i relative to j
- n is the number of observations (points or areal units)

Weight Matrices

Weight matrices serve as a fundamental tool in spatial analysis that establish a structured framework that defines the neighboring relationships for each location within the spatial dataset. By assigning specific weights, the matrix quantifies the strength or intensity of the connections between pairs of spatial units, allowing for a comprehensive exploration of the spatial relationships present in the data.

Weight matrices are created when we have n observations and form an nxn table which summarizes all the pairwise relationships in the dataset. These weight matrices are used in the estimation of spatial regression and the calculation of spatial autocorrelation indices. While different types of weight matrices exist, this report primarily employs Cliff and Ord's queen's matrix. Queen's matrix is a contiguity-based measure of proximity which checks if polygons share a boundary at either a point or a segment.

While this report will primarily utilize one specific weight matrix, it is common practice among statisticians to experiment with various types of weight matrices to ensure that the results are not purely an artifact of the matrix being used.

Moran's I Significance Test

The significance test for spatial autocorrelation (Moran's I) helps to determine whether the observed spatial pattern is statistically significant or whether it can be attributed to randomness. By conducting this test, we can either accept or reject the null hypothesis, providing us with insights into the presence or absence of spatial autocorrelation in the data. Here we are testing whether the dependent variable LNMEDHVAL is significantly spatially autocorrelated. The null and alternative hypotheses are stated as:

- H_0: LNMEDHVAL is not spatially autocorrelated.
- H_{a_1}: LNMEDHVAL is positively spatially autocorrelated.
- H_{a_2}: LNMEDHVAL is negatively spatially autocorrelated.

To test the significance of spatial autocorrelation we employ a permutation test approach. The first step in significance testing is to compute Moran's I of LNMEDHVAL. We then create a shuffle map which randomly shuffles (or permutes) the values of LNMEDHVAL and calculates the Moran's I of for this shuffle map. We then repeat this shuffling process a total of 999 times and calculate the Moran's I for each resulting permutation (shuffle). We then arrange the calculated Moran's I for each of the 999 permutations plus the original one calculated from LNMEDHVAL in descending order to see where the Moran's I value for the LNMEDHVAL stands in comparison to the Moran's I for the 999 random permutations.

Next, the pseudo p-value is calculated by taking the rank of Moran's I and dividing it by 1000. A high Pseudo P-value that is greater than Moran's I of the LNMEDHVAL implies that no spatial autocorrelation is present under the assumption of spatial randomness, and we fail to reject the null hypothesis. Conversely, if the pseudo p-value is lower than that of the Moran's I of LNMEDHVAL, significant spatial autocorrelation is present, and we can reject the null hypothesis of no spatial autocorrelation.

Local Spatial Autocorrelation

While global Moran's I is effective in detecting spatial processes such as clustering and dispersion, it does not provide specific information about the precise locations of individual clusters or outliers within our overall dataset. For this information, we use local indices of spatial autocorrelation (LISA), which measure the degree to which values at neighboring locations are associated with the value at a specific site or location or to what extent are values at sites in vicinity of i associated with i. In local spatial autocorrelation (local Moran's I), the significance testing involves shuffling or permuting the values of the variable X among the various locations in the dataset while keeping the value at the specific location of interest (location i) unchanged. Typically, these reshufflings are performed multiple times, ranging from 9 to 999, and the li value is computed for each location for every reshuffling.

The li value for the original dataset is then ranked relative to the list of values generated by the reshuffling process. If the li values for the original dataset are notably low or high in comparison to the list of results obtained through the shuffling process, they are deemed significant. A pseudo significance value is determined by noting the rank of the actual value of li relative to the permuted results. Positive local spatial autocorrelation tells us of clustering of similar values near location i, while negative local spatial autocorrelation tells us that location i is a spatial outlier relative to neighboring locations.

A Review of OLS Regression and Assumptions

Ordinary least square (OLS) regression is a statistical method used to examine the linear relationship between a variable of interest (dependent variable) and one or more explanatory variables (predictors). OLS tests the strength of the

relationship, the direction of the relationship (positive, negative, or no relationship) and goodness of model fit - how well a model will predict a future set of observations. Regressions can also calculate the amount that the dependent variable changes when a predictor variable changes by one unit (holding all other predictors constant). However, if an explanatory variable is a significant predictor of the dependent variable, it does not imply causation. Prior to making conclusions about the model estimates or using the model for predictions, certain assumptions of OLS must be met. These assumptions include linearity, independence of observations, normality of residuals, homoscedasticity, and no multicollinearity. For a more detailed report on OLS regression please refer to "Assignment 1 – OLS Regression".

OLS Assumptions & Spatial Components

When spatial components are present in the data, the OLS assumptions of independence of observations and homoscedasticity (randomness of errors) do not hold. When spatial autocorrelation is present the values of a variable in nearby areas are related to each other and are not independent. Additionally, if we have spatially autocorrelated OLS residuals, there is systemic under/over prediction in certain parts of the data: furthermore, the significance estimates for the beta coefficients (β_i) in OLS may be inflated. We can test these assumptions by examining the spatial autocorrelation of the residual using Moran's I.

Another way to test OLS residuals for spatial autocorrelation is to regress them on nearby residuals. In this report, these nearby residuals are residuals at neighboring block groups, as defined by the queen matrix. The first step is to generate standardized OLS residuals by dividing the OLS model residuals by an estimate of their standard deviation. Then, we regress the OLS standardized residuals on the spatial lag of the OLS residuals (i.e., OLS residuals at the queen's neighbor). This test produces the beta coefficient (β_i) of the lagged residuals which is referred to as slope b.

Testing Regression Assumptions

This report utilizes the open-source statistical software R to conduct our OLS regressions, which offers various methods for testing OLS assumptions, including assessments of homoscedasticity and the normality of errors. When we violate the assumption of homoscedasticity, which is tied to assumption of independence of errors, we have heteroscedasticity. Heteroscedasticity is the variance in the residual that may change with the values of another variable. We can test heteroscedasticity in R using the following:

- The Breusch-Pagan Test - for heteroscedasticity and random coefficient variation, this test calculates its test statistic using the squared residuals from the OLS regression model. This test compares the variance of the residuals with the variance predicted by a simple auxiliary regression model that uses the same independent variables to predict the squared residuals.
- The Koenker-Basset Test (The studentized Breusch-Pagan test) - for heteroscedasticity based on regression quantiles is a modification of the Breusch-Pagan test that takes into account the potential influence of individual data points. A high value for this statistic indicates the variance of the residuals is related to the magnitude of the predicted values.
- The White Test - a heteroscedasticity-consistent covariance matrix estimator and also assesses heteroscedasticity. It does so without requiring a specific form for the alternative hypothesis of changing variances. This makes White's test a more general check for heteroscedasticity than the Breusch-Pagan test, which relies on the model's independent variables.

For these three tests, the null hypothesis is that of homoscedasticity. If the p-value is less than 0.05, we can reject the null hypothesis for the alternative hypothesis of heteroscedasticity.

The normality of residuals (errors) should not contain any systematic meaningful information and they should be normally distributed. The Jarque Bera test is used in R to examine the null hypothesis that the residuals are from a normal distribution. If $p < 0.05$, we reject the null hypothesis of normality for the alternative hypothesis of non-normality.

Spatial Lag and Spatial Error Regression

We use the open source software R to run a spatial lag regression and a spatial error regression on our data set.

Spatial Lag

The spatial lag regression model builds on the OLS regression model by associating nearby values of the dependent variable as a predictor in the model. We use a queen-neighbor weights matrix in our model to associate an individual median home value of a block group with nearby home value block groups. To account for this spatial lag of y in the model, the regression equation includes a y-lag variable with rho as a coefficient for each observation in the data set.

$$LNMEDHVAL = \rho WLNMEDHVAL + \beta_0 + \beta 1 PCTVACANT + \beta 2 PCTSINGLES + \beta 3 PCTBACHMOR + \beta 4 LNNBELPOV + \epsilon$$

Spatial lag regression includes the spatial lag term of the dependent variable y (e.g. LNMEDHVAL) as a predictor in addition to the terms included in an OLS regression. The coefficient rho, denoted by p, is limited to a value between -1 and 1. The predictor Wy represents the weights matrix W applied to the dependent variable y, which allows us to associate the nearby ys (the log of median home values in nearby Census block groups) with an individual observation. Taken together, rho and Wy are the spatial lag of y as a predictor for y.

The coefficient β_i of each predictor is more complicated to interpret for spatial lag models compared to OLS regression models. Our interpretation of the spatial lag model will focus on when the lag regression term is positive or negative.

The variable epsilon, denoted by ϵ, is commonly referred to as the residual term or random error term in the model. The residual term allows the regression surface to fall above ($\epsilon > 0$) or below ($\epsilon < 0$) the actual data points. Epsilon is the difference between observed values of y (e.g. LNMEDHVAL) and the values of y predicted by the regression model (denoted by $\hat{y}$).

Spatial Error

The spatial error regression model also builds on the OLS regression model by associating nearby residuals as a predictor using an individual observation's residual term. For our spatial error model, we also use queen neighbors to associate nearby residual values with an observation residual value. The spatial error model first runs an OLS regression, regressing y on the predictor variables. Residuals are then regressed on neighbor residuals to filter out the spatial component of the OLS residuals. The epsilon is split into a spatial component term and a term that represents random noise.

$$LNMEDHVAL = \beta_0 + \beta 1 PCTVACANT + \beta 2 PCTSINGLES + \beta 3 PCTBACHMOR + \beta 4 LNNBELPOV + \lambda W\epsilon + u$$

Spatial error regression includes the spatial error term of the lagged residuals as a predictor that replaces the epsilon term in OLS regression. The coefficient lambda, denoted by λ (lambda), is limited to a value between -1 and 1. If the term is significant, the closer the data set is to being spatially autocorrelated. The predictor $W\epsilon$ represents the weights matrix W applied to the residual term from the initial regression run on the data set. Taken together, lambda and the spatial lag of the residual term make up the spatially lagged residuals as a predictor for y (e.g. LNMEDHVAL). Through this process of filtering out the spatial information from the residual terms of y, we are left with some leftover value denoted by u which represents random noise in the data.

The Beta coefficient β_i of each predictor is interpreted in the same way as OLS regression – the amount by which the dependent variable changes as the independent variable increases by one unit, holding all other independent variables constant. The sign indicates whether the relationship between the dependent variable and the independent variables is positive (direct) or negative (inverse). It is important to look at the sign and value of the β_i when the coefficient is statistically significant and different from zero. The β_i is considered statistically significant and different from zero when the p-value falls below our alpha threshold of 0.05.

Assumptions and Goals

Apart from spatial independence of observations, the primary assumptions needed to run an OLS regression model are still needed for both spatial lag and spatial error regression models. To reiterate these assumptions, linearity, normality of residuals, no heteroscedasticity, and no multicollinearity are expected to be met to run these models. The goal of using a spatial lag or spatial error regression model is to account for potential spatial patterns in the data or residuals, which OLS regression is unable to account for. Using these methods can result in less heteroscedasticity for residuals or no spatial autocorrelation. We use these techniques to try to adjust for spatial patterns that may occur in geographically-influenced data sets.

Choosing a Spatial Regression Technique

To determine if a spatial regression model should be used, a Lagrange Multiplier (LM) diagnostic is provided as part of an initial OLS regression output. The decision of if and what model type to use is based on LM and Robust LM diagnostic probabilities and values that are provided for both lag and error, which are denoted in parentheses.

- If neither LM probability for lag and error are significant, we do not use spatial regression and keep with the results of the OLS regression.
- If the LM (lag) or LM (error) is significant and the other is not, we use the spatial regression model type that corresponds with the significant Lagrange Multiplier. If both LM (lag) and LM (error) are significant, the Robust LM values and probabilities are compared - the one with the lower p-value or higher test statistic is chosen.

Comparing Spatial Regressions to OLS Regression

Both spatial lag and spatial error models provide regression diagnostics that are similar to outputs of our OLS regression model. We obtain an R^2 value in our spatial regressions, but this value is considered a pseudo R^2 and is not directly comparable to an OLS R^2 - it does not have the same interpretation as the OLS R^2 value so we do not use it to compare between models. We also obtain a p-value for the Breusch-Pagan test for each spatial regression model. If $p < 0.05$, then heteroscedasticity is still present in our spatial regression model, which violates one of the assumptions of running the regressions.

To see which spatial regression model better accounts for spatial autocorrelation in our data set, we will compare the results of the spatial lag regression with the OLS regression and the results of the spatial error regression with the OLS regress to determine which performs better based on a number of different criteria.

These criteria include:

- Akaike Information Criterion and/or Schwarz Criterion;
- Log Likelihood;
- Likelihood Ratio Test

The Akaike Information Criterion (AIC) and Schwarz Criterion (SC) are two measures of goodness of fit for an estimated statistical model, which are relative measures of lost information when a model attempts to describe reality, which allow us to compare the quality of models to each other. While not useful as standalone measures, choosing one criterion and comparing the values for two or more models allows us to determine which model better fits the data – a lower value indicates a better fit. These criteria are usable to compare across all models used in this statistical analysis.

The log likelihood method uses the maximum likelihood method of determining the best fitting data given the parameters of a statistical model, where a higher value indicates a better fit. We are only able to use log likelihood to compare nested models – the OLS regression model is a nested version of the spatial lag and spatial error regression models that does not include the spatial lag or spatial error terms. We can compare our OLS model to each spatial regression separately, but we are unable to compare the spatial regressions to each other using the log likelihood criteria.

The likelihood ratio test conducts a hypothesis test to compare the OLS regression model with the spatial regression models. Like the log likelihood method, this test can only be used for nested models and so the OLS model must be separately compared to the spatial lag and spatial error models. We reject the null hypothesis if p is less than 0.05.

The null and alternate hypotheses of the likelihood ratio test are:

- H_0: the spatial lag (or spatial error) model is not a better specification than the OLS model
- H_a: the spatial lag (or spatial error) model is a better specification than the OLS model

Additionally, we can compare the results of our OLS model to the results of the spatial lag and spatial error models through the Moran's I of regression residuals. Moran's I allows us to examine the degree to which there is spatial autocorrelation present in the data of our models. If the Moran's I value is lower for the spatial regression model compared to the OLS model, then the spatial regression model better accounts for spatial autocorrelation in the data.

Geographically Weighted Regression

We use the open source software R to run a Geographic Weighted Regression (GWR) on the data set provided. We will compare the GWR model to the spatially lag, spatial error, and OLS model. AIC and Global Moran's I will be used to examine the model performance relative to the other models.

Overview of GWR Methods

GWR is based on a concept called Simpson's Paradox, which states that the pattern present in data when data is divided into thematic groups will be different than the pattern present in the entire dataset. Simpson's Paradox is an issue which is often present in spatial data - for instance the relationship between home value and home size might vary in different parts of a city.

Running a local regression instead of a global regression can help mitigate the challenges presented by Simpson's Paradox. Local regression methods involve running a separate regression for each location in a dataset. GWR is a type of local regression that we will apply to our data.

The equation for GWR equation used in our analysis is written for each observation as:

$$LNMEDHVAL_i = \beta_{i0} + \beta_{i1} PCTVACANT_i + \beta_{i2} PCTSINGLES_i + \beta_{i3} PCTBACHMOR_i + \beta_{i4} LNNBELPOV_i + \epsilon_i$$

The equation closely resembles the equation for an OLS regression.

β_{i0} is the value of the dependent value (y) (e.g. LNMEDHVAL) when all predictors (x_{i1} to x_{i4}) (e.g. PCTVACANT, PCTSINGLES, PCTBACHMOR, LNNBELPOV) are equal to 0. β_{i1} to β_{i4} represent the amount that the dependent variable changes by when the specified independent variable increases by one unit, holding all other independent variables constant. ϵ_i is the residual term or error. The subscript i is included in front of each term to indicate the regression equation is describing the relationship between y and the predictors around the location of i. The relationship is specific to location i.

To run a regression for each location, we need to have multiple observations. GWR uses data for neighboring observations in the regression. Neighboring observations are weighted according to their proximity to the location being analyzed and points which are closer to the location are given higher weights. In other words, observations which are closer to location i have a stronger influence on the coefficient estimates for that location i.

Fixed and Adaptive Bandwidth

Running GWR requires determining which of the locations around location i should be considered neighboring locations. There are two main approaches to define neighboring locations: fixed bandwidth and adaptive bandwidth.

When a fixed bandwidth kernel is used, the search distance (i.e: the search height) is kept constant. When a fixed bandwidth kernel is used, the number of neighboring observations around each location i will vary by location. As an example, a local regression on the price of homes sold in Philadelphia might consider all properties which are located within 500 feet of each home as part of the local regression. In this example, a fixed search distance of 500 feet is used. The number of homes neighboring each home is variable as the number of homes within 500 feet will be different for each home.

When an adaptive bandwidth kernel is used, the number of neighboring observations included in the regression is kept constant and the search distance will vary by location. Going back to the example of homes sold in Philadelphia, an example of an adaptive bandwidth would be to consider the nearest 20 homes to be neighbors. In this case, the search distance is variable because the distance to the 20th closest home will not be constant. However, the number of neighbors is constant and will always be equal to 20.

In an analysis, the assumptions about how to calculate weights and the type of bandwidth to use can have a major impact on results. A fixed bandwidth kernel is generally more appropriate when the distribution of data across space is relatively constant. An adaptive bandwidth kernel is generally more appropriate when the location of data is clustered in space or when you are working with polygons that have a non-uniform shape or size.

We will use an adaptive bandwidth in our GWR regression of median home sales by census block. An adaptive bandwidth is more appropriate because census block polygons do not have a uniform shape or size.

Assumptions of GWR

All the assumptions used in OLS regression also apply to GWR regression. Like OLS, GWR residuals should be normally distributed and there should be no severe multicollinearity in predictors. Additionally, the relationship between predicted values and the standardized residuals should be homoscedastic.

GWR is likely to be problematic when a regression analysis includes two or more predictors that have very similar spatial patterns in a region. Including two variables which have high and low values in the same locations will likely result in severe multicollinearity, which is a violation of the GWR assumptions. We can examine the condition number in the GWR results to determine if results are unstable for each local regression - a condition number is outputted for each local regression. As a general rule, the local regression results should not be trusted when the condition number is greater than 30, is null, or is equal to $-1.7976931348623158e + 308$.

GWR and p-values

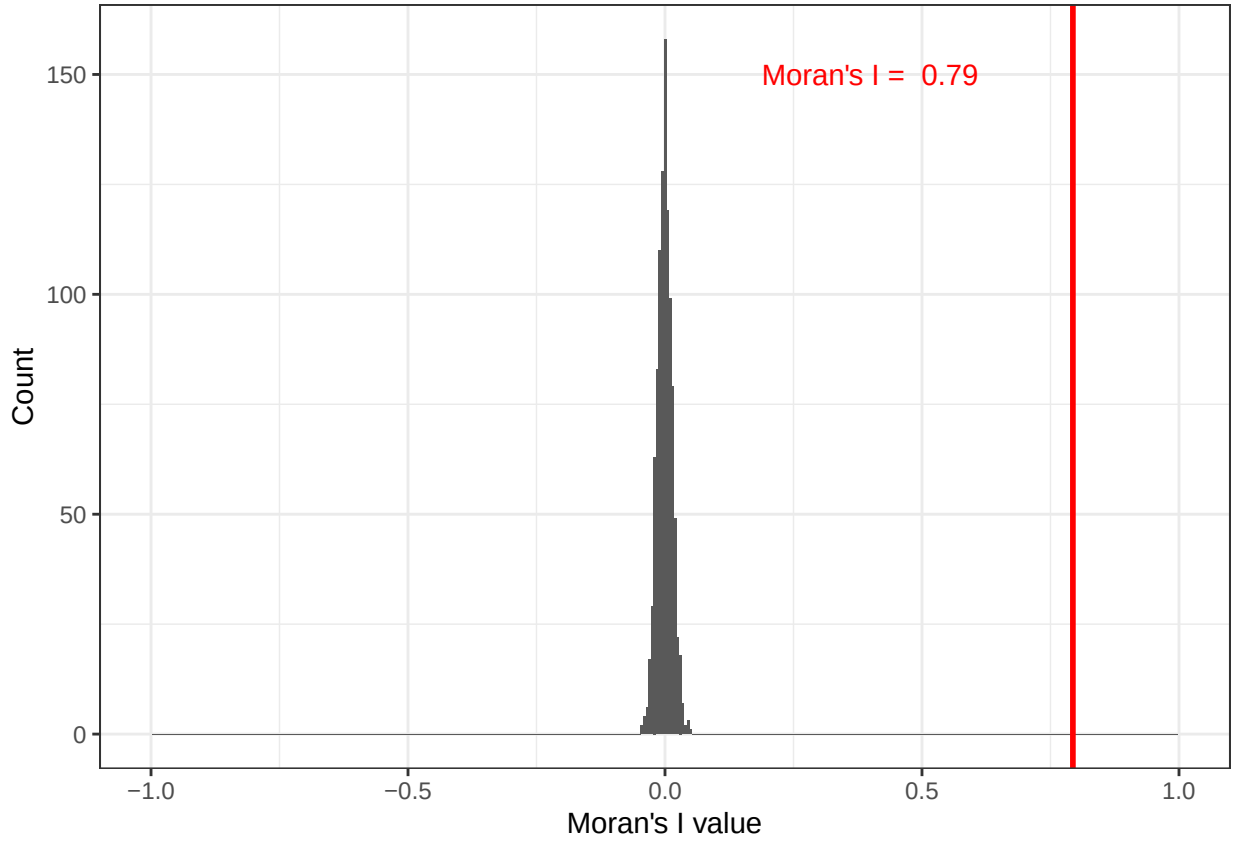
GWR does not provide estimates of statistical significance (i.e: p-values). This is because GWR includes a set of regression coefficients and errors for each local regression. As a result, there are potentially thousands of tests which are needed to identify if local parameters are significant. For a GWR regression with four predictors and 1000 regression points, there would be $5,000$ significance tests required, one for the intercept coefficient at each location and one for each of the four predictors at each location. In this situation, we would expect 250 of the $5,000$ tests to return a significant result by random chance if using a 0.05 significance level, which is problematic. For this reason, we will not consider p-values when reviewing our GWR regression results. There are methods to adjust for multiple testing, but these approaches are beyond the scope of our analysis.

Results

Spatial Autocorrelation

Global Moran's I

The global Moran's I value for the dependent variable LNMEDHVAL is calculated to be 0.79. After comparing this value with the Moran's I values obtained from 999 random permutations, it becomes evident that LNMEDHVAL exhibits a significantly high degree of spatial autocorrelation. This strongly suggests that the spatial distribution of LNMEDHVAL is not random but instead exhibits a spatial pattern. Therefore, we can confidently conclude that there is a substantial spatial autocorrelation present in LNMEDHVAL, leading us to reject the null hypothesis of no spatial autocorrelation. The histogram shows that the value of Moran's I for LNMEDHVAL of 0.79 is significantly higher that the values shown in the histogram of the Moran's I values generated from the 999 random permutations.



Local Moran's I

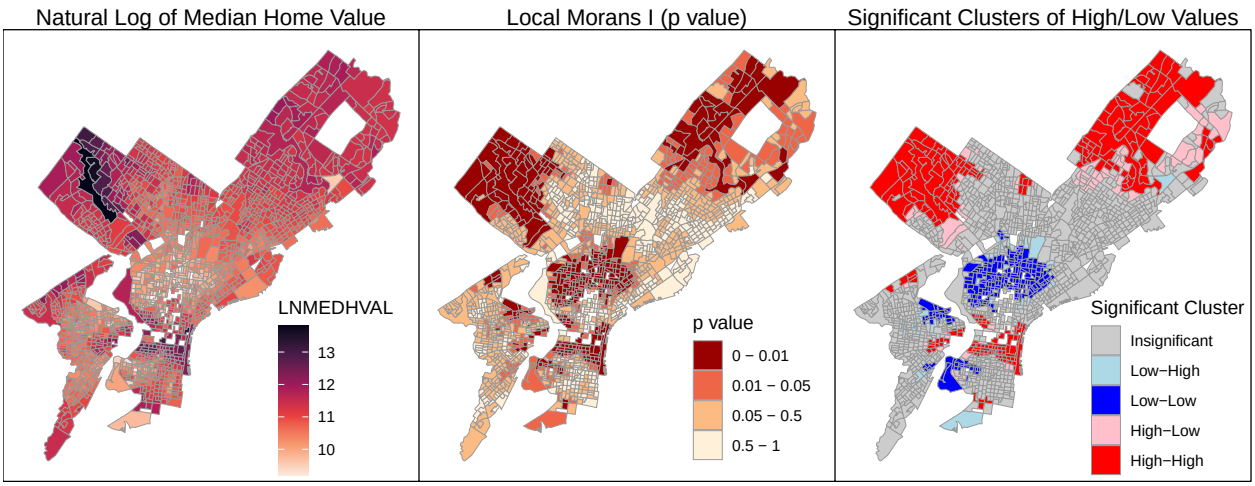
The Local Moran's I results are presented using maps of the local Moran's I p-values and a significant cluster map. These maps reveal distinct patterns within the spatial distribution of the variable of interest LNMEDHVAL. We consider clusters to be significant if the local Moran's I p value is less than 0.05.

The 'not significant' areas represent regions where no significant spatial clustering or outliers were identified, suggesting a relatively homogeneous spatial pattern.

In contrast, the 'high-high' clusters indicate areas where high values LNMEDHVAL are surrounded by other high values, signifying spatial clusters of high values. Conversely, the 'low-low' clusters represent areas characterized by low values surrounded by other low values, indicating spatial clusters of low values.

The ‘high-low’ areas depict locations where high values of the variable are surrounded by neighboring areas with low values, suggesting the presence of stark spatial outliers. Similarly, the ‘low-high’ areas denote regions where low values are surrounded by neighboring areas with high values, representing contrasting spatial outliers.

Based on the Local Moran’s I results, there are significant clusters of high-high values in Center City, Society Hill, and Rittenhouse Square, Roxborough, and parts of Northeast Philadelphia. Significant clusters of low-low values can be found in North Philadelphia, Kensington, Southwest Philadelphia. Areas with high-low clustering include parts of East Falls and localized areas of Northeast Philadelphia located along the Delaware river. Areas with low-high clustering include Navy Yard and Mill Creek. Not significant areas include most of South Philadelphia, Kensington, and West Philadelphia.



OLS Regression

OLS Results

The OLS regression results identified that poverty levels, education status, home vacancy rates, and single-resident homes are statistically significant predictors of the natural log of median home values in Philadelphia, with an R^2 value of 0.6623, explaining 66.23 of the variance in median home values.

```
##
## Call:
## lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
##     LNNBELPOV, data = data)
##
## 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  < 2e-16 ***
## PCTVACANT   -0.0191563  0.0009779 -19.590  < 2e-16 ***
## PCTSINGLES   0.0029770  0.0007032   4.234 2.42e-05 ***
## PCTBACHMOR   0.0209095  0.0005432  38.494  < 2e-16 ***
## LNNBELPOV   -0.0789035  0.0084567  -9.330  < 2e-16 ***
## ---
## 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: < 2.2e-16
```

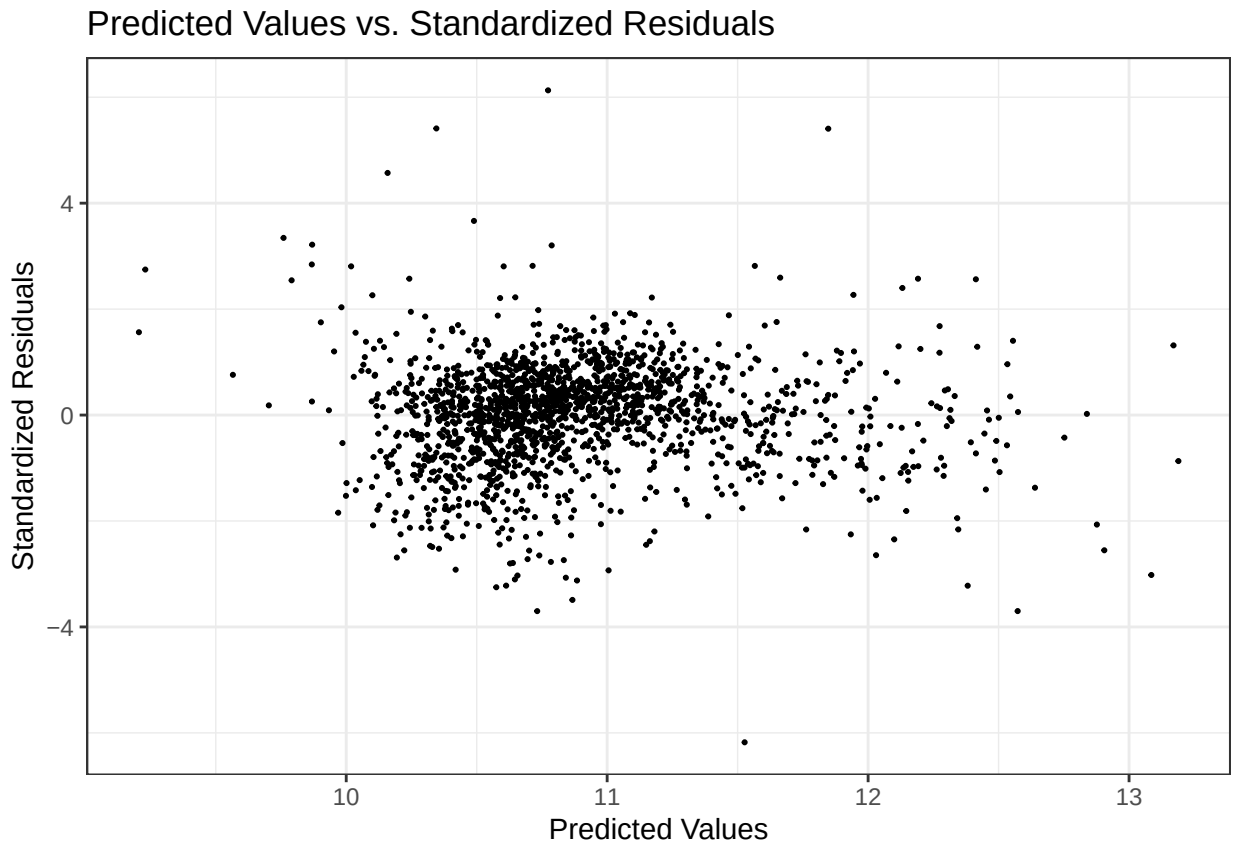
Heteroscedasticity tests

The results from the Breusch-Pagan, Koenker-Bassett, and White's tests for heteroscedasticity are consistent with one another, each providing strong evidence against the null hypothesis of homoscedasticity in the regression model. All of the test's p-values suggest that the variance of the residuals is not constant across observations, with p-values significantly below the common alpha level of 0.05, which in this case the p-values for the Breusch-Pagan and Koenker-Basset test are both near 0. The Koenker-Bassett test produces a high-test statistic of 43.94, which suggests strong evidence of heteroscedasticity. These very low p-values and high test statistic lead to the rejection of the null hypothesis of homoscedasticity of the model's residuals (i.e., having a constant variance).

```
##
##  Breusch-Pagan test
##
## data:  reg
## BP = 113.19, df = 4, p-value < 2.2e-16
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 42.868, df = 4, p-value = 1.102e-08
```

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

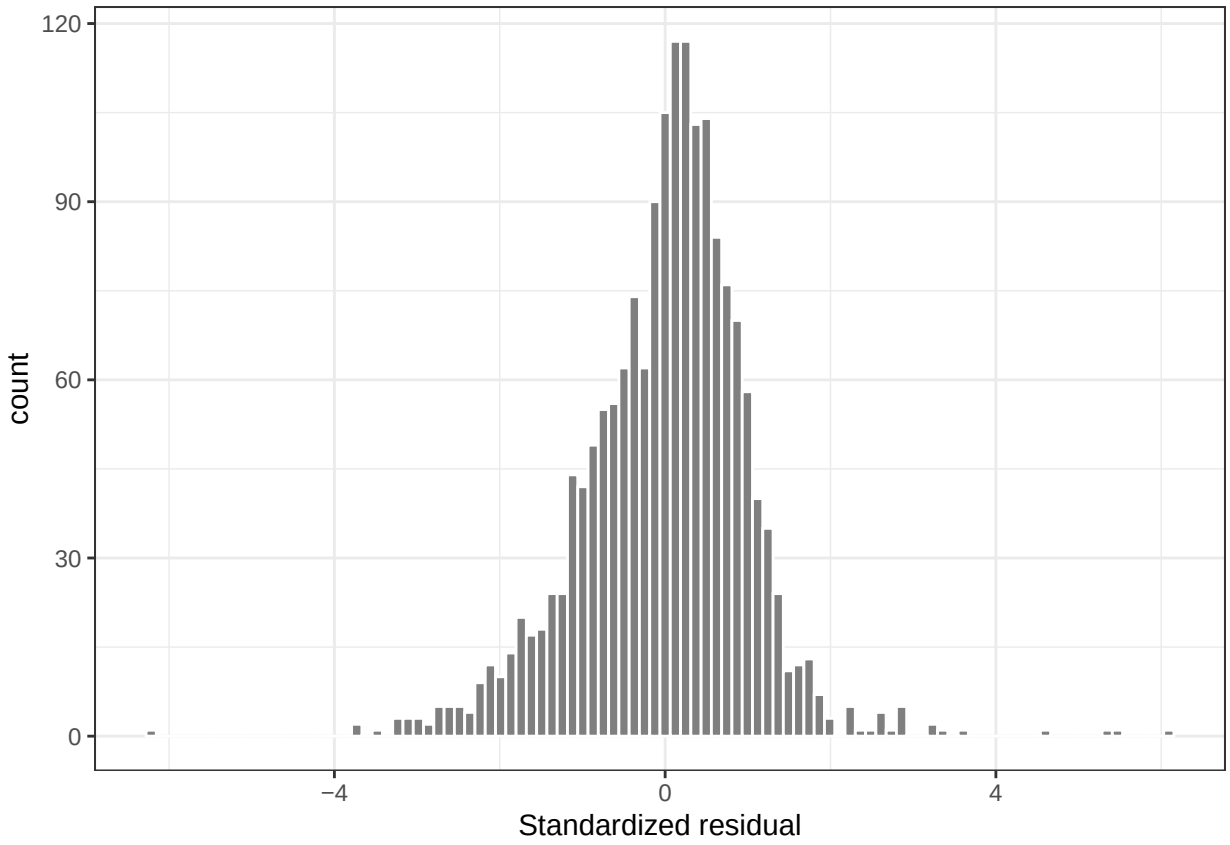


The conclusion from the Breusch-Pagan, Koenker-Bassett, and White's tests for heteroscedasticity is consistent with the observations made from the residual by predicted plot presented in to "Assignment 1 – OLS Regression". The "bow-tie" shape observed in the scatter plot around the predicted value of 11.5 is a visual indication of heteroscedasticity, where the spread of residuals increases at certain levels of the predicted variable, deviating from what would be expected in a homoscedastic relationship. This analysis complements the statistical evidence provided by these tests, which also suggests non-constant variance in the residuals.

Normality of errors (Jarque-Bera test)

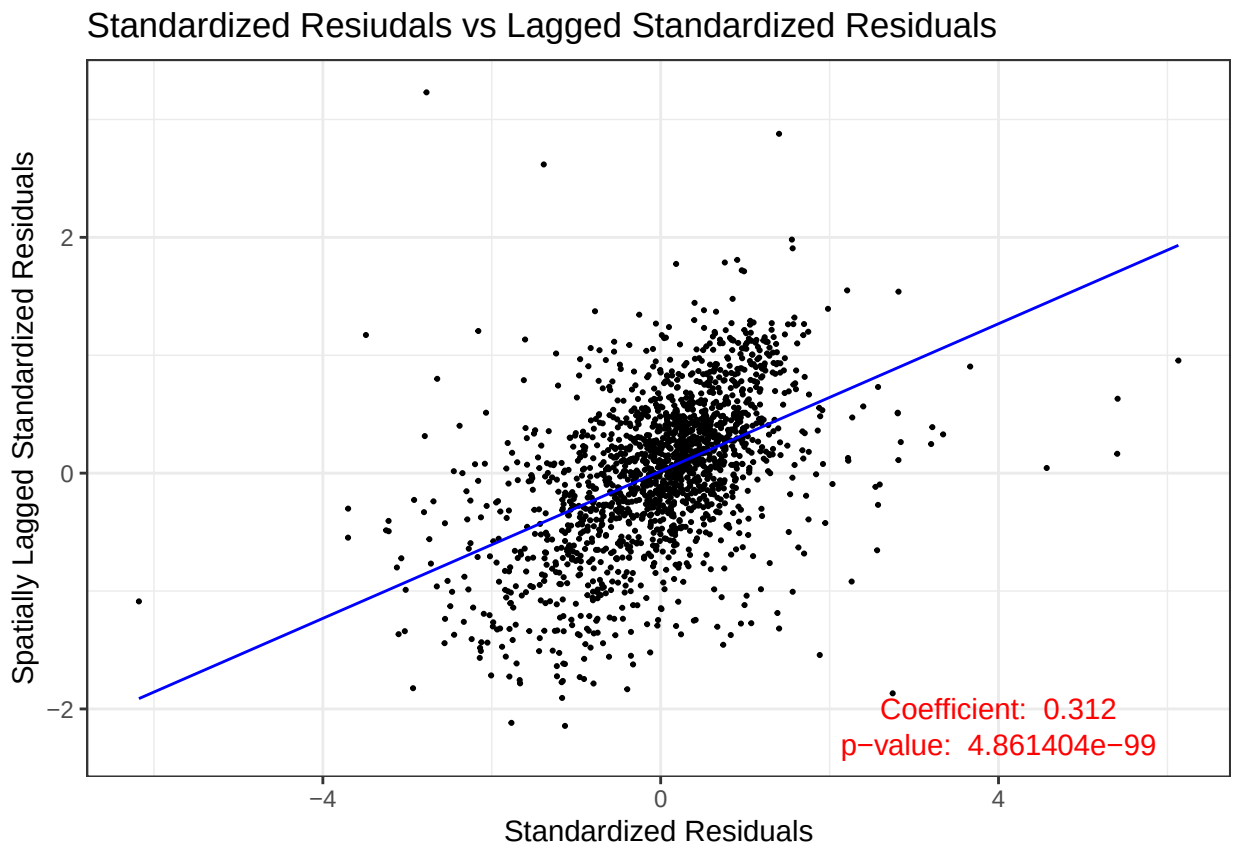
According to the Jarque-Bera test results, the assumption of normality for the regression errors is violated. The p-value for the Jarque-Bera test is close to 0 and is far below the conventional significance level ($p < 0.05$). This extremely low p-value leads to the rejection of the null hypothesis that the residuals are normally distributed. These results are not consistent with the histogram of residuals presented in "Assignment 1 – OLS Regression". This inconsistency can be attributed to the limits and accuracy of visual interpretation of histogram results. While the histogram of the residuals suggests a roughly normal distribution at a glance, the Jarque-Bera test is a nuanced tool that captures subtle deviations in skewness that are not immediately evident in the visual representation.

```
##
##  Jarque Bera Test
##
## data:  data$residual
## X-squared = 778.96, df = 2, p-value < 2.2e-16
```



Moran's I Plot

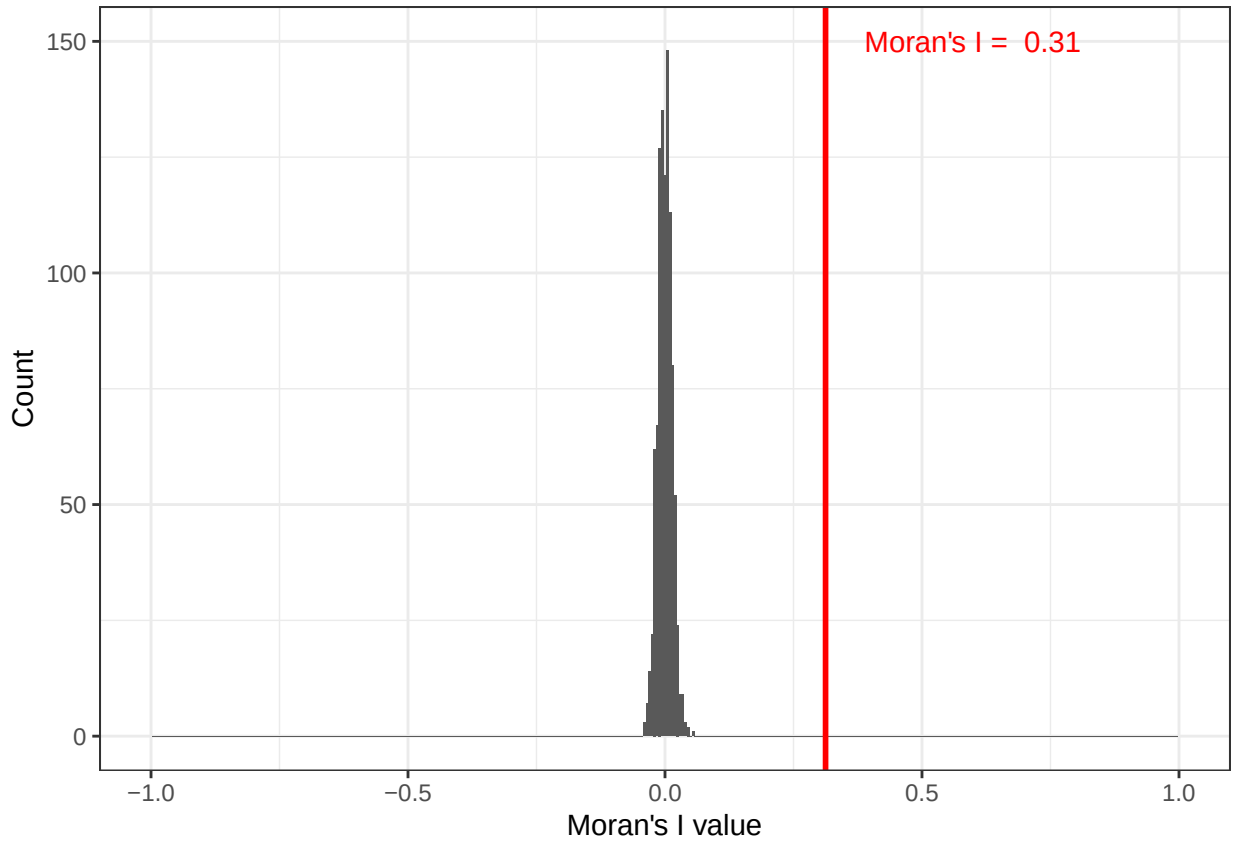
By regressing the standardized residuals on the spatially lagged residuals we can see that the beta coefficient (slope b) of the lagged residual is 0.312. As the lagged residual changes by 1 unit, the standardized residual changes by 0.312 units. This is a key indicator of spatial autocorrelation, which suggests that nearby or neighboring residuals are similar. When this coefficient is significantly different from zero, it means that there's a statistically significant spatial autocorrelation in the residuals of the model. In our analysis, the p-value for the beta coefficient (slope b) is close to zero which leads us to conclude that statistically significant positive correlation is present between the residuals and that spatial lag of the residuals. Thus, spatial autocorrelation is present in the OLS residuals.



Moran's I Permutation Results

Another way to check for spatial autocorrelation is by comparing the observed Moran's I value to randomly generated Moran's I value. The observed Moran's I value of 0.31, significantly distanced from the permutation distribution's center, revealing that substantial spatial autocorrelation is present in the OLS residuals. This is problematic because it violates the assumption that the residuals are independently distributed.

Both Moran's I and the beta coefficient of the spatially lagged residuals tell a similar story in that they both indicate the presence of spatial autocorrelation. The beta coefficient quantifies the relationship between the residuals and their spatial lags, while Moran's I provides an overall measure of spatial autocorrelation. The results of both statistics indicate that positive spatial autocorrelation is present and reinforce the conclusion that the OLS model's residuals are not independent and that there is a pattern to their spatial distribution.



Spatial Lag and Spatial Error Regressions

Spatial Lag Regression

The spatial lag regression model includes a separate line of results that provide the rho value of 0.65 and the p-value for this spatial lag term. Our p-value is close to 0, which means our spatial lag term is significant - as a result, we can say that median home values in an area are associated with nearby median home values.

Our other independent variables (LNNBELPOV, PCTBACHMOR, PCTSINGLES, PCTVACANT) all have probabilities near 0, meaning that they remain significant predictors for median home values in an area similar to the OLS model. The intercept value in the spatial lag model is nearly a third of the value in the OLS model, with the coefficients becoming slightly lower for LNNBELPOV, PCTBACHMOR, and PCTSINGLES and slightly higher for PCTVACANT.

The spatial lag model's Breusch-Pagan test results show a p-value near 0. Given that this p-value is less than 0.05, we conclude that the spatial lag regression residuals are still heteroscedastic in our model.

We also compare the AIC/SC, log likelihood, and likelihood ratio test values between our OLS regression model and spatial lag model to determine which model is a better fit for our data:

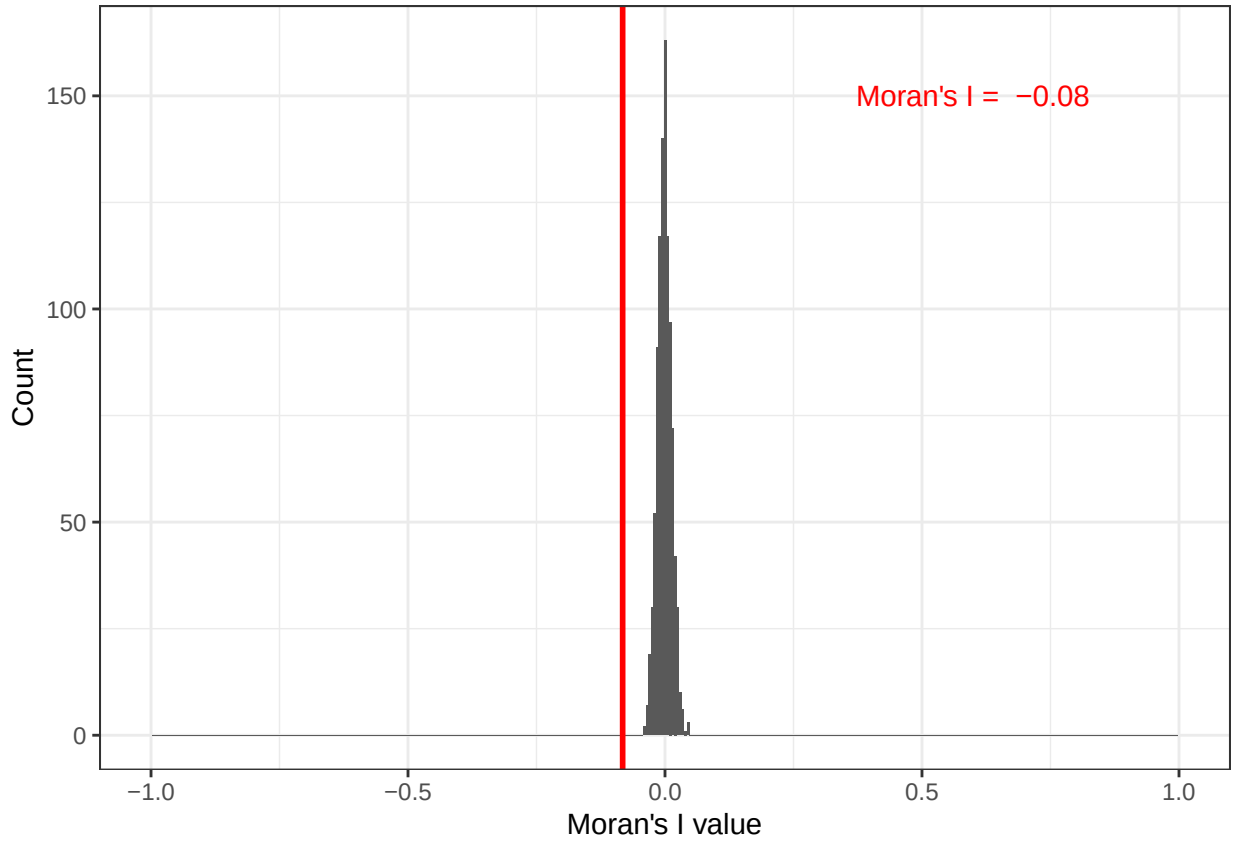
- The AIC value is 525 for the spatial lag model which is significantly lower than the AIC of 1435 for the OLS model, indicating that the spatial lag model is a better fit based on that metric since it is lower.
- The log likelihood value of the spatial lag model is -256 which is higher than the log likelihood value of the OLS model at -711, indicating that the spatial lag model is also a better fit based on that metric since it is higher.
- The p-value for the likelihood ratio test is near 0, so we reject the null hypothesis that the spatial lag model is not a better fit than the OLS model. As a result, the spatial lag model is a better fit for the data than the OLS model based on the likelihood ratio test.

```
##
## Call:lagsarlm(formula = LNMEDHVAL ~ LNNBELPOV + PCTBACHMOR + PCTSINGLES +
##     PCTVACANT, data = data, 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 < 2.2e-16
## LNNBELPOV   -0.03405466  0.00629287  -5.4116 6.246e-08
## PCTBACHMOR   0.00851381  0.00052193  16.3120 < 2.2e-16
## PCTSINGLES   0.00203342  0.00051577   3.9425 8.064e-05
## PCTVACANT   -0.00852940  0.00074367 -11.4694 < 2.2e-16
##
## Rho: 0.6511, LR test value: 911.51, p-value: < 2.22e-16
## Asymptotic standard error: 0.01805
##     z-value: 36.072, p-value: < 2.22e-16
## Wald statistic: 1301.2, p-value: < 2.22e-16
##
## 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: 2.2204e-16

##
##  Likelihood ratio for spatial linear models
##
## data:
## Likelihood ratio = 911.51, df = 1, p-value < 2.2e-16
## sample estimates:
## Log likelihood of lagreg    Log likelihood of reg
##                -255.7400                -711.4933

##
##  Breusch-Pagan test
##
## data:
## BP = 210.76, df = 4, p-value < 2.2e-16

##
##  studentized Breusch-Pagan test
##
## data:
## BP = 51.411, df = 4, p-value = 1.832e-10

##
##  Jarque Bera Test
```

```
##
## data:  lagreg$residuals
## X-squared = 2756.9, df = 2, p-value < 2.2e-16
```

Spatial Lag Moran's I Results The Moran's I value of the spatial lag model is -0.08. The negative Moran's I value and the p-value of 0.002 lead us to conclude that negative spatial autocorrelation is present in the spatial lag model residuals. However, the p-value of -0.08 is still a large improvement over the Moran's I value of the OLS model which was 0.31. The spatial lag model better accounts for spatial autocorrelation in the data and shows less spatial autocorrelation in the residuals compared to the OLS model.

Our comparison criteria show that the spatial lag model appears to better account for spatial autocorrelation in our data compared to the OLS model. The spatial lag model has a lower AIC value, higher log likelihood value, statistically significant likelihood ratio test, and lower Moran's I value. However, the Breusch-Pagan test shows that there is still heteroscedasticity present in the model's residuals.



Spatial Error Regression

The spatial error regression model includes a separate line of results that provide the lambda value of 0.81 and the p-value for this spatial error term. Our p-value is close to 0, which means our spatial error term is significant - as a result, we can say that the residuals of median home values in an area are associated with the residuals of nearby median home values.

Our other independent variables (LNNBELPOV, PCTBACHMOR, PCTSINGLES, PCTVACANT) all have probabilities near 0, meaning that they remain significant predictors for median home values in an area similar to the OLS model. The intercept value in the spatial lag model is similar to the value in the OLS model, with the coefficients becoming slightly lower for PCTBACHMOR and PCTSINGLES and slightly higher for LNNBELPOV and PCTVACANT.

The spatial error model's Breusch-Pagan test results show a p-value of 0.0001. Given that this p-value is less than 0.05, we conclude that the spatial error regression residuals are still heteroscedastic in our model.

We also compare the AIC/SC, log likelihood, and likelihood ratio test values between our OLS regression model and spatial error model to determine which model is a better fit for our data:

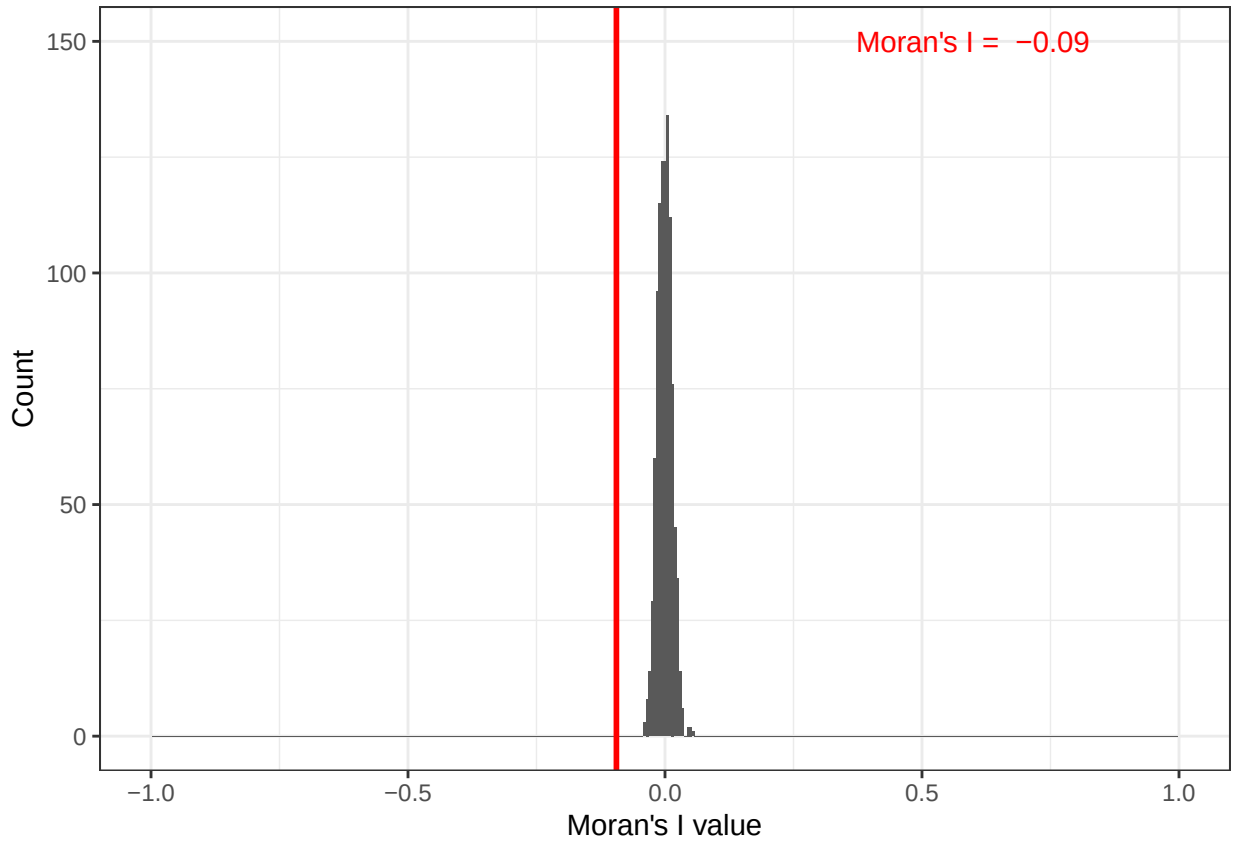
- The AIC value is 755 for the spatial lag model which is significantly lower than the AIC of 1435 for the OLS model, indicating that the spatial lag model is a better fit based on that metric since it is lower.
- The log likelihood value of the spatial error model is -3734 which is higher than the log likelihood value of the OLS model at -711, indicating that the spatial error model is also a better fit based on that metric since it is higher.
- The p-value for the likelihood ratio test is near 0, so we reject the null hypothesis that the spatial error model is not a better fit than the OLS model. As a result, the spatial error model is a better fit for the data than the OLS model based on the likelihood ratio test.

```
##
## Call:errorsarlm(formula = LNMEDHVAL ~ LNNBELPOV + PCTBACHMOR + PCTSINGLES +
##     PCTVACANT, data = data, listw = queenlist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.926477 -0.115408  0.014889  0.133852  1.948664
##
## Type: error
## Coefficients: (asymptotic standard errors)
##                Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept) 10.90643423  0.05346779 203.9814 < 2.2e-16
## LNNBELPOV   -0.03453408  0.00708933  -4.8713 1.109e-06
## PCTBACHMOR   0.00981293  0.00072896  13.4615 < 2.2e-16
## PCTSINGLES   0.00267792  0.00062083   4.3134 1.607e-05
## PCTVACANT   -0.00578308  0.00088670  -6.5220 6.937e-11
##
## Lambda: 0.81492, LR test value: 677.61, p-value: < 2.22e-16
## Asymptotic standard error: 0.016373
##     z-value: 49.772, p-value: < 2.22e-16
## Wald statistic: 2477.2, p-value: < 2.22e-16
##
## 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: NA (not available for weighted model), (AIC for lm: 1435)

##
##  Likelihood ratio for spatial linear models
##
## data:
## Likelihood ratio = 677.61, df = 1, p-value < 2.2e-16
## sample estimates:
## Log likelihood of errreg    Log likelihood of reg
##                -372.6904                -711.4933

##
##  Breusch-Pagan test
```

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

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

##
##  Jarque Bera Test
##
## data:  errreg$residuals
## X-squared = 3507, df = 2, p-value < 2.2e-16
```

Spatial Error Moran's I Results The Moran's I value of the spatial error model is -0.09. The negative Moran's I value and the p-value which is again lower than 0.05 lead us to conclude that negative spatial autocorrelation is present in the spatial error model residuals. However, the Moran's I value is an improvement over the Moran's I value of the OLS model which was 0.31. The spatial error model better accounts for spatial autocorrelation in the data and shows less spatial autocorrelation in the residuals compared to the OLS model.

Our comparison criteria show that the spatial error model appears to better account for spatial autocorrelation in our data compared to the OLS model. The spatial error model has a lower AIC value, higher log likelihood value, statistically significant likelihood ratio test, and lower Moran's I value. However, the Breusch-Pagan test shows that there is still heteroscedasticity present in the model's residuals.



Comparison Between Spatial Regression Models

We can compare the results of our spatial regression models based on the values of the AIC/SC results. The spatial lag regression's AIC value of 525 is lower than the spatial error regression's AIC value of 755, indicating that the the spatial lag model is a better spatial regression model for our data.

Geographically Weighted Regression

Results

The output shows the results of the GWR analysis in which the median home value by block group is regressed against our four predictors. We examine the quasi-global R^2 which is a measure of the goodness of fit of the GWR model. The quasi-global R^2 value is 0.847 for the GWR analysis, the R^2 value for the OLS regression analysis is 0.6623. The GWR regression has a higher R^2 and appears to be better at explaining the variance in the natural log of the median home sale value by block group.

```
## Call:
## gwr(formula = LNMEDHVAL ~ LNNBELPOV + PCTBACHMOR + PCTSINGLES +
##     PCTVACANT, data = datas, 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
## LNNBELPOV    -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
## PCTSINGLES   -0.0249706 -0.0075550 -0.0016626  0.0042280  0.0143340  0.0030
## PCTVACANT    -0.0317407 -0.0142383 -0.0089599 -0.0035770  0.0167916 -0.0192
## 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
```

Akaike Information Criteria (AIC)

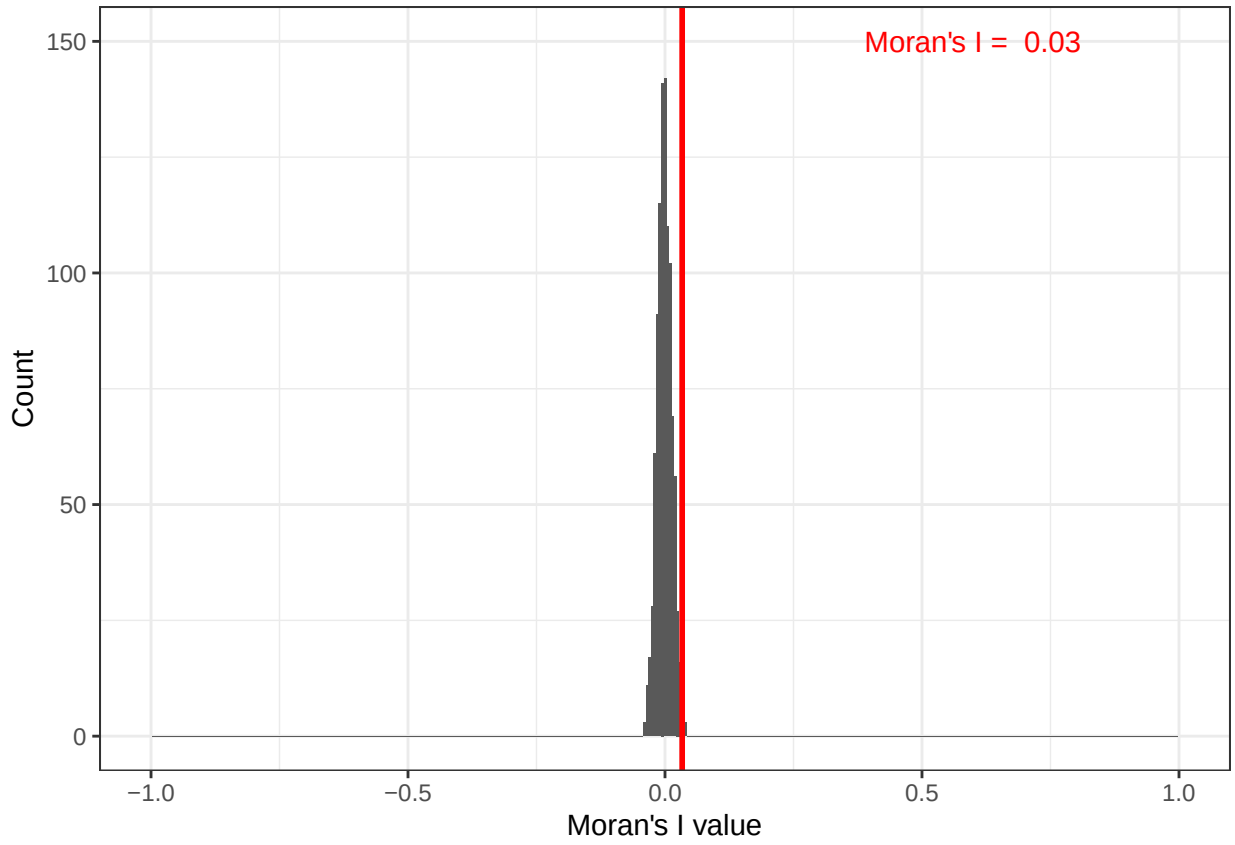
The table below shows the AIC values for each of our four models. Models with a lower AIC value have a better fit. Based on the AIC values, we can conclude that the GWR model seems to be doing the best job predicting the natural log of the median home value.

Model	AIC
GWR	308.7123
OLS	1434.9867
Spatial Lag	525.4800
Spatial Error	754.9850

Geographically Weighted Regression Moran's I Results

The plot below shows the global Moran's I value for the residuals of the GWR. The global Moran's I value of the GWR residuals is 0.03. By comparing the Moran's I value to the histogram of random permutations, we can conclude that there is clustering in the GWR residuals. However, there is a slight possibility that this clustering is the result of random chance as a small number of the random Moran's I permutations have Moran's I values which are higher than 0.03. The p-value for the Moran's I statistic of the GWR residuals is 0.014, indicating that we are only 97.2 confident the clustering is not the result of random chance and there is a very small possibility the clustering of the GWR residuals could be random.

When compared to other models, the GWR residuals are less clustered than the OLS residuals. The Spatial Lag and Spatial Error residuals have negative Moran's I value indicating spatial dispersion (i.e: negative spatial autocorrelation). Because the absolute value of the Moran's I statistic for the GWR regression is lower than the other three models, we can conclude that the residuals of the GWR regression have the least amount of spatial autocorrelation.

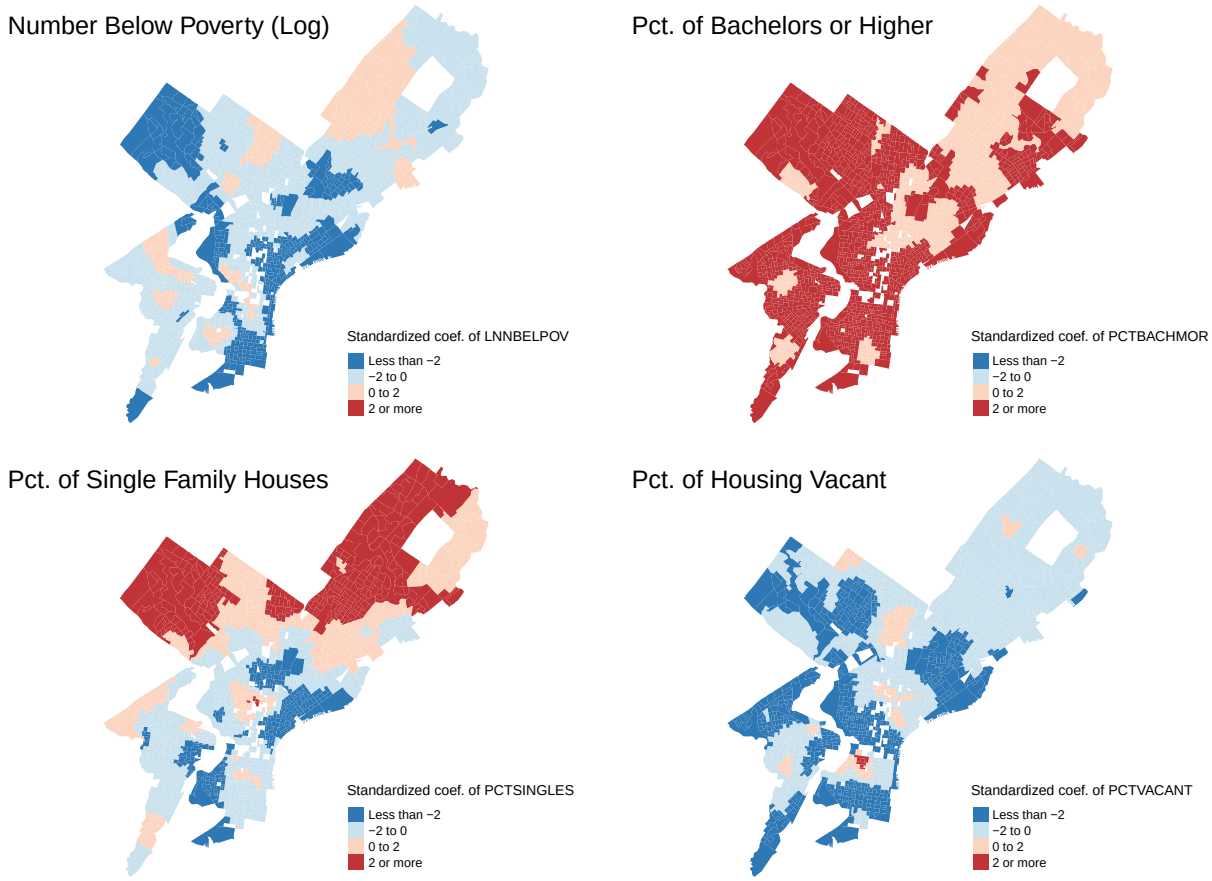


GWR Local Regression Results

The maps below show the beta coefficient divided by the standard error for each of our four dependent variables. This figure is referred to as a standardized coefficient.

If the standardized coefficient is greater than two, we can conclude that there is possibly a significant local positive relationship between the predictor and the natural log of the median home sales value. If the coefficient is less than negative two, we can conclude that there is possibly a significant negative relationship between the predictor and the natural log of the median home sale value.

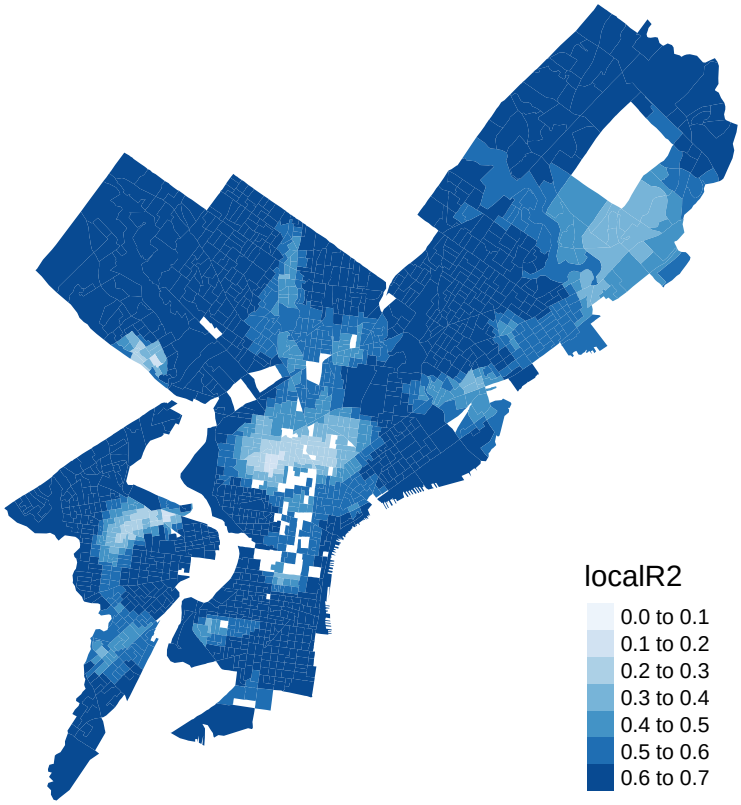
From the maps below, we can conclude that there is a possibly significant positive relationship between the percent of households with a bachelor's degree and the natural log of median home sale value in most of the city. There is also possibly a statistically significant negative relationship between the percent of houses which are vacant and the natural log of the median home sale value in many parts of Philadelphia including areas of Northwest Philadelphia, South Philadelphia, and parts of West Philadelphia. There is a possibly statistically significant positive relationship between the percent of single family houses and the natural log of the median home values in large sections of Northeast and Northwest Philadelphia. This possibly significant positive relationship is not present in other parts of Philadelphia, where the number of single family homes tends to be smaller.



Local R2 Map

The map below shows the local R^2 values for the GWR regression. In areas where the R^2 is higher, our local regression model explains more of the variation in median home sales values. The Local R^2 for most block groups in South Philadelphia, Northwest Philadelphia, and Northeast Philadelphia is greater than 0.5, indicating our model does a good job explaining the variance of the dependent variable. However, there are localized areas of North Philadelphia

and West Philadelphia where the R^2 is less than 0.2, indicating that the model does not do a good job explaining the variance in the median home sale value.



Discussion

Through this analysis, we built upon our previous multiple regression analysis of median home values in Philadelphia through spatial regression techniques to account for spatial autocorrelation in the data. This report explores multiple spatial regression techniques, including spatial lag regression, spatial error regression, and geographically weighted regression, to examine which technique best fits our Philadelphia data set. We can conclude that spatial regression techniques do provide a better model fit for our data compared to OLS regression based on the comparisons of different model metrics. The GWR model had the lowest AIC value and a Moran's I value that was closest to 0, which means the model was the most effective at accounting for spatial autocorrelation out of the four regression techniques presented in this analysis.

The spatial lag and spatial error methods have statistically significant values for their Breusch Pagan tests, which indicate that the residuals still show the heteroscedasticity that the OLS model does. Having homoscedastic residuals remains an important assumption for running these models, so these methods do not account for this assumption.

While GWR does have a lowest AIC, the model performance is mixed and the predictors we use do a good job accounting for the variation in mean household value in some neighborhoods but not others. Notably, the GWR model appears to do a very poor job explaining the variation in median home sale prices in areas with a large minority population like North Philadelphia and West Philadelphia. Additionally, local GWR regressions predictors are likely to display some degree of local multicollinearity. For example, the proportion of residents in a block group with at least a bachelor's degree and the proportion of homes which are single family homes are likely to be collinear in parts of Northwest Philadelphia even though they are not collinear at a city wide scale.

Spatially weighted residuals (i.e., spatially lagged) are the residuals from a regression model that have been adjusted for spatial autocorrelation. These residuals are weighted based on a spatial weights matrix (Queen's neighbor). Spatially lagged residuals show how observations may be related to neighboring observations. In contrast a spatial lag model directly incorporates spatial dependence into regression through a spatially lagged dependent variable. Meaning, that the dependent variable for one observation is explained not only by its own predictors but also by the values of the dependent variable for its neighbors. The residuals from a spatial lag model are the differences between the observed values of the dependent variable and the values predicted by the model, which includes this spatially lagged dependent variable. Spatial lag model residuals are expected to show less spatial autocorrelation than ordinary least squares residuals because the model itself accounts for spatial dependence.

ArcGIS Pro, an industry standard for GIS and spatial analysis, is problematic to use for GWR models because it uses unclear methodologies to run the regression and returns either unhelpful or nonsensical results. Instead of an AIC value, ArcGIS Pro only returns an AIC corrected (AICc) value, which cannot be used to compare with AIC outputs for other regression methods. ArcGIS Pro optimizes bandwidth for the model using a confusing methodology called Golden search, and can return outputs that include negative R^2 values, which should be instead bounded between zero and one. Other tools, such as R, are a more consistent and appropriate choice to use when developing a GWR model.