

Gabriel Hernandez, Timothy Oliver, Ziyi Tang

Professor Eugene Brusilovskiy

MUSA 5000/CPLN 6710

19 October 2023

Assignment 2: Spatial Lag, Spatial Error, & Geographically Weighted Regression

1. Introduction

Market research was professionalized in the 1920s and has grown into a ubiquitous and targeted effort which includes demographic analysis focusing on an area of study (Booker). This extends to the simple examination of buying preferences for certain goods and at other times more complex relationships like those between housing values and neighborhood characteristics. The housing market is based on the agglomeration of individual decisions that take into account some form of analysis. The success of real-estate planning and development requires an understanding of the community and the local built environment; however, understanding the spatial relationship of nearby market prices greatly aids understanding of the price under review (Anselin). As in our previous report, we will firstly examine and discuss the relationship between median house values and characteristics of Philadelphia's neighborhoods through the use of the conventional tool of Ordinary Least Squares (OLS) regression. This approach often stumbles when confronted with datasets imbued with a significant spatial component. The value of one observation can be significantly influenced by its neighbors, leading to spatial autocorrelation that OLS fails to account for.

In light of these insights, the current report aims to pivot towards more spatially attuned methodologies. We explore the realms of Spatial Lag, Spatial Error, and Geographically Weighted Regression (GWR). These methods promise a more nuanced understanding of spatial autocorrelation, a phenomenon frequently overlooked yet crucial in urban socioeconomic contexts.

2. Methods

In the preceding report, we employed Ordinary Least Squares (OLS) regression to scrutinize the interplay between median home values (our dependent variable) and a suite of predictors. These predictors included the proportion of residents with a bachelor's degree or higher (PCTBACHMOR), the vacancy rate (PCTVACANT), the percentage of single-family houses (PCTSINGLES), and the number of households below the poverty line (NBELPOV100). While OLS regression is an important analytical tool, it often proves inadequate when spatial processes influence the values observed in the dataset. The inherent assumption of OLS—that observations are independent—is frequently violated in spatial datasets where the value of a given observation might be influenced by the value of its neighbors. This shortcoming was pointed out in the conclusion of our previous report: when we account for spatial relationships we can decrease the variance in our model.

This report pivots to alternative methodologies that explicitly account for spatial autocorrelation in the dependent variable as shown in the residuals of our previous OLS model. The inclusion of Spatial Lag and Spatial Error models in our analysis addresses spatial autocorrelation directly, decreasing the influence of unseen spatial processes on our data. The Spatial Lag model incorporates a neighboring effect on the dependent variable as a predictor, and the Spatial Error model adjusts for correlation within the model's residuals.

Moreover, Geographically Weighted Regression (GWR) allows for a nuanced, location-specific, analysis that provides a more-accurate reflection of the spatial heterogeneity inherent in urban socioeconomic patterns. In short, GWR accounts for spatial non-stationarity of the observed data and its relationship to predictors in different contexts.

Overall, this report endeavors to make a more compelling case on the relationships between our predictors and median house values in Philadelphia. We will engineer new features, such as our lagged dependent variable, to have a model that accounts for spatial processes. It will then make a more compelling explanation of the factors affecting median home values in Philadelphia compared to the traditional OLS approach used in our first report.

2.1 A Description of the Concept of Spatial Autocorrelation

We retrieved the original Philadelphia block-group level dataset used in our report from the American Census. It contained 1,816 total observations. After removing those block groups with a population under 40, 0 housing units, and median house values below \$10K. We removed one outlier block group in North Philadelphia that had both a disproportionately high median house value (>\$800,000) and a low median household income (<\$8,000). The final, cleaned, dataset contains 1,720 observations. In addition to the spatial polygon feature of each block group in our analysis, our other features/predictors/variables include:

- 1) PCTBACHMOR - the proportion of residents in the block group with a bachelor's degree or higher education
- 2) PCTVACANT - the proportion of housing units that are vacant
- 3) PCTSINGLES - the percent of housing units that are detached single family houses
- 4) NBELPOV100 - the number of households with incomes below 100% of the poverty level—denoting living in poverty

- 5) MEDHHINC - the median household income
- 6) MEDHVAL - the median value of all owner-occupied housing units in the block group and the dependent variable for the regression models used.
- 7) Created features (SPatial Lag, lagged errors) used in regressions Mention in another section; this is the original data set. can explain autocorrelation models in a following paragraph that states the name of the feature and its interpretation

Further description of the dataset is value-specific and discovered through exploratory data analysis.

2.1.1 Moran's I and 1st Law of Geography

The very premise of spatial autocorrelation lies in the observation that near things are more similar to one another than to things farther away. This is Tobler's First Law of Geography. Therefore, we begin by defining neighbors for each of the block groups in Philadelphia. In this study, we will be using Queen Neighbors.

Moran's I is a measure of spatial autocorrelation, quantifying the degree to which a variable is similar to itself in nearby locations. The formula for Moran's I is:

$$I = \frac{n}{W} \times \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Where:

- n is the number of spatial units indexed by i and j ,
- W is the sum of all spatial weights w_{ij} ,
- x_i and x_j are the values of the variable at locations i and j ,
- \bar{x} is the mean of the variable.

Each term in the formula represents a key component in understanding spatial relationships, with the spatial weights w_{ij} playing a pivotal role.

The weight matrix, denoted as W , represents the spatial structure of the data, with w_{ij} signifying the spatial weight between units i and j . The choice of this matrix is crucial as it influences the measurement of spatial autocorrelation. Throughout this report, this specific weight matrix will be used consistently. However, statisticians often utilize more than one spatial weight matrix in analyses to explore different aspects of spatial relationships. This approach helps in understanding the robustness of the results against different spatial configurations.

2.1.2 Testing the Significance of Moran's I


To test whether the observed spatial autocorrelation is significant, we formulate two hypotheses:

- Null Hypothesis (H_0): There is no spatial autocorrelation.
- Alternative Hypothesis (H_1): There is significant spatial autocorrelation.

The significance of Moran's I is assessed using a random permutation process (999 in our work). This involves randomly rearranging the spatial units and recalculating Moran's I for these permutations to create a reference distribution. The observed Moran's I is then compared to this distribution to assess its significance.

2.1.3 Local Spatial Autocorrelation

Local spatial autocorrelation focuses on the correlation within localized areas, allowing for the identification of spatial clusters or outliers. The local spatial autocorrelation is similar to Moran's I, involving permutation tests that compare local statistics against a

random distribution to determine significant spatial clusters or outliers. This aspect is crucial for detailed spatial analysis, providing insights into the localized patterns and anomalies in the spatial data. 

2.2 A Review of OLS Regression and Assumptions

2.2.1 Overview of OLS Regression

Ordinary Least Squares (OLS) regression is a fundamental statistical method used for estimating the relationships between a dependent variable and one or more independent variables. The use of multiple linear regression requires fulfillment of five assumptions.


- **Linearity:** The relationship between independent and dependent variables should be linear.
- **Independence of Observations:** Each observation in the dataset is independent of other observations.
- **Normality of Residuals:** The residuals (or errors) of the regression should be normally distributed.
- **Homoscedasticity:** The variance of the errors should remain consistent across all levels of the independent variables.
- **No Multicollinearity:** Independent variables should not be too highly correlated with each other. Here, a correlation r value at or above the 0.8 threshold is considered highly correlated.

2.2.2 Data with Spatial Component

When dealing with spatial data, the assumption that errors are random and independent often does not hold. This non-independence of errors can significantly impact the validity of OLS regression results.

One way to test the independence of errors is by examining the spatial autocorrelation of the residuals using Moran's I. This test will indicate if there is a pattern in the residuals that could violate the OLS assumptions.

Another method involves regressing OLS residuals on nearby residuals. This is particularly relevant when dealing with spatial data like census tracts and block groups, as defined by the Queen matrix. The slope (b) at the bottom of the scatterplot between OLS_RESIDU and WT_RESIDU provides insight into this relationship. This slope is calculated through a regression analysis between the residuals of a location and the weighted average residuals of its neighbors.

In this study, we used GeoDa to run the OLS regression. For heteroscedasticity tests, we used Breusch-Pagan to examine  data. The null hypothesis states that errors exhibit constant variance (homoscedasticity), while the alternative hypothesis indicates the presence of heteroscedasticity (variable error variance across observations).

For normality of errors tests, we chose Jarque-Bera tests in GeoDa. The null hypothesis asserts that the errors are normally distributed, whereas the alternative hypothesis suggests a deviation from normality.

2.2.3 Summary Statistics and Distribution Analysis

We generated histograms (figure 1) to visualize the distributions of the data's variables in order to discern whether the distribution of each approximates a normal distribution. We explore this because the Ordinary Least Squares (OLS) regression contains an underlying assumption of normally distributed residuals, and viewing a normal

distribution in values for each variable can quickly affirm such. Given the potential pitfalls of non-normality in linear regression, we will also examine if a logarithmic transformation of any of the variables results in a more normal distribution, which will improve our analysis if we are to incorporate the variable. While this might normally be done for only those variables observed to be not normally distributed, we transformed all of our features except our dependent variables. This resulted in the following additional variables and observed distributions (figure 2):

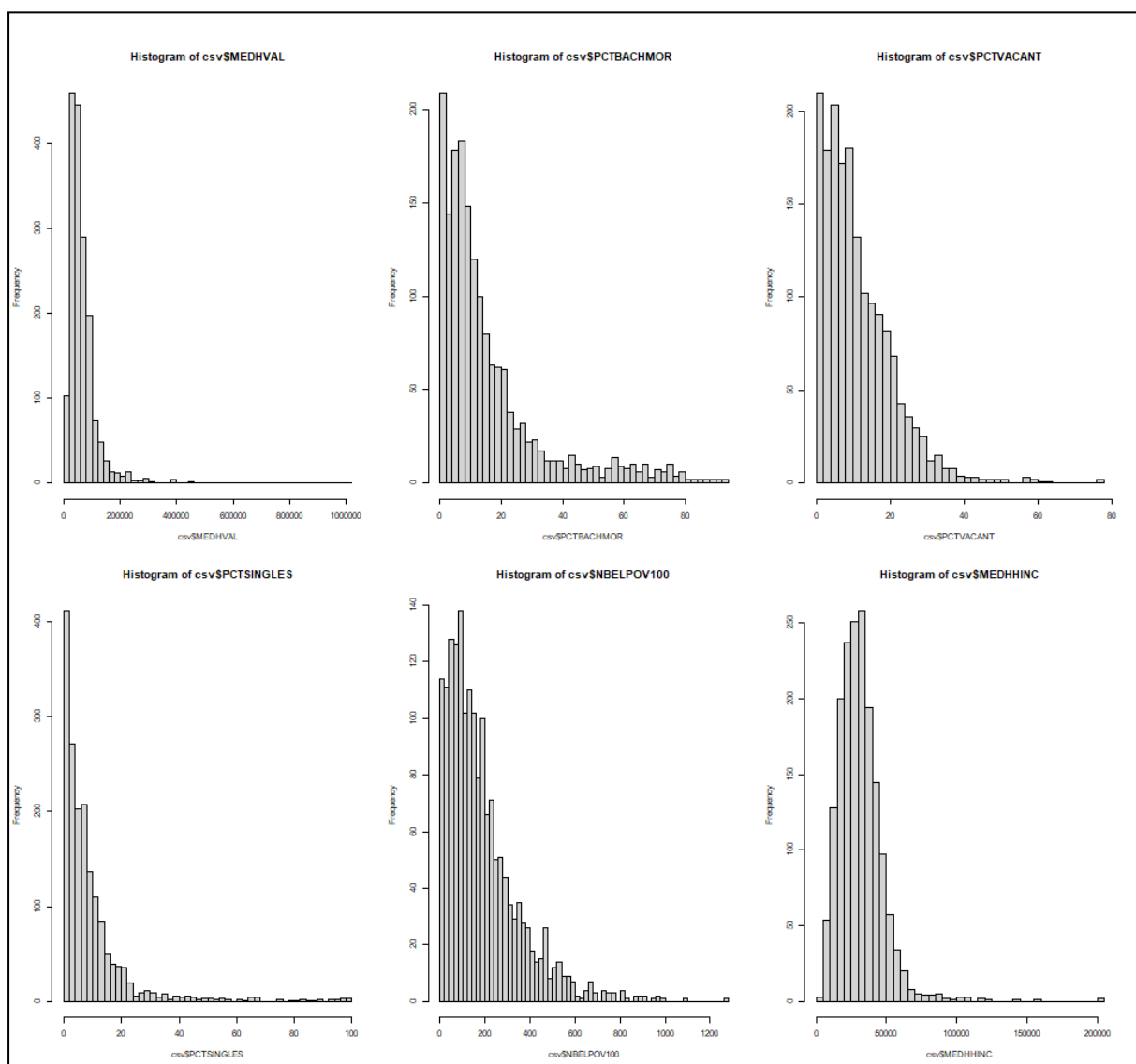


Figure 1: *Histograms of variables as-given*

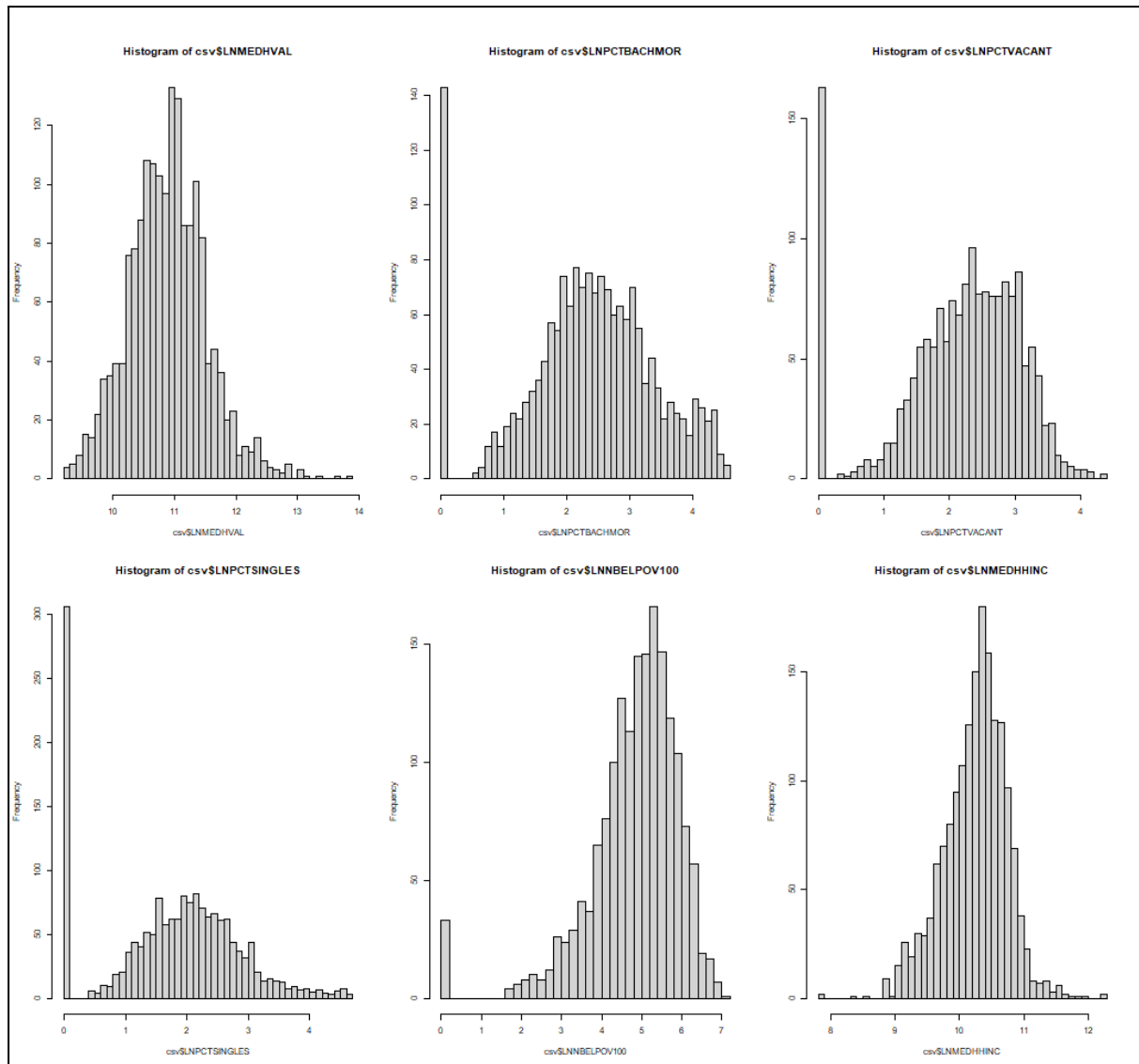


Figure 2: *Histograms of variables Transformed*

2.2.4 Correlation Analysis

Next, we evaluate the pairwise correlations amongst the predictors. Understanding the degree to which our predictors are correlated gives us insight into potential multicollinearity, which impacts the reliability of our regression coefficients. Correlation quantifies the linear relationship between two quantitative variables. A positive correlation indicates that as our predictor increases, so too does our dependent variable and vice versa for a negative correlation.

The degree of linear association is quantified using the sample correlation coefficient that is denoted by r and represented through the equation:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Where:

X_i and Y_i are individual data points.

\bar{X} and \bar{Y} represent the mean values of X and Y respectively.

The value of r can lie between -1 and 1. An r value of -1 signifies a perfect negative linear relationship where one data point changes inversely to another often at an equal or multiple rate. Similarly, a value of 1 denotes a perfect positive linear relationship with the change in one variable being an increase as the other increases also. Importantly, an r value of 0 implies no linear correlation between the variables, meaning changes in one variable do not predict changes in the other in any consistent direction.

These correlations in each of the predictor variables are found using the `cor` command in R to compute Pearson correlations. The correlation between each predictor and the dependent variable, LNMEDHVAL, is observed in the resulting r value, and confirmed visually with scatter plots. Multicollinearity is observed using the numerical r value between pairs of predictor variables.

2.3 Spatial Lag and Spatial Error Regression

2.3.1 Spatial Lag Regression

For this analysis, we will be using GeoDa for running spatial lag and spatial error regressions, which are essential for handling spatial dependencies in regression models.

Spatial lag regression is used to model spatial dependencies in regression analysis. It incorporates the influence of neighboring areas into the regression model. The model equation for spatial lag is:

$$\text{LNMEDHVAL} = \rho \cdot W\text{LNMEDHVAL} + \beta_1 \cdot \text{LNNBELPOV} + \beta_2 \cdot \text{PCTBACHMOR} + \beta_3 \cdot \text{PCTSINGLES} + \beta_4 \cdot \text{PCTVACANT} + \epsilon$$

Where:

- LNMEDHVAL is the natural logarithm of median home value, the dependent variable.
- ρ represents the spatial autoregressive coefficient, capturing the relationship between a region's median home value and the median home values of its neighbors.
- W denotes the spatial weights matrix, defining the structure of spatial relationships.
- LNNBELPOV, PCTBACHMOR, PCTSINGLES, and PCTVACANT are the independent variables reflecting neighborhood poverty levels, educational attainment, single households, and vacancy rates, respectively.
- ϵ is the error term.



2.3.2 Spatial Error Regression

Spatial error regression addresses spatial dependencies in the error terms of the regression model. The spatial error model used in this assignment is:

$$\text{LNMEDHVAL} = \beta_1 \cdot \text{LNNBELPOV} + \beta_2 \cdot \text{PCTBACHMOR} + \beta_3 \cdot$$

$$\text{PCTSINGLES} + \beta_4 \cdot \text{PCTVACANT} + \lambda \cdot Wu + \epsilon$$

Where:

- LNMEDHVAL is the natural logarithm of median home value, the dependent variable.
- λ is the spatial autoregressive coefficient for the error term, quantifying the spatial autocorrelation in errors.
- u represents the vector of error terms.
- W is as previously defined.
- LNNBELPOV , PCTBACHMOR , PCTSINGLES , and PCTVACANT are the independent variables reflecting neighborhood poverty levels, educational attainment, single households, and vacancy rates, respectively.
- ϵ is the stochastic error term.

2.3.3 Assumptions and Objectives for Spatial Regression

While spatial lag and spatial error models rely on the same foundational assumptions as OLS regression, including linearity, no endogeneity, homoscedasticity, and normality of error terms, with the exception of spatial independence.

The aim of using spatial regression models is to improve the estimation accuracy by accounting for spatial dependencies. This should lead to residuals that do not exhibit spatial autocorrelation, which in turn, would result in more reliable statistical inferences.

2.3.4 Model Comparison and Evaluation Criteria

The performance of spatial lag and spatial error regression models will be compared with the OLS model using the following criteria:

Akaike Information Criterion (AIC) and Schwarz Criterion (BIC): These criteria measure the relative goodness of fit of a model while penalizing for the number of estimated parameters, with lower values indicating a better model.

Log Likelihood: Reflects the probability of the data given the model, with higher values signifying a better fit.



Likelihood Ratio Test: Compares the fit of two nested models, with the null hypothesis positing that the simpler model is true, against an alternative hypothesis that the more complex model is better.

Further, the Moran's I statistic for regression residuals will be analyzed as a measure of spatial autocorrelation. A lower Moran's I value for the spatial models, as compared to the OLS model, would suggest a better fit for capturing spatial processes in the data.

2.4 Geographically Weighted Regression

The Geographically Weighted Regression (GWR) analyses for this study will be conducted using the statistical programming environment R, which provides robust tools for spatial data analysis, including the `spgwr` package for running GWR models.

GWR is a local form of linear regression used to model spatially varying relationships. This method addresses the limitations highlighted by Simpson's paradox, where aggregated data can suggest misleading associations that do not hold when examining local subsets of the data. GWR allows for regression coefficients to vary across space, thus enabling the exploration of local rather than global relationships between variables.

The GWR model is formulated as:

$$y_i = \beta_0 \cdot (u_i, v_i) + \sum \beta_k \cdot (u_i, v_i) \cdot x_{ik} + \epsilon_i$$

Where:

- y_i is the dependent variable at location i
- $\beta_0 \cdot (u_i, v_i)$ is the intercept term that varies with location i , denoted by coordinates (u_i, v_i) .
- $\beta_k \cdot (u_i, v_i)$ are the local regression coefficients for each predictor x_{ik} at location i .
- ϵ_i is the random error term at location i .

In essence, GWR extends the traditional regression framework by allowing the regression coefficients to be functions of spatial location, thus providing a local model of the variable relationships.

Local regression in GWR is run for each feature in the dataset. For each location, a separate regression equation is estimated, where the data points are weighted according to their proximity to the location in question. This means that observations near the location have more influence on the local model than those farther away.

2.4.1 Bandwidth

Bandwidth is a critical parameter in GWR as it determines the scale of the local analysis by controlling how much the data is smoothed or localized. There are two types of bandwidth:

Fixed Bandwidth: Every location is influenced by a fixed radius of neighbors.

Adaptive Bandwidth: The number of neighbors included varies such that each location is influenced by a fixed number of nearest neighbors.

For this analysis, an adaptive bandwidth will be used because it accounts for varying densities of data points across the study area, which is particularly important in unevenly distributed populations or when dealing with urban and rural divides.

2.4.2 OLS Assumptions and Multicollinearity in GWR

OLS assumptions are also applicable to GWR, including normality of residuals, homoscedasticity, no multicollinearity, normal dependent variable or residuals. However, due to the local nature of the regression, issues such as multicollinearity may manifest differently. The Condition Number is used to assess multicollinearity in GWR. A high Condition Number is required for GWR (at least 300 observations).

2.4.3 P-values and GWR Output

P-values are not traditionally part of GWR output because the local nature of the model means that standard statistical tests of significance are not applicable. The local regression coefficients vary across the study area, and thus the concept of a single p-value for the whole model does not hold. Instead, GWR focuses on the variation and significance of the relationships at each location.



3. Results

3.1 Spatial Autocorrelation

3.1.1 Global Moran's I

The Global Moran's I (Queen Matrix) statistic is 0.793565, which is close to 1, suggesting a high level of positive spatial autocorrelation. This means that similar values of the dependent variable LNMEDHVAL are clustered together in space. The p-value is 0, which is less than the conventional alpha level of 0.05, indicating that the observed spatial pattern is very unlikely to be the result of random chance. Therefore, we can reject the null hypothesis of spatial randomness.

The method used is a Monte-Carlo simulation of Moran's I with 1000 permutations. This approach repeatedly randomizes the location of values and calculates the Moran's I for each permutation to establish a reference distribution. The actual Moran's I is then compared to this reference distribution to calculate the p-value.

Based on the Global Moran's I value and the p-value, it can be concluded that LNMEDHVAL is significantly spatially autocorrelated. The positive Moran's I value indicates that the spatial pattern exhibits clustering of similar values rather than being dispersed or random.

3.1.2 Local Moran's I

Significance Map: Figure 3 depicts areas of statistical significance concerning spatial autocorrelation. Darker shades of purple indicate areas where the Local Moran's I statistic is very low, suggesting strong evidence against the null hypothesis of random distribution. These areas are likely to exhibit significant spatial clustering. The gradation of purple represents different levels of significance, with the darkest purple showing p-values between 0.000 to 0.001, which is highly significant.

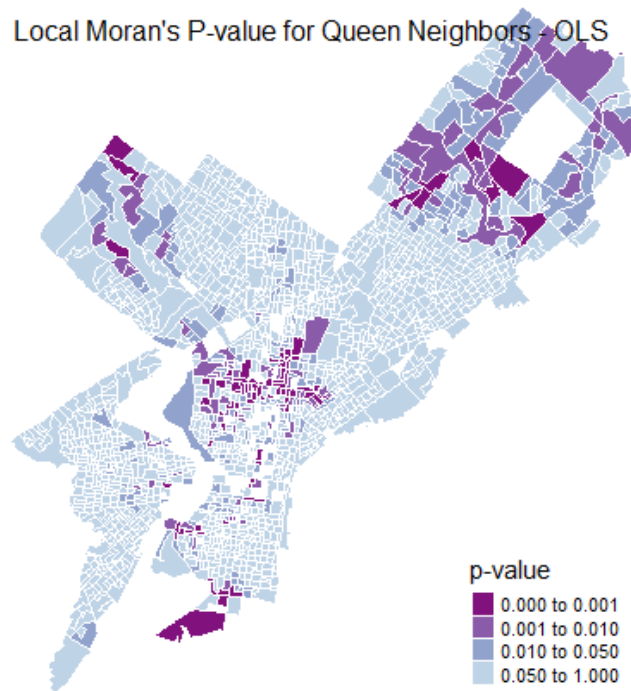


Figure 3: *Local Moran's P-value for Queen Neighbors*

Cluster Map: Figure 4 provides more insight into the type of spatial clusters present in our model. High-High (Red): These areas, shown in red, are clusters where the value of LNMEDHVAL is high and is surrounded by other areas with high values. These areas may represent neighborhoods with higher home values, possibly affluent or rapidly developing areas. Northeast Philadelphia and Chestnut Hill are "High-High" clusters, possibly corresponding to well-established or high-value real estate markets.

Low-Low (Blue): Shown in blue, these are areas where the value of LNMEDHVAL is low and surrounded by areas with similarly low values. These could be regions with lower home values, which may correlate with lower-income neighborhoods or less developed areas. North Philadelphia includes "Low-Low" clusters, potentially reflecting less developed or less expensive neighborhoods.

Low-High (Light Blue): These areas have low values of LNMEDHVAL but are surrounded by high-value areas. They could be pockets of less expensive homes within generally affluent areas or perhaps transitional zones where development is not uniform.

High-Low (Pink): These regions have high values but are surrounded by low-value areas, potentially indicating gentrifying neighborhoods or areas that stand out due to higher home values amongst less valuable regions.

Not Significant (Grey): These are areas where the Local Moran's I did not find significant spatial autocorrelation at the chosen significance level, indicating a more random spatial pattern for LNMEDHVAL in these locations.

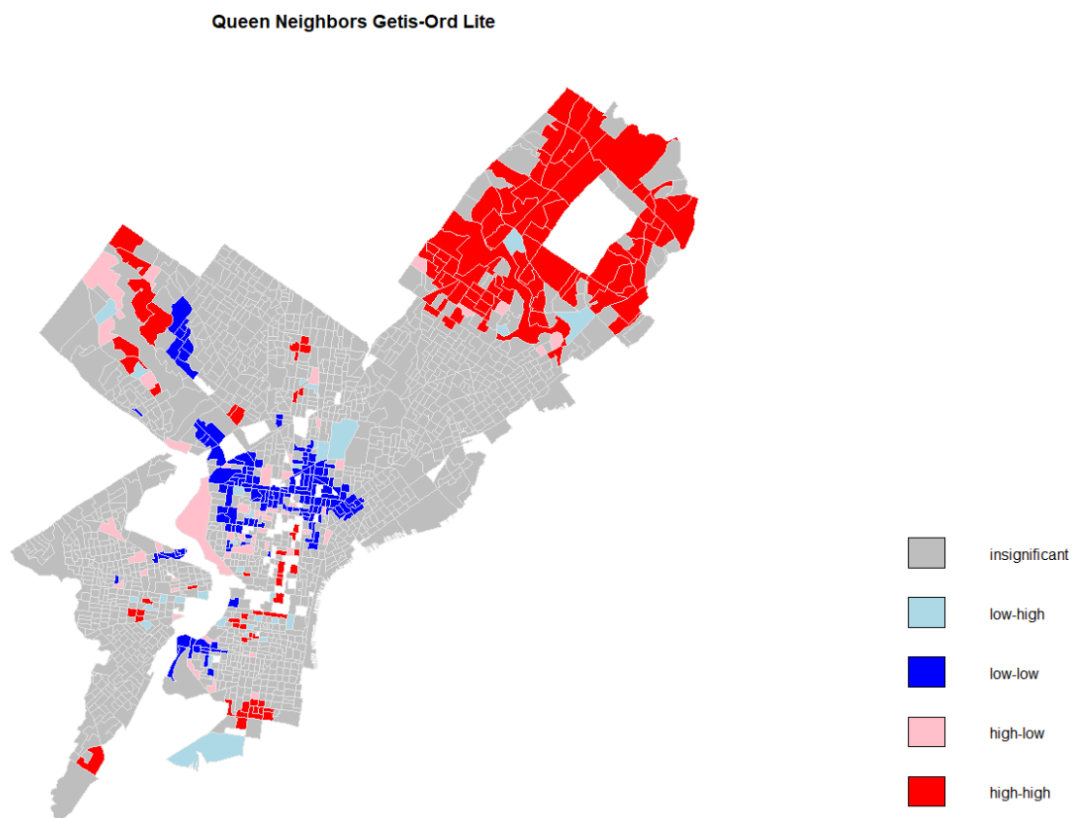


Figure 4: *Cluster Map*

3.2 A Review of OLS Regression and Assumptions: Results

3.2.1 Overview of OLS Regression results

The regression output (figure 5) reveals the following:

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION				
Data set	:	RegressionData		
Dependent Variable	:	LNMEDHVAL	Number of Observations:	1720
Mean dependent var	:	10.882	Number of Variables	: 5
S.D. dependent var	:	0.62972	Degrees of Freedom	: 1715
R-squared	:	0.662300	F-statistic	: 840.869
Adjusted R-squared	:	0.661513	Prob(F-statistic)	: 0
Sum squared residual	:	230.332	Log likelihood	: -711.493
Sigma-square	:	0.134304	Akaike info criterion	: 1432.99
S.E. of regression	:	0.366475	Schwarz criterion	: 1460.24
Sigma-square ML	:	0.133914		
S.E of regression ML	:	0.365942		
Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	11.1138	0.0465318	238.843	0.00000
LNMBELPOV	-0.0789035	0.0084567	-9.3303	0.00000
PCTBACHMOR	0.0209095	0.000543184	38.4944	0.00000
PCTSINGLES	0.00297695	0.000703155	4.23371	0.00002
PCTVACANT	-0.0191563	0.000977851	-19.5902	0.00000
REGRESSION DIAGNOSTICS				
MULTICOLLINEARITY CONDITION NUMBER		12.990609		
TEST ON NORMALITY OF ERRORS				
TEST	DF	VALUE	PROB	
Jarque-Bera	2	778.9646	0.00000	
DIAGNOSTICS FOR HETEROSKEDASTICITY				
RANDOM COEFFICIENTS				
TEST	DF	VALUE	PROB	
Breusch-Pagan test	4	162.9108	0.00000	
Koenker-Bassett test	4	61.6992	0.00000	

Figure 5: OLS Regression Output

The association between the percentage of Vacant Houses (PCTVACANT) and LNMEDHVAL is highly significant and negative as displayed by a p-value less than 0.001; The percentage of Single Houses (PCTSINGLES) has a significant and positive association with LNMEDHVAL where the p-value is lower than 0.001 and β_2 approximately equals



3. The relationship between the percentage with Bachelor's Degrees or higher (PCTBACHMOR) and LNMEDHVAL is highly significant and positive with a p-value lower than 0.001 and coefficient β_3 that equals 0.021; There is a highly significant and negative relationship between the natural log of individuals below the poverty line (LNNBELPOV100) and LNMEDHVAL with a p-value lower than 0.001 and coefficient β_4 that equals 0.079.

The model explains about two-thirds (66.23%) of the variance in LNMEDHVAL, as evidenced by the R^2 value of 0.6623 and the Adjusted R^2 value of 0.6615. The latter communicates that the larger number of predictor variables was not especially detrimental to our model. Additionally, this is considerably higher than the provided example. The extremely low p-value associated with the F-ratio, which is $2.2E^{-16}$, indicates that we can reject the null hypothesis that all coefficients in the model are 0, affirming that the predictors in the model collectively exhibit a significant relationship with LNMEDHVAL.

3.2.2 Heteroscedasticity Tests

The Breusch-Pagan and White's tests for heteroscedasticity both indicate the presence of heteroscedasticity in the residuals of the model. The Breusch-Pagan test yielded a test statistic of 113.19 with a p-value less than 0.00000000000000022, and the studentized Breusch-Pagan test showed a test statistic of 42.868 with a p-value of 0.00000001102. Similarly, White's test provided a test statistic of 43.94 with a p-value of 0. These consistent results across different tests strongly suggest that the variance of the residuals is not constant, indicating a violation of the homoscedasticity assumption in the OLS model. This finding aligns with the visual evidence from the residual by predicted plot presented in Homework 1.

3.2.3 Normality of Errors

The Jarque-Bera test applied to the residuals of the model yielded a test statistic of 778.96, with a p-value less than 0.00000000000000022, suggesting a significant departure from normality. This result points to a potential issue with the assumption that the error terms are normally distributed. The histogram of residuals from Homework 1 (figure 6) shows a normal distribution, which suggest that the Jarque-Bera test's sensitivity to large sample sizes might be detecting subtler departures from normality that are not easily visualized.

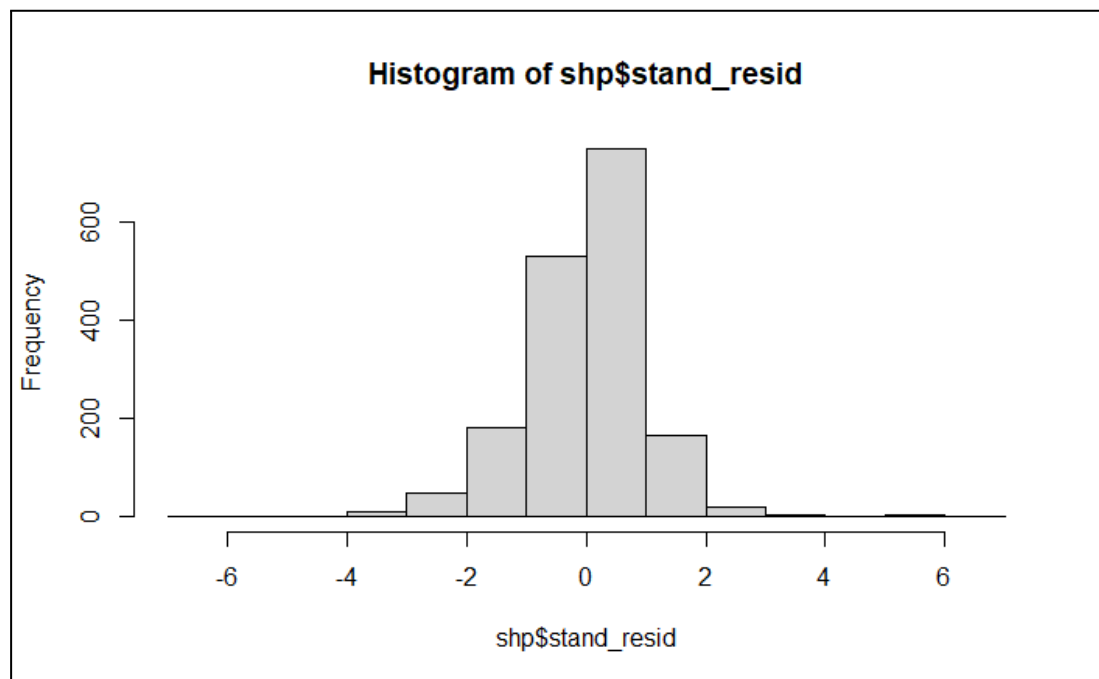


Figure 6: *Histogram of Standardized Residuals*

3.2.4 Scatterplot of OLS_RESIDU by WT_RESIDU

The Slope (b) at the bottom of the scatter plot (figure 7) represents the coefficient of the spatially lagged variable (WT_RESIDU) in the regression model where OLS residuals are regressed on WT_RESIDU. This slope is calculated by the least squares method which

minimizes the sum of squared differences between the observed values (OLS_RESIDU) and the values predicted by the linear regression model.

The scatterplot of OLS_RESIDU by WT_RESIDU (figure 7) shows the relationship between the OLS residuals and the spatially lagged residuals. The points show a discernible pattern, such as a linear ascent trend, which suggest that there is spatial autocorrelation.

The p-value for Slope b is 0, which strongly suggests that there is significant spatial autocorrelation in the OLS residuals. This implies that the residuals are not randomly distributed but instead exhibit a pattern that is related to the spatial configuration of the data points.

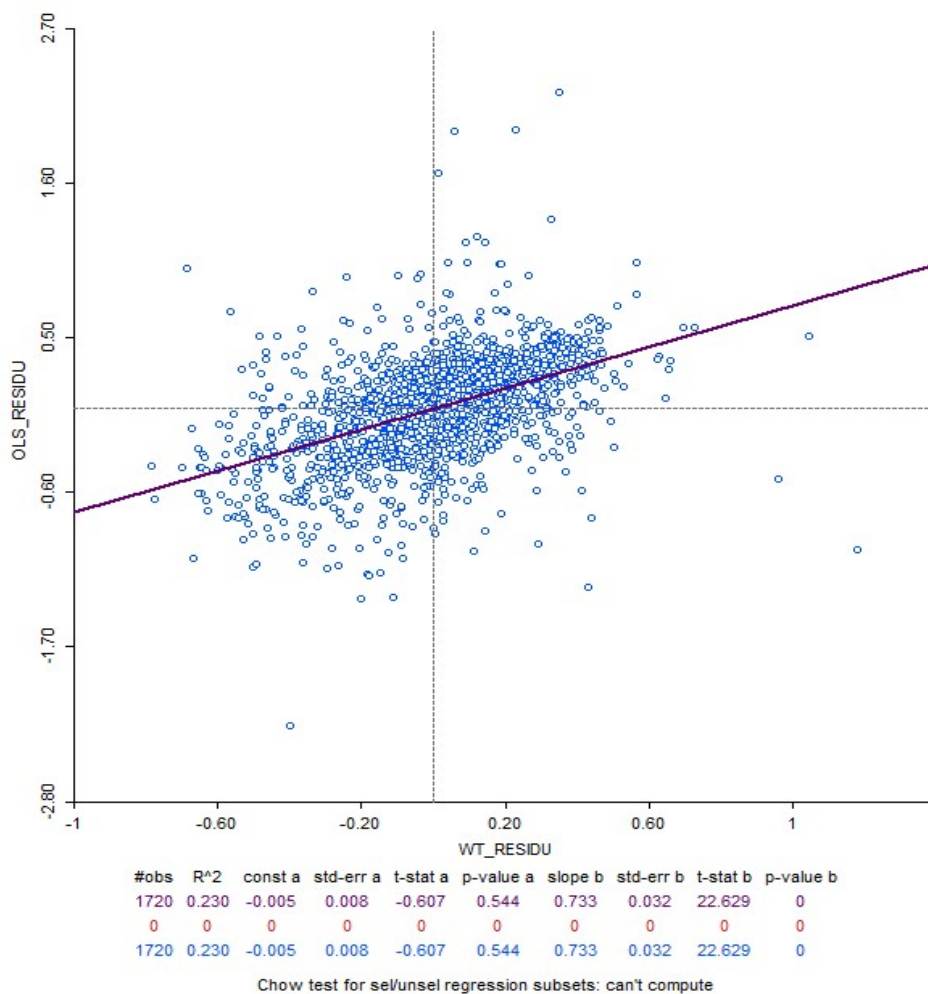


Figure 6: Scatterplot of OLS_RESIDU by WT_RESIDU

3.2.5 Moran's I scatterplot and results of OLS regression residuals.

In Moran's I scatter plot (figure 7), we observe the standardized residuals from the OLS regression on the x-axis plotted against the spatially lagged standardized residuals on the y-axis. The slope of the line in the scatterplot indicates the Moran's I statistic.

From the scatterplot, there appears to be a positive slope, suggesting a positive spatial autocorrelation among the OLS residuals. This means that locations with high residuals tend to be near other locations with high residuals, and similarly for low residuals.

The presence of significant spatial autocorrelation in the residuals is problematic because it violates one of the key OLS assumptions — that of independently distributed errors. This violation can lead to biased standard errors and, consequently, unreliable hypothesis tests and confidence intervals. It suggests that the OLS model may be missing important spatial predictors or that a spatial process is at work that the model does not account for.

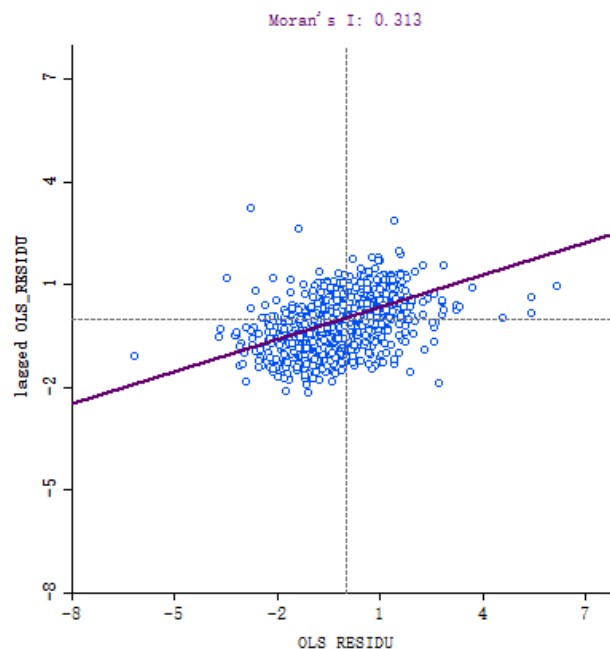


Figure 7: Moran's I scatter plot and results of OLS regression residuals.

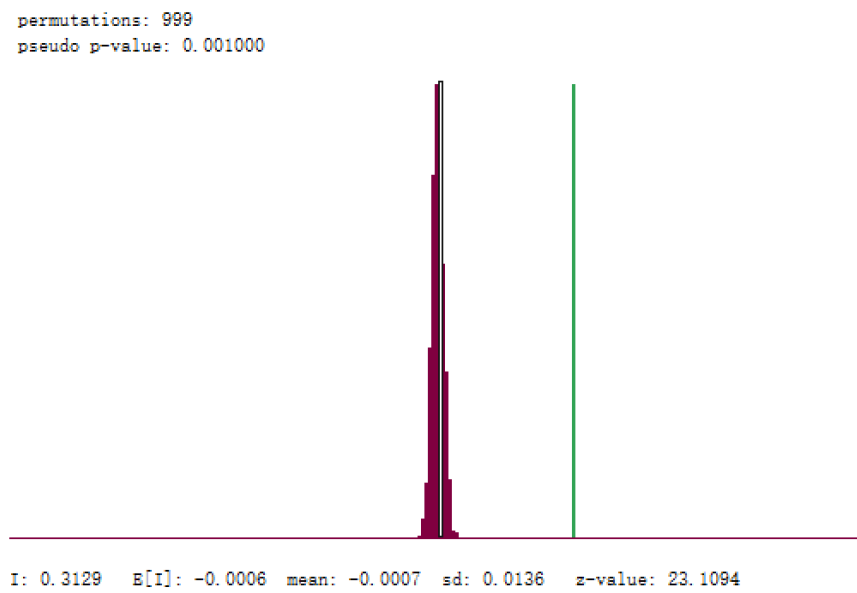


Figure 8: *Histogram of Moran's I values for 999 permutations*

3.3 Spatial Lag Regression Results

SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION				
Data set	: RegressionData			
Spatial Weight	: RegressionData			
Dependent Variable	: LNMEDHVAL	Number of Observations	: 1720	
Mean dependent var	: 10.882	Number of Variables	: 6	
S.D. dependent var	: 0.62972	Degrees of Freedom	: 1714	
Lag coeff. (Rho)	: 0.651097			
R-squared	: 0.818564	Log likelihood	: -255.74	
Sq. Correlation	: -	Akaike info criterion	: 523.48	
Sigma-square	: 0.071948	Schwarz criterion	: 556.18	
S.E of regression	: 0.268231			
Variable	Coefficient	Std. Error	z-value	Probability
W_LNMEDHVAL	0.651097	0.0180501	36.0716	0.00000
CONSTANT	3.89846	0.201114	19.3843	0.00000
LNNBELPOV	-0.0340547	0.00629287	-5.41163	0.00000
PCTBACHMOR	0.00851381	0.000521935	16.312	0.00000
PCTSINGLES	0.00203342	0.00051577	3.9425	0.00008
PCTVACANT	-0.0085294	0.000743667	-11.4694	0.00000
REGRESSION DIAGNOSTICS				
DIAGNOSTICS FOR HETEROSKEDASTICITY				
RANDOM COEFFICIENTS				
TEST		DF	VALUE	PROB
Breusch-Pagan test		4	220.3884	0.00000
DIAGNOSTICS FOR SPATIAL DEPENDENCE				
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : RegressionData				
TEST		DF	VALUE	PROB
Likelihood Ratio Test		1	911.5067	0.00000

Figure 9: Result table of Spatial Lag Regression

3.3.1 Summary Table

The spatial lag term, W_LNMEDHVAL, is highly significant in the spatial lag regression model with a coefficient of 0.651 and a p-value of 0.000. This signifies a strong spatial dependence in median home values. Specifically, the coefficient indicates that a unit increase in the median home values in neighboring areas is associated with a 65.1% increase in the median home value in the focal area, suggesting that nearby areas have a substantial influence on a given area's home values.

The predictors LNNBELPOV, PCTBACHMOR, PCTSINGLES, and PCTVACANT retain their statistical significance in the spatial lag model. The coefficients of these variables show similar significance levels to those in the OLS regression model. Specifically:

LNNBELPOV has a negative association with LNMEDHVAL, indicating that higher levels of the population below poverty correlate with lower home values; PCTBACHMOR remains positively associated, suggesting that areas with higher percentages of individuals with a bachelor's degree or more have higher median home values; PCTSINGLES also shows a positive association, indicating that areas with a higher percentage of singles correlate with higher home values; PCTVACANT has a negative association, which suggests that higher vacancy rates correspond to lower median home values. Comparing these results to the OLS model, the significance and direction of the relationships remain consistent, although the magnitudes of the coefficients might have been altered due to the inclusion of the spatial lag term, which adjusts for the influence of spatially correlated errors.

3.3.2 Heteroscedasticity

The Breusch-Pagan test for heteroscedasticity provides a test statistic value of 220.3884 with a p-value of 0.0000. This indicates that there is significant heteroscedasticity present in the residuals of the spatial lag model. Despite accounting for spatial autocorrelation, the model still exhibits non-constant variance across the residuals, which can affect the efficiency of the estimators and the robustness of standard errors and test statistics. This suggests that the model may benefit from further refinement or the use of robust standard errors to account for this heteroscedasticity.

3.3.3 Compare the Spatial Lag regression and OLS regression models

Akaike Information Criterion (AIC) and Schwarz Criterion (BIC): The AIC for the OLS model is 1432.99, and the BIC is 1460.24, whereas for the Spatial Lag Model, the AIC is 523.48, and the BIC is 556.18. The Spatial Lag Model has much lower values for both criteria, indicating a better fit relative to the number of parameters in the model.

Log Likelihood: The log likelihood for the OLS model is -711.03, compared to -255.74 for the Spatial Lag Model. The higher (closer to zero) log likelihood of the Spatial Lag Model suggests it is a better fit for the data.

The Likelihood Ratio Test result provided indicates a highly significant improvement of the Spatial Lag Model over the OLS model, with a test statistic of 911.5067 and a p-value of 0.00000. This test evaluates whether the inclusion of the spatial lag term significantly improves the model fit, and the results here strongly suggest that it does.

3.3.4 Moran's I scatter plot of spatial lag regression residual

The Moran's I scatterplot for the OLS regression residuals (figure 7) shows a Moran's I value of 0.313, which indicates a moderate positive spatial autocorrelation. In contrast, the Moran's I scatterplot for the Spatial Lag Model residuals (figure 10), which showed a Moran's I statistic of -0.082, suggests almost no spatial autocorrelation. This value, being closer to zero and negative, indicates that the Spatial Lag Model has effectively reduced the spatial autocorrelation present in the OLS model residuals.

The histogram of the Moran's I values from 999 permutations (figure 10) shows a distribution centered very close to zero, with the observed Moran's I value lying in the tail, which corroborates the significance found in the permutation test. The negative z-value of -5.8778 further confirms the statistical significance of the negative spatial autocorrelation.

When comparing the two scatterplots, it is evident that there is less spatial autocorrelation in the residuals of the Spatial Lag Model than in the OLS model residuals.

This reduction is a big improvement, as spatial autocorrelation in the residuals of a regression model violates the assumption of independent errors and can lead to incorrect inferences.

Overall, the comparison of Moran's I values and the patterns in the scatterplots suggest that the Spatial Lag Model is better suited for data with spatial dependency and provides a more reliable understanding of the underlying spatial processes affecting the dependent variable. This conclusion supports the use of the Spatial Lag Model over the OLS model for this particular dataset.

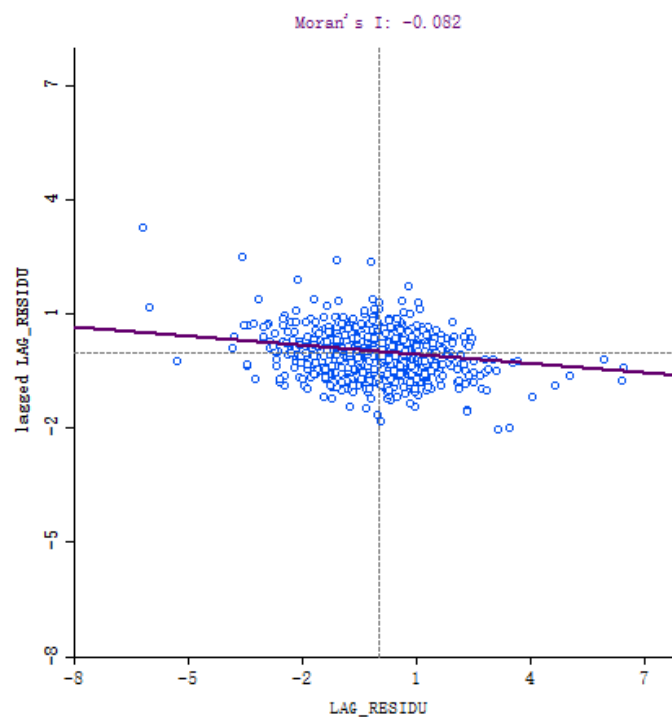


Figure 10: *Moran's I scatter plot of spatial lag regression residual*

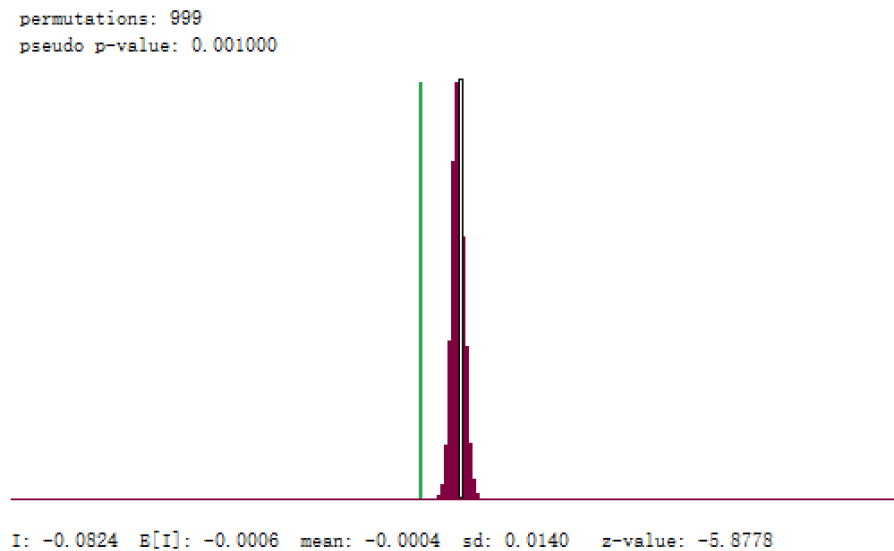


Figure 10: Histogram of Moran's I values for 999 permutations

3.3.4 Overall Performance of the Spatial Lag Model:

The Spatial Lag Model has lower AIC and BIC values than the OLS model, indicating a better fit to the data when accounting for the number of parameters. For Log Likelihood, the Spatial Lag Model has a higher log likelihood than the OLS model, suggesting it is more likely to have produced the observed data. Regarding to the Moran's I , the Spatial Lag Model has a Moran's I value closer to zero (-0.082) than the OLS model (0.313), indicating less spatial autocorrelation in the residuals. Despite the negative spatial autocorrelation being statistically significant, it is still an improvement over the positive spatial autocorrelation found in the OLS residuals.

Considering these aspects, the Spatial Lag Model is performing better than the OLS model. It not only improves the fit according to information criteria but also addresses the issue of spatial autocorrelation to a considerable extent, which is a key concern in spatial data analysis. The significant negative spatial autocorrelation in the Spatial Lag Model suggests that while it has corrected for the positive spatial autocorrelation present in the OLS model,

there may be room for further model refinement. However, in terms of overall model performance, the Spatial Lag Model is superior based on the provided criteria.

3.4 Spatial Error Regression Results

SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION				
Data set	:	RegressionData		
Spatial Weight	:	RegressionData		
Dependent Variable	:	LNMEDHVAL	Number of Observations:	1720
Mean dependent var	:	10.882000	Number of Variables	: 5
S.D. dependent var	:	0.629720	Degrees of Freedom	: 1715
Lag coeff. (Lambda)	:	0.814918		
R-squared	:	0.806957	R-squared (BUSE)	: -
Sq. Correlation	:	-	Log likelihood	: -372.690368
Sigma-square	:	0.0765508	Akaike info criterion	: 755.381
S.E of regression	:	0.276678	Schwarz criterion	: 782.631
<hr/>				
Variable		Coefficient	Std. Error	z-value
				Probability
CONSTANT		10.9064	0.0534678	203.981
LNMBELPOV		-0.0345341	0.00708933	-4.87127
PCTBACHMOR		0.00981293	0.000728964	13.4615
PCTSINGLES		0.00267792	0.000620832	4.31343
PCTVACANT		-0.00578308	0.000886701	-6.52201
LAMBDA		0.814918	0.016373	49.7719
<hr/>				
REGRESSION DIAGNOSTICS				
DIAGNOSTICS FOR HETEROSKEDASTICITY				
RANDOM COEFFICIENTS				
TEST		DF	VALUE	PROB
Breusch-Pagan test		4	210.9923	0.00000
 DIAGNOSTICS FOR SPATIAL DEPENDENCE				
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : RegressionData				
TEST		DF	VALUE	PROB
Likelihood Ratio Test		1	677.6059	0.00000

Figure 11: Result table of Spatial Error Regression

3.4.1 Summary Table

The spatial error regression output includes a term LAMBDA with a coefficient of 0.814918. This term is highly significant, with a z-value of 49.7719 and a probability

effectively at zero. The LAMBDA coefficient represents the spatial autoregressive parameter for the error component of the model. Its significance indicates that there is a substantial spatial correlation in the error terms of the regression model. The positive coefficient suggests that if a nearby location has a positive error term, it is likely that the focal location will also have a positive error term, and vice versa for negative error terms.

All the independent variables—LNNBELPOV, PCTBACHMOR, PCTSINGLES, and PCTVACANT—are significant in the spatial error model. Their coefficients and significance levels are consistent with the OLS model results, suggesting that these factors consistently influence the dependent variable across different model specifications.

3.4.2 Heteroscedasticity

The Breusch-Pagan test for the spatial error model indicates a test statistic value of 210.9923 with a probability of 0.00000, showing that the residuals are still heteroscedastic. This suggests that even after accounting for spatial correlation in the error terms, the model does not have a constant variance among the residuals.

3.4.3 Model Comparison:

The Akaike Information Criterion (AIC) and Schwarz Criterion (BIC) are lower for the spatial error model than for the OLS model, suggesting a better fit. The Log Likelihood is also improved in the spatial error model (-372.690368) compared to the OLS model (-711.03), indicating that the spatial error model is more probable in explaining the observed data. The Likelihood Ratio Test strongly rejects the simpler model in favor of the spatial error model, with a value of 677.6059 and a probability effectively at zero.

3.4.4 Moran's I Scatterplot for Spatial Error Model:

The Moran's I scatterplot for the spatial error model residuals shows a value of -0.095, which, alongside a pseudo p-value of 0.00100, indicates a statistically significant negative spatial autocorrelation. When compared to the OLS residuals, which had a positive Moran's I value of 0.313, there is a reduction in spatial autocorrelation, suggesting an improvement in model specification.

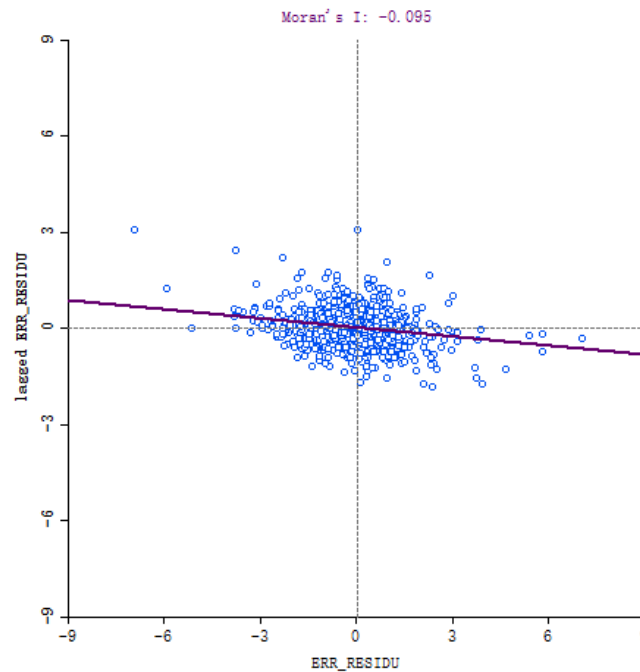


Figure 12: *Moran's I scatter plot of spatial Error regression residual*

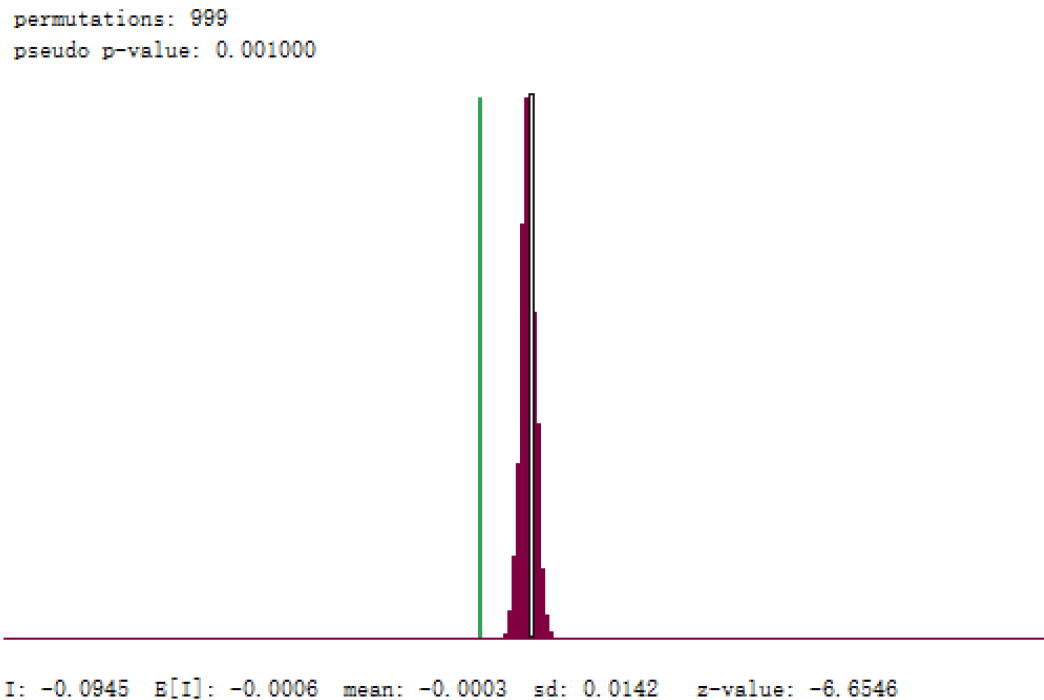


Figure 13: *Histogram of Moran's I values for 999 permutations*

3.4.5 Overall Performance of the Spatial Error Model:

Considering the significant improvement in AIC, BIC, and Log Likelihood, along with the reduction in spatial autocorrelation as evidenced by a negative Moran's I value, the spatial error model is performing better than the OLS model. However, the presence of heteroscedasticity is still a concern, indicating that additional model refinement could be necessary. Despite this, the spatial error model better accounts for the spatial structure in the data and provides a more reliable analysis based on the criteria evaluated.

3.4.6 Comparison of Spatial Lag and Spatial Error Models:

When assessing the performance of non-nested models like the Spatial Lag and Spatial Error models, we look to the Akaike Information Criterion (AIC) and the Schwarz Information Criterion (BIC) for guidance.

- For the Spatial Lag Model, the AIC is 523.48 and the BIC is 556.18.
- For the Spatial Error Model, the AIC is 755.831 and the BIC is 782.631.

According to the AIC and BIC values, the Spatial Lag Model has a lower AIC and BIC compared to the Spatial Error Model. Lower values of AIC and BIC indicate a better balance of model fit and complexity, suggesting that the Spatial Lag Model is more efficient in terms of information loss and parsimony. Based on the Akaike Information Criterion and the Schwarz Information Criterion, the Spatial Lag Model is the preferred model over the Spatial Error Model.

3.5 Geographically Weighted Regression Results

3.5.1 Bandwidth Selection

We used the `gwr.sel` function to calculate the optimal bandwidth for GWR. We used both adaptive bandwidth and fixed bandwidth. Adaptive bandwidth refers to varying the distance, but fixes the number of neighbors for each observation. Fixed bandwidth is more closely approximate distance.

For adaptive bandwidth, it is recommended that 0.008130619 of the observations be used for each location (with AIC= 660.7924) - the bandwidth should be adapted to capture about 14 observations (0.00813×1720) for each location. For fixed bandwidth, AIC reaches its lowest value of 700.3524 when bandwidth = 2863.493. We chose adaptive bandwidth of 0.008130619 to run the GWR due to lower AIC value.

3.5.2 GWR Results and Model Fit Comparison:

```
gwr(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
     LNNBELPOV, 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
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
PCTBACHMOR     0.0010974  0.0101380  0.0149279  0.0202187  0.0347258  0.0209
LNNBELPOV     -0.2365244 -0.0733572 -0.0401186 -0.0126657  0.0948768 -0.0789
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
```

Figure 14 Result table of Geographically Weighted Regression

The R-squared value for the GWR model is reported as a quasi-global R2 of 0.8479244, suggesting a high degree of variance in the dependent variable explained by the model. When compared to the R-squared of the OLS regression, which was 0.662300, the GWR model appears to provide a better fit overall, capturing the local variations in the relationship between predictors and the dependent variable that a global model like OLS may miss.

The AIC value for the GWR model is 308.7123, which is a measure of the model's relative quality and complexity. In comparison:

- The AIC for the OLS model was 1432.99,
- For the Spatial Lag model, it was 523.48,
- And for the Spatial Error model, it was 755.831.

Based on the AIC values, the GWR model shows a significantly better fit than the other three models, with the lowest AIC value indicating it is the most efficient at explaining the variance in the data.



3.5.2 Moran's I Scatter Plot Analysis for GWR Residuals:

From the scatterplot, it appears that the residuals from the GWR model do not exhibit a strong pattern of spatial autocorrelation. The scatter is fairly random without a clear trend, which would suggest that the GWR model has accounted for much of the spatial structure that was present in the data. When comparing this to the previously mentioned Moran's I values of the OLS residuals (0.313), Spatial Lag (-0.052), and Spatial Error (-0.095), the GWR residuals (0.033) suggests less spatial autocorrelation than the OLS residuals, and they are comparable to the Spatial Lag and Spatial Error models.

The histogram displays a normal distribution centered around zero, with the majority of the permutation-based Moran's I values clustering close to zero, indicating a general lack of spatial autocorrelation. The vertical red line, representing the observed Moran's I value, is close to the mean of the distribution, further supporting the conclusion of minimal spatial autocorrelation.

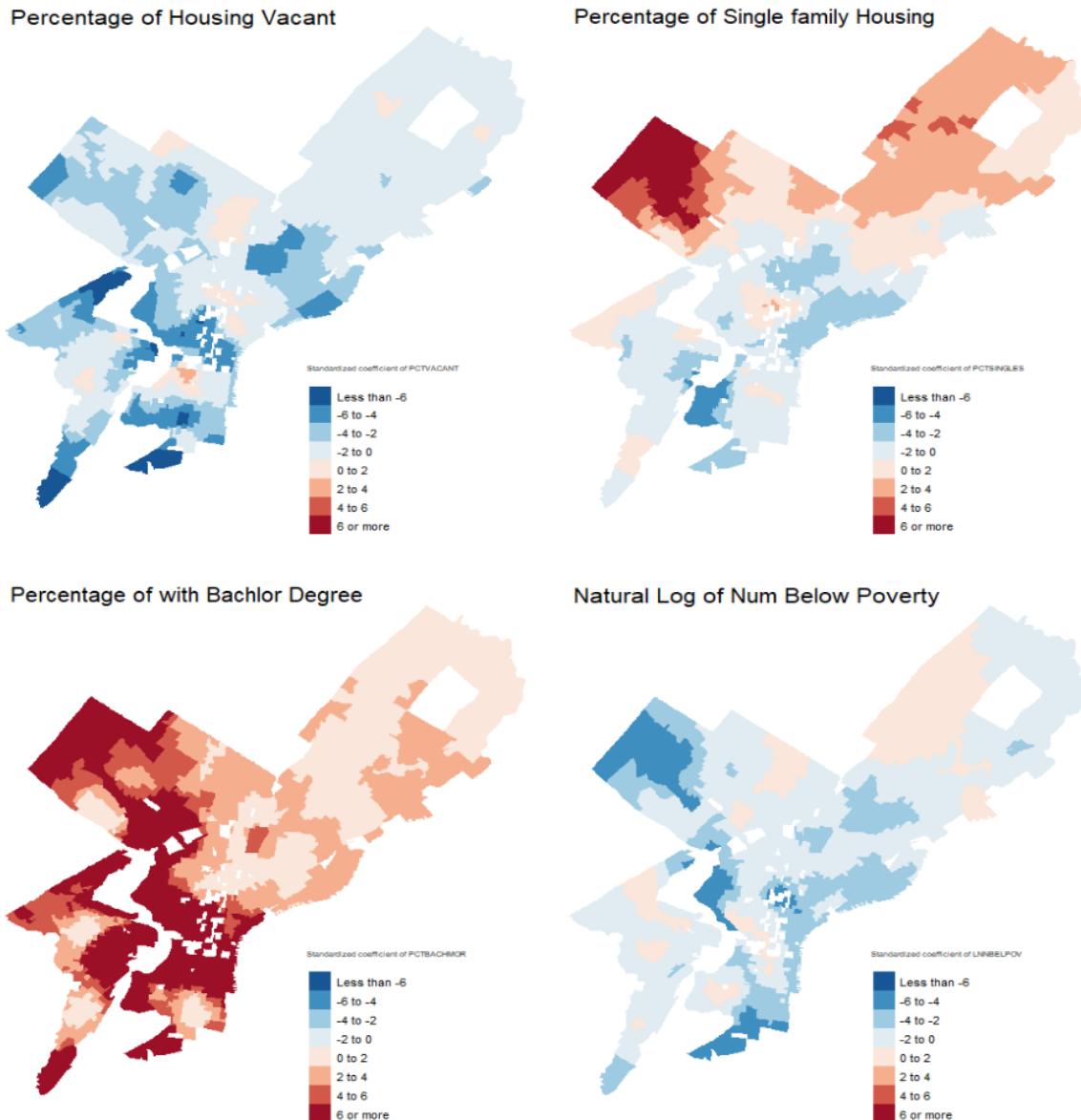


Figure 16 *Moran's I scatter plot of GWR residuals*

Figure 17 shows the spatial distribution of the local R-squares. The gradation of colors from light to dark blue represents the range from low to high local R-squared values. The areas shaded in darker blue signify locations where the model's predictors have a stronger explanatory power for the variance of the dependent variable, while lighter shades indicate where the model explains less of the variance. We can see that the four predictors do

a good job explaining the variance in our dependent variable in the most parts of Philadelphia, but not in many other parts of the city.

The north and northeastern parts of Philadelphia, depicted in lighter blue, indicate a lower explanatory power, where the four predictors do not account for much of the variance in the dependent variable house value. This could be due to demographic, economic, or spatial factors unique to this area that are not captured by the current model.

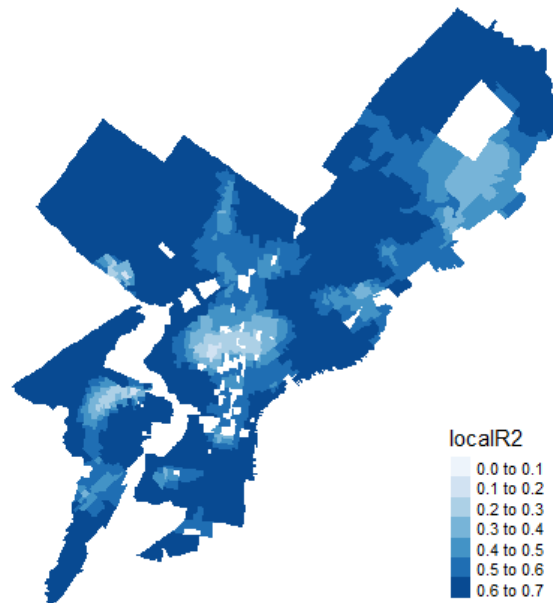


Figure 17 *Choropleth map of local R-squares*

4. Discussion

This paper aimed to investigate spatial patterns in median home values across Philadelphia block groups using various regression models that account for spatial autocorrelation. Our analysis included Ordinary Least Squares (OLS), Spatial Lag, Spatial Error, and Geographically Weighted Regression (GWR) methods. We found significant spatial autocorrelation in the dataset, as evidenced by a high Global Moran's I value. The GWR model, which accounts for local variations, provided the best fit based on the lowest

Akaike Information Criterion (AIC) value and the highest quasi-global R-squared, indicating its superior performance in explaining the variance in median home values.

The limitations of our study include the presence of heteroscedasticity in the Spatial Lag and Spatial Error models, suggesting that variance in the residuals was not consistent across all levels of the independent variables. Additionally, the Global Moran's I indicated significant spatial autocorrelation in the dataset, which challenges the assumption of independent observations required for OLS regression.

Weighted residuals refer to the residuals from an OLS regression model that have been adjusted by neighboring values. This is different from spatial lag model residuals, which arise from a model that includes a spatial lag of the dependent variable as an independent variable. The spatial lag model residuals are the remaining errors after accounting for both the local effects of the predictors and the spatially lagged dependent variable.

ArcGIS has issues with GWR, specifically in its calculation of local R-squared values. The software has been shown to generate negative local R-squared values, which is conceptually incorrect as R-squared values, indicating the proportion of variance explained by a model, should logically fall between 0 and 1. Such discrepancies undermine the reliability of GWR results produced by ArcGIS, suggesting a fundamental error in the software's statistical computations for GWR. Given these concerns, R is recommended for GWR analyses due to its proven consistency and reliability in statistical computations related to spatial data.

Works Cited

Booker, Bel. "The 100-Year History of Market Research - 1920 to 2020." *Attest*, 7 Apr. 2023, www.askattest.com/blog/articles/history-of-market-research.

Ryan, Bill, et al. "Demographics & Lifestyle Analysis." *Community Economic Development*, economicdevelopment.extension.wisc.edu/articles/demographics-lifestyle-analysis/. Accessed 13 Oct. 2023.

Anselin, Luc. "Spatial econometrics." *A companion to theoretical econometrics* 310330 (2001).