Using Spatial Lag, Spatial Error and Geographically Weighted Regression to Predict Median House Values in Philadelphia Block Groups

Fall 2025/MUSA 5000

Jun Luu

Alex Stauffer

Tessa Vu

1. Introduction

This report examines spatial relationships within Philadelphia, Pennsylvania. In the previous analysis, Ordinary Least Squares (OLS) regression was used to explore the relationship between the dependent variable (median home value, MEDHVAL) and a set of socioeconomic predictors (*e.g.* PCTVACANT, PCTBACHMOR, PCTSINGLES, NBELPOV100).

While OLS regression provides valuable baseline insights, it assumes that observations are independent. This assumption is often violated when analyzing spatial data, where nearby areas tend to be similar (spatial autocorrelation). Therefore, this report applies Spatial Lag, Spatial Error, and Geographically Weighted Regression (GWR) models to assess whether these spatial methods better explain the data compared to OLS.

2. Methods

2.1 Concept of Spatial Autocorrelation

As stated by Waldo Tobler, "Everything is related to everything else, but near things are more related than distant things" (Jones, 1998, pp. 1). This is the 1st Law of Geography and a core principle of spatial statistics. The concept of spatial autocorrelation examines the association between two variables much like non-spatial correlation but examines the relationship of values within a single variable at nearby locations.

To calculate the existence of spatial dependencies, we look to Moran's I. The equation is as follows:

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

Where $\bar{X}$ is the mean of 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 part of the weight matrix indexing location of i relative to j , and n is the number of observations (points or areal units). W is representative of a weight matrix that may either be Queen (eight neighbors) or Rook (four neighbors) distance-based, summarizing all pairwise relationships in the dataset.

Throughout the report we will be using a Queen weight matrix. Statisticians sometimes prefer to use multiple spatial matrixes on their datasets to assess whether their variables are truly spatially correlated. Queen or Rook weight contiguities could potentially expose different relationships among the variables and evaluate the consistencies of the spatial relationships.
The general parameters for testing the significance of Moran's I is to account for these definitions: First, large positive values near 1 indicate a strong positive correlation among variables, referred to as clustering. Second, large negative values near -1 indicate a strong negative correlation among variables, referred to as dispersion. Third, values near 0 tend to delineate no spatial autocorrelation, referred to as randomness. The range of -1 to +1 is not always apparent in Moran's I interpretations compared to the Pearson correlation coefficient. To reiterate, if a Moran's I value were 0.84, we can conclude it is an indication of strong positive spatial correlation.

The hypotheses in this report are as follows:

H_0: *No spatial autocorrelation.*
H_{a1}: *Positive spatial autocorrelation.*
H_{a2}: *Negative spatial autocorrelation.*

We can perform a random shuffle or permute the values of a given variable n times and then calculate Moran's I for each permutation. Then, the original variable's Moran's I and the permutated Moran's I will be arranged in descending order. Once this is complete, we can then compare the original observed variable's Moran's I value compared to that of the random permutation's.

Local spatial autocorrelation allows for the analysis of clustering at individual locations rather than across the entire study area. In this case, rather than assessing spatial clustering across all of Philadelphia, we calculate the Local Moran's I for each spatial unit (census tract, block group, neighborhood) to identify where local spatial clustering patterns occur. A common method for assessing local clustering is the Local Indicator of Spatial Association (LISA). LISA measures the extent to which the value of a variable at a given location i is similar to the values at neighboring locations j, as defined by the spatial weights matrix. The results of LISA are typically categorized into four spatial association types: High–High (HH) and Low–Low (LL) clusters, where both x_i and its neighbors x_j have values above or below the global mean (indicating positive spatial autocorrelation), and High–Low (HL) and Low–High (LH) outliers, where x_i differs substantially from its neighbors x_j (indicating negative spatial autocorrelation). These classifications reveal whether a location is part of a statistically significant cluster or a spatial outlier relative to its surrounding area.

2.2 Review of OLS Regression and Assumptions

OLS regression is a statistical method that examines the relationship between a dependent variable, also termed the response variable, and at least one other or more predictor variables, also termed the explanatory variables in two different types: the simple regression with one predictor and the multiple regression with more than one predictor. This technique has several assumptions: linear relationship(s) between y and x or multiple x variables, normally distributed residuals, random residuals, homoscedastic residuals, independence of observations and residuals, no multicollinearity if there are multiple predictors, and that y is continuous. For more information on OLS, refer to the previous paper, HW1.

Considering the above statements, when data possesses a spatial component (i.e. collected with location data like specific points or aggregated blocks in this case), the OLS assumption that errors are random does not often hold. This particular assumption can be tested by examining spatial autocorrelation of the data's residuals by using Moran's I, which is explained in section 2.1. Other ways to test OLS residuals for spatial autocorrelation is regressing them on nearby residuals, in this case, the neighboring block groups utilizing the Queen spatial matrix. In the Moran's I scatter plots throughout this paper is a line that cuts through the residuals, this is the "linear fit through the point cloud" which corresponds to the Moran's I value; this specific line is also known as “slope b” when using GeoDa and is “the slope of a regression of $\Sigma_j w_{ij} z_j$ on z_i” (Anselin, 2020).

The tool being used to run this OLS regression, R, has several different libraries that provide other ways to test other regression assumptions. This paper will go over two different assumption test alternatives in R: first, homoscedasticity, which is linked to the assumption of independence of errors; second, normality of errors, tying in with the homoscedastic assumption.

The following tests are for homoscedasticity:

Breusch-Pagan Test, which is preferably used when it is known that the residuals are suspected to be normal and assumes that the residual variance is linear, meaning it is not very robust encountering more complex relationships (Price, 2022, p. 11).

Koenker-Bassett Test (Studentized Breusch-Pagan Test), which is a modified version of Breusch-Pagan, but the main difference being that this version is less stringent on the normal distribution of errors and adjusts the test statistic. This makes it more robust against non-normal distributions of errors (e.g. heavy tail / skewed) (MetricGate, n.d.).

White’s Test, which is considerably well-rounded compared to the previous two, can be used for both normal and non-normal distributions of errors, able to test complex relationships because it includes interactions and squares (Statistics How To, n.d.-b).

All the above homoscedasticity tests have these hypotheses:

H_0: *Homoscedastic.*
H_a: *Heteroscedastic.*

The following test is for normality of errors:

Jarque Bera Test, which diagnoses normality of residuals, testing whether the skewness and kurtosis resemble a normal distribution. With that being said, it seems to focus on the shape of error distribution rather than their spread like the tests for heteroscedasticity (Statistics How To, n.d.-a).

This test has the hypotheses:

H_0: *Normal distribution.*
H_a: *Non-normal distribution.*

2.3 Spatial Lag and Spatial Error Regression

This paper will conduct spatial lag and spatial error regressions in R, both of which are described below.

Spatial lag regression observes how a dependent variable in a given space is influenced by its adjacent neighbors with that same dependent variable while considering how the values are related or bleed out into their neighbors. That is to say, it investigates whether a block's given value is a result of its neighbors' values by retrieving the average dependent variable value of the neighbors and adding it into a regression model—beyond the current predictors, this is accounting for the spatial idiosyncrasies of the block's local area.

$$y = \rho Wy + \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon$$

In this case, the equation would be:

$$LNMEDHVAL = \rho Wy + \beta_0 + \beta_1(LNBELPOV) + \beta_2(PCTBACHMOR)$$
$$\beta_3(PCTSINGLES) + \beta_4(PCTVACANT)$$

The beta coefficient β_0 is the predicted value of y, or LNMEDHVAL, when all other predictors equal zero.

Compared to spatial lag regression, the spatial error regression model addresses spatial autocorrelation in the error terms of the regression model. Spatial error regression assumes that the residual of a variable at one location is associated with residuals at nearby locations. The "nearby locations" are determined by the applied weights matrix *W.* The equation for the spatial error regression model for our purposes is:

$$LNMEDHVAL = \beta_0 + \beta_1(LNNBELPOV) + \beta_2(PCTBACHMOR) + \beta_3(PCTSINGLES)$$
$$+ \beta_4(PCTVACANT) + \lambda W\varepsilon + u$$

where $\varepsilon = \lambda W\varepsilon + u$

The beta coefficient β_0 represents the intercept and is the predicted value of our *y* or, LNMEDHVAL, when all other predictors are equal to zero. In this context, each β_i quantifies the expected change in the dependent variable associated with a one-unit increase in the corresponding independent variable, holding the remaining variables constant. Lamda, λ , is the coefficient of the spatially lagged errors $W\varepsilon$, indicating how much the residuals at one location is influenced by neighboring residuals. The range for λ is –1 to +1. The random noise error term is delineated as *u.* The error term comprises the spatially lagged errors plus random noise.

The assumptions for OLS are still needed for both spatial lag and spatial error regression models, excluding the necessity of spatial independence of observations. The assumptions include a linear relationship between the dependent variable and predictors, homoskedasticity, normality of residuals, and no multicollinearity. However, unlike an OLS regression, spatial lag and error models do not require spatial independence of observations. The primary goal of spatial lag and spatial error models is to take into account that there may be spatial dependencies in the residuals. The methods of spatial lag and spatial error aim to decrease heteroscedasticity and eliminate spatial autocorrelation which can be detected by calculating the regression's global Moran's I.

To decide which regression approach we will select, we will first compare the results of the spatial lag model to OLS, and then the results of the spatial error model to OLS. There are three criteria that can be used to decide which spatial model performs the best: Akaike Information Criterion/Schwarz Criterion (AIC/SC), Log Likelihood, and Likelihood Ratio Test. The AIC/SC measure the goodness of fit of an estimated statistical model. They can describe the tradeoff between precision and complexity of the model.

The AIC and Schwarz Criterion are not particularly meaningful when viewed in isolation; they only become valuable tools in the context of model comparison, where lower values indicate the better model. The next criterion to analyze is Log Likelihood, which is associated with the maximum likelihood method of fitting a statistical model to the data and estimating model parameters. The higher (less negative) the log likelihood value, the better the model fits the data. It should be noted that log likelihood comparison and the related Likelihood Ratio Test are only valid for comparing nested models—that is, models where one is a special case of the other. Since spatial lag and spatial error models are not nested models, these criteria can be used to compare each spatial model against OLS, but not to compare spatial lag directly against spatial error.

The Likelihood Ratio Test formally assesses whether a spatial model provides a statistically significant improvement over OLS. For the spatial lag model, the null hypothesis is that $\rho = 0$ (the spatial lag coefficient is zero), meaning the OLS model is adequate. For the spatial error model, the null hypothesis is that $\lambda = 0$ (the spatial error coefficient is zero), again meaning the OLS model is adequate. The alternative hypothesis in both cases is that the spatial model provides a significantly better fit. If the p-value from the likelihood ratio test is less than 0.05, we reject the null hypothesis and conclude that the spatial lag (or spatial error) model performs significantly better than the OLS model.

Another method to compare OLS results with spatial lag and spatial error models is by examining Moran's I of the regression of residuals. The better-performing model would account for spatial autocorrelation most effectively, resulting in residuals with Moran's I values closer to zero and ideally not statistically significant. If spatial autocorrelation persists in the residuals (indicated by a significant Moran's I), this suggests the model has not fully addressed the spatial dependencies in the data.

2.4 Geographically Weighted Regression

This paper will conduct a Geographically Weighted Regression (GWR) analysis in R. GWR tackles the assumption of stationarity held by OLS. Rather than having a global regression, GWR allows for separate, local regressions for each location in the model. We can visually conceptualize this notion with Simpson’s paradox. This phenomenon in statistics is where a trend that appears in aggregate data reverses or disappears when the data is divided into subgroups. For example, when examining the relationship between two variables across an entire city, we might observe one particular pattern. However, when we disaggregate the data and examine the relationship within different neighborhoods or regions of the city, we may find that the pattern differs substantially—or even reverses—within each local area. This occurs because the aggregate relationship can be driven by differences between regions, while the local relationships reflect dynamics within each region. Simpson's paradox demonstrates that relationships observed at the global scale may not hold at the local scale, which is why GWR uses local regression to capture these spatially varying relationships rather than forcing a single global model onto heterogeneous spatial data.

The GWR equation for our model is as follows:

$$y_i(LNMEDHVAL) = \beta_{i0} + \beta_{i1}(LNNBELPOV) + \beta_{i2}(PCTBACHMOR)$$
$$\beta_{i3}(PCTSINGLES) + \beta_{i4}(PCTVACANT) + \varepsilon_i =$$
$$\beta_{i0} + \sum_{k=1}^{m} \beta_{ik}\, x_{ik} + \varepsilon_i$$

The use of subscript *i* in this equation indicates that the regression model describes the relationship between y and predictors at various locations *i*. The β_{i0} is the intercept at location *i*, the predicted value of y when all predictors equal 0. Each β_{ik}, where k ranges from 1 to *m* (number of predictors), shows how much the dependent variable is expected to change with a one-unit increase in predictor *k* at location *i,* holding all other variables constant. The x_{ik} represents the actual values of predictor variables at location *i*. Finally, ε_i is the error term at location *i*, or the difference between the observed value of y_i and the predicted value of *y*.

Local regression requires the existence of multiple locations to operate. GWR will use other observations (locations) in the dataset to run the regression, and observations that are closer to location *i* receive greater weights. The weights will vary based on location *i,* and observations closer to *i* have a stronger influence on the estimation parameters for location *i.*

There are two ways to weigh these nearby locations, fixed bandwidth and adaptive bandwidth. Fixed bandwidth implies that thhe number of observations will vary around each point *i,* but the bandwidth distance *h* and the area will remain constant. Adaptative bandwidth implies the opposite in that the number of observations will remain constant, but the area *h* will not be the same. For our purposes, we will be using adaptive bandwidth. Adaptive bandwidth ensures each local regression uses a consistent number of observations regardless of local spatial density, which produces more reliable and meaningful local coefficients across the entire study area. Adaptive bandwidth is more appropriate for this problem because the spatial distribution of block groups across Philadelphia is highly uneven. Some neighborhoods have many block groups concentrated in small areas while others are more spread out. Fixed bandwidth would include too many observations in dense areas and too few in sparse areas, making adaptive bandwidth the better choice

The assumptions for OLS still hold in GWR in that we must have normality of residuals, homoscedasticity, and no multicollinearity. GWR may give us some issues with multicollinearity if we have two or more variables that have similar patterns of clustering in a certain region. We can use the condition number to in the local results table to identify when results are unstable due to local multicollinearity. If the condition number is greater than 30, it is best not to trust the results for that feature.

A feature of GWR outputs is the absence of traditional p-values and t-statistics for local coefficients. Because there is more than one set of parameters associated with each regression point, as well as one set of standard errors, then there are potentially hundreds of thousands of tests that would be required to determine whether parameters are actually locally significant. With a 4-predictor model estimated at 2,000 regression points, there would be 10,000 significance tests (5 per point—1 for the intercept and 4 for the predictors). Even if none of the relationships were truly significant, we would expect 500 of these tests to return a significant result simply by chance. This makes it virtually impossible to distinguish genuine local effects from false positives, rendering traditional p-values unreliable and potentially misleading in the GWR context.

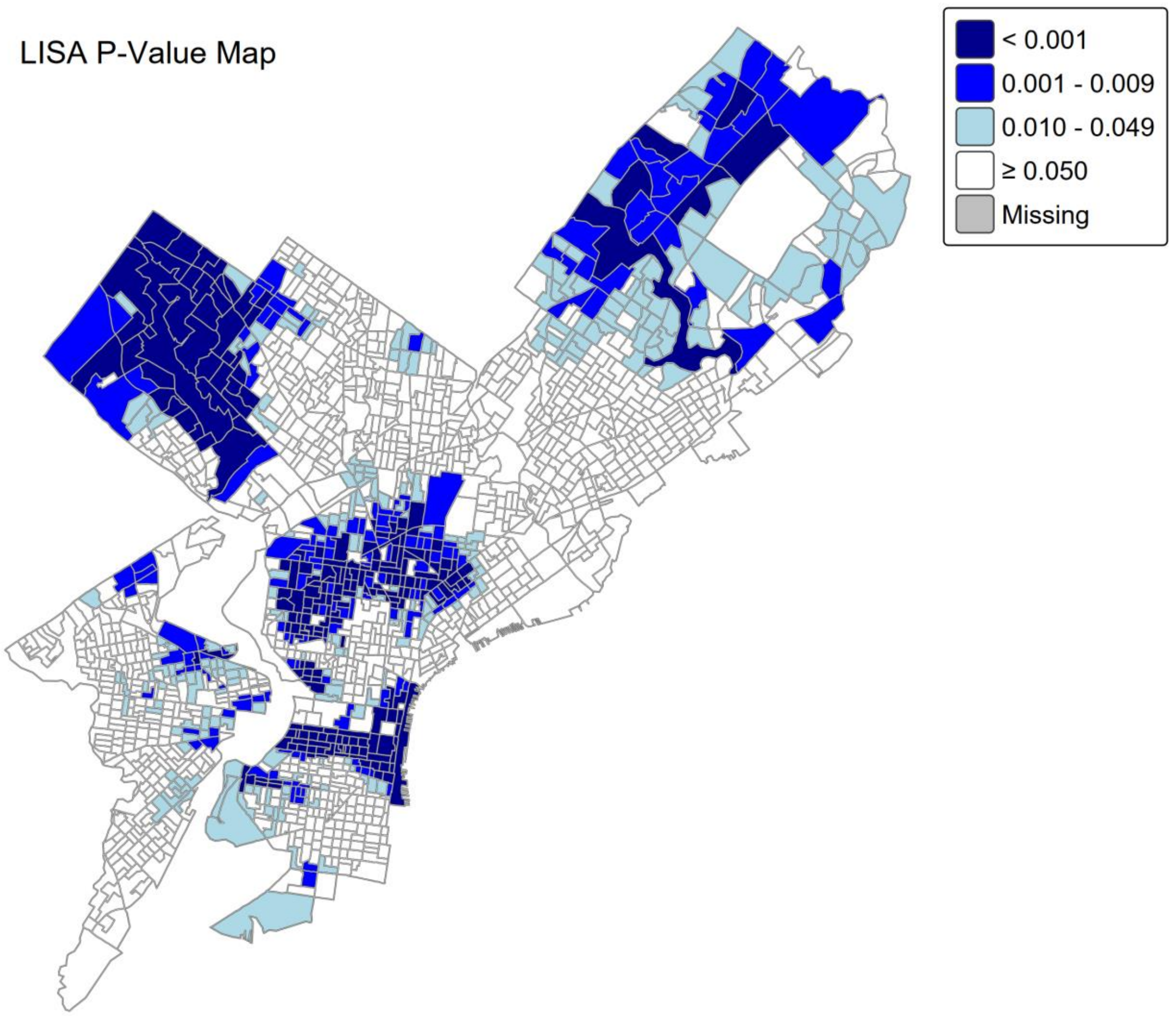
3. Results

3.1 Spatial Autocorrelation

```
        Monte-Carlo simulation of
        Moran I

data:  regress_data$LNMEDHVAL
weights: queenlist
number of simulations + 1: 1000

statistic = 0.79356, observed
rank = 1000, p-value <
0.00000000000000022
alternative hypothesis: two.sided
```

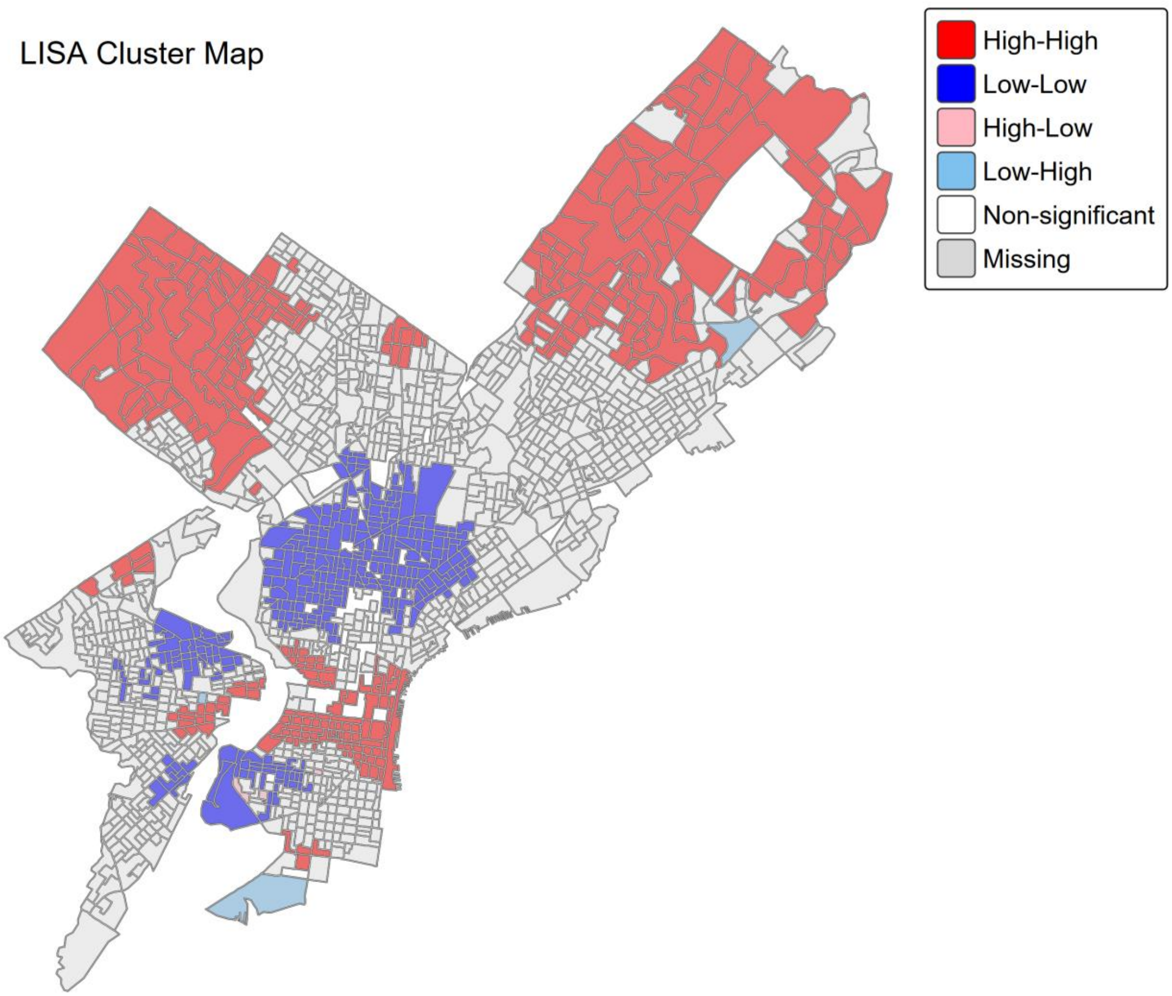
Figure 1: LNMEDHVAL Moran's I Output

The 0.79356 Moran's I value indicates very strong positive spatial autocorrelation, meaning many of the block groups are clustered together with high-income areas aside other high-income areas, low-income areas aside other low-income areas, or inverses where a low-income block is surrounded by high-income neighbors and vice versa. However, this statistic does not provide any visual depth to affirm those patterns in different Philadelphia areas. In addition, the pseudo p-value is astronomically small, making the results statistically significant in comparison to the 999 random permutations it was observed with. That is, there is a less than 1% chance that Philadelphia's positive spatial autocorrelation occurred by random chance, so the null hypothesis of no spatial autocorrelation is rejected as this indicates strong positive spatial autocorrelation.



Map 1: LISA P-Values

According to the above map 1 of LISA's p-values, many statistically significant block clusters are deeper in northeast and northwest of Philadelphia, as well as parts of West Philadelphia in the Mantua neighborhood above Market St, North Philadelphia in the Strawberry Mansion and Kensington neighborhoods, and in Center City where the business districts are. This presents the regions of block groups and their neighbors within Philadelphia County that are outliers relative to the rest of the block groups that are colored in white. Whether these are high-priced areas near other high ones or low-priced areas near low or some other spatial context is not indicated on this map, but the map below.



Map 2: LISA Cluster

Aligning with the p-Value map 1 above, map 2 presents the same clustered spaces in Northeast and Northwest Philadelphia as well as within Center City in the county's urban core, but this map gives context that they are high-high clusters compared to the rest of the block groups. The rest of the significant block group clusters are where North Philadelphia and West Philadelphia areas are, which possess low-low clusters; this reflects long-term disinvestment within these neighborhoods. However, the University City area in West Philadelphia holds three private non-profit anchor institutions, Drexel University, Saint Joseph's University, and University of Pennsylvania, which has high-high clusters that are almost sandwiched by the Mantua, Mill Creek, and Powelton neighborhoods in the north and a portion of Kingsessing in the south.

3.2 A Review of OLS Regression and Assumptions: Results

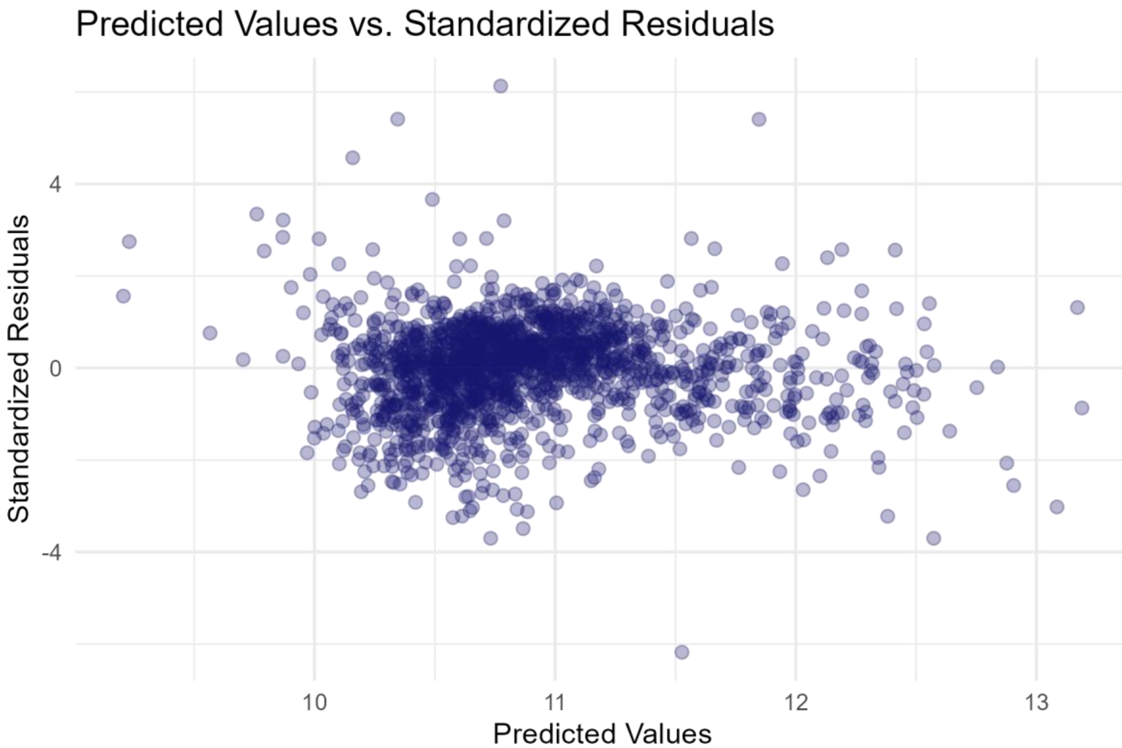
```
Call:
lm(formula = LNMEDHVAL ~ LNNBELPOV + PCTBACHMOR + PCTSINGLES +
    PCTVACANT, data = regress_data)

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 < 0.0000000000000002 ***
LNNBELPOV   -0.0789035  0.0084567  -9.330 < 0.0000000000000002 ***
PCTBACHMOR   0.0209095  0.0005432  38.494 < 0.0000000000000002 ***
PCTSINGLES   0.0029770  0.0007032   4.234             0.0000242 ***
PCTVACANT   -0.0191563  0.0009779 -19.590 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3665 on 1715 degrees of freedom
Multiple R-squared:  0.6623,	Adjusted R-squared:  0.6615
F-statistic: 840.9 on 4 and 1715 DF,  p-value: < 0.00000000000000022
```

Table 1: OLS Regression Output

Table 1 shows the OLS regression results, where all the predictors are statistically significant with $p < 0.001$. According to these results, the adjusted R^2 is 0.6615, so the model explains 66.15% of the variance in LNMEDHVAL.

Test	Statistic	df	p-value
Breusch-Pagan	113.19	4	0.00000000000000022
Koenker-Bassett	42.868	4	0.00000001102
White's	43.94	4	0

Table 1.1: Heterokedasticity Test—OLS Breusch-Pagan, Koenker-Bassett, and White's

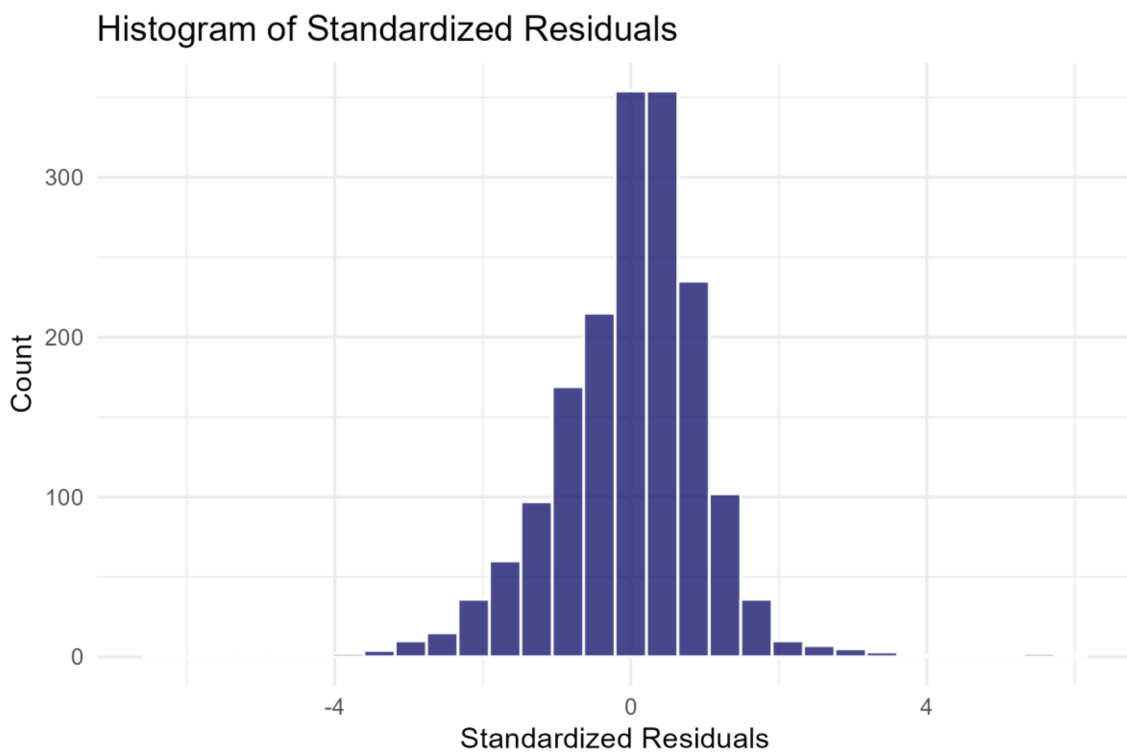


Graph 1: OLS Predicted Values vs. Standardized Residuals Scatter Plot

First, according to table 1.1, all tests mentioned analyze heteroscedasticity in a regression model. The most important part of the table is the p-values, which are all less than 0.001, making the test statistics significant, and favoring the alternative hypothesis that heteroscedasticity exists. The test statistics themselves are not as important as the p-values, but they are different values because they have different mathematical considerations when it pertains to X^2, so they should not be compared directly with one another. Of course, this means that the model violates one of the OLS assumptions for normality of residuals, especially since all three tests are consistent with one another, which does not align with the residual by predicted plot from the previous HW 1 paper as Graph 1 visually implies normality of residuals.

Test	**X-Squared**	**df**	**p-value**
Jarque Bera	778.96	2	0.00000000000000022

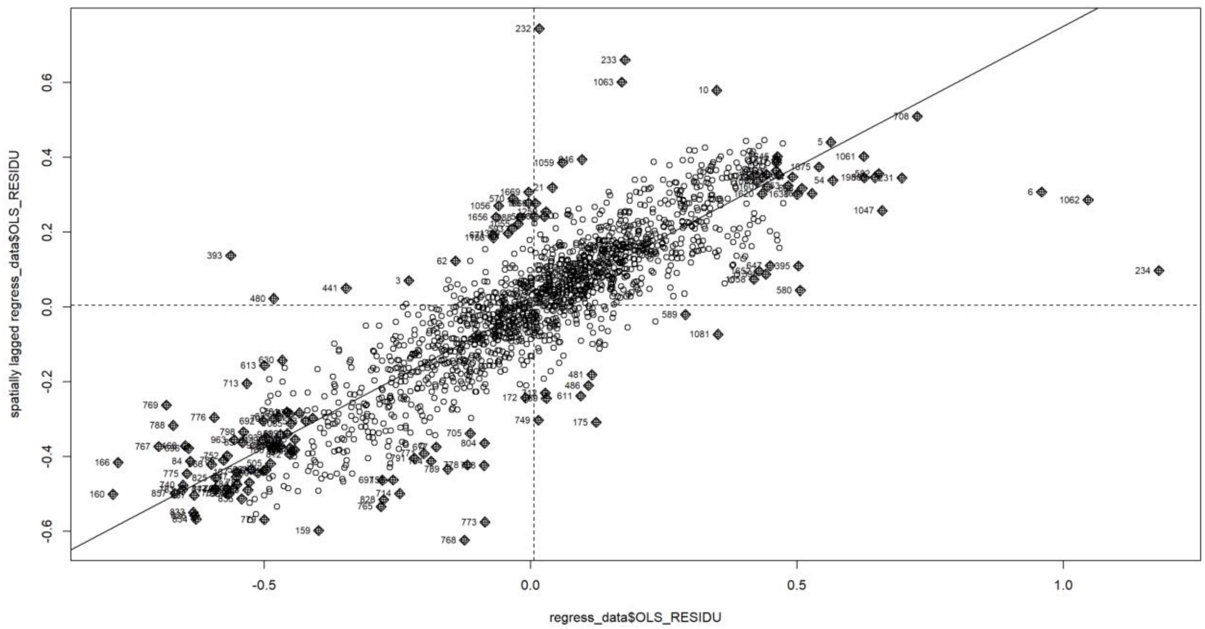
Table 1.2: Normal Errors Test—OLS Jarque Bera



Graph 2: OLS Standardized Residuals Histogram

According to the results of the Jarque Bera test, the p-value is leagues less than 0.001, meaning we reject the null hypothesis in favor of the alternative hypothesis. The reports in table 1.2 do not align with graph 2 from the previous HW 1 paper that visually indicates normal residuals.

Clearly, from the above tests, this indicates that mathematical testing methods to analyze heterocedasticity and normality of residuals versus just visualization have extremely deep gaps, meaning the tests are much more perceptive, or sensitive, to unseen patterns humans cannot inspect from plots and histograms. Apparently, this “issue is especially common with large samples, where even tiny, random, and inconsequential deviations from an expected distribution often lead to statistically significant differences from that distribution” (Shatz, 2024), and this paper deals with almost 1,720 observations. It seems that even though some OLS assumptions are implied to be violated at a more granular level, on a practical level this may be inconsequential.



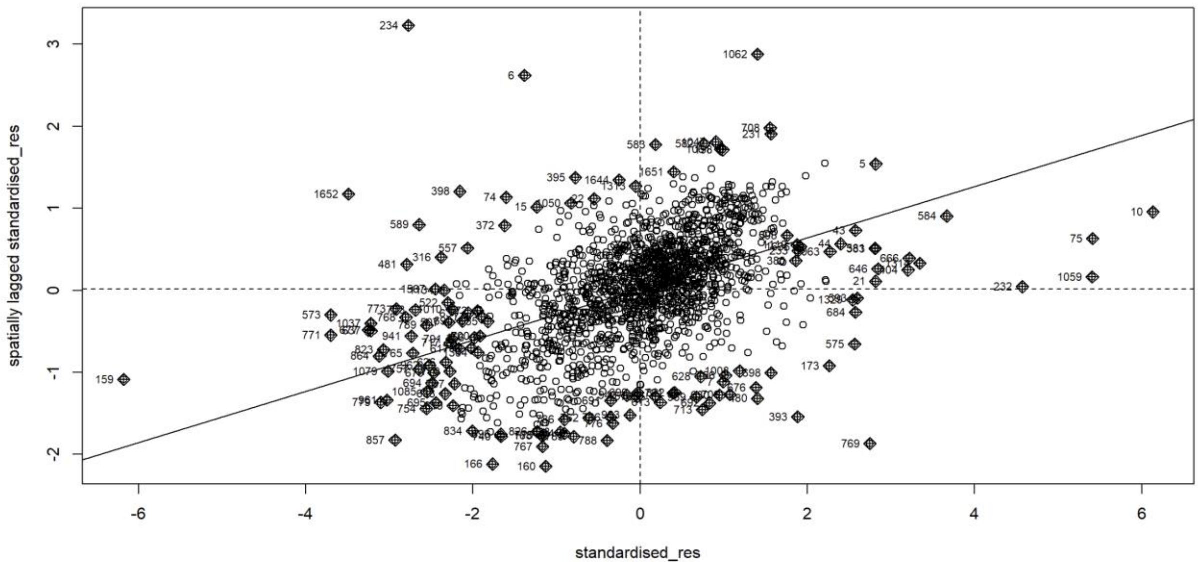
Graph 3: OLS Residuals vs. Weighted OLS Residuals Moran's I Scatter Plot

```
Call:
lm(formula = moran_RESIDU$x ~ moran_RESIDU$wx)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72972 -0.06072 -0.00055  0.05815  1.08051 

Coefficients:
                Estimate Std. Error t value            Pr(>|t|)    
(Intercept)     0.001938   0.002879   0.673               0.501    
moran_RESIDU$wx 0.999994   0.013893  71.980 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1194 on 1718 degrees of freedom
Multiple R-squared:  0.751,	Adjusted R-squared:  0.7508 
F-statistic:  5181 on 1 and 1718 DF,  p-value: < 0.00000000000000022
```

Table 1.3: OLS Residuals vs. Weighted OLS Residuals Regression Output

The statistics in table 1.3 indicate, from the adjusted R2 value 0.7508, that 75.08% of the variation in the spatially lagged OLS residuals are explained by the OLS residuals. In addition, the associated t-value with the OLS residuals predictor variable and its very small p-value also indicate that there is a less than 1% chance of this being random. Graph 3 depicts the OLS residuals on the x-axis and the weighted OLS residuals on the y-axis, and this Moran's I plot visually indicates that there is indeed spatial autocorrelation, with most points aligned to form a positive-sloped pattern in low-low and high-high quadrants, as indicated by the steep line.

However, while this plot tells us that the data is spatially autocorrelated, it does not tell us what type of spatial regression model to use.



Graph 4: OLS Residuals Moran’s I Scatter Plot

```
        Monte-Carlo simulation of Moran I

data:  standardised_res
weights: queenlist
number of simulations + 1: 1000

statistic = 0.3124, observed rank = 1000, p-value < 0.00000000000000022
alternative hypothesis: two.sided
```

Table 1.3: OLS Residuals Moran’s I Output

The 0.3124 test statistic indicates spatial autocorrelation exists among the OLS's residuals as Moran’s I is significant due to the extremly small p-value. This means that the data is misspecified as an OLS model and that another statistical method should be considered. Again, the very small p-value signals to reject the null hypothesis. Also, when compared to the Moran’s I and β coefficient of the weighted (spatially lagged) residuals, the outputs align with one another and the spatial pattern from their Moran’s I plots are very similar. As a result, the next steps in this paper will run the Spatial Lag Model, Spatial Error Model, and Geographically Weighted Regression (GWR).

3.3 Spatial Lag and Spatial Error Regression Results

```
Call:
lagsarlm(formula = LNMEDHVAL ~ LNNBELPOV + PCTBACHMOR + PCTSINGLES +
    PCTVACANT, data = regress_data, listw = queenlist)

Residuals:
      Min        1Q    Median
-1.655421 -0.117248  0.018654
       3Q       Max
 0.133126  1.726436

Type: lag
Coefficients: (asymptotic standard errors)
               Estimate
(Intercept)  3.89845505
LNNBELPOV   -0.03405466
PCTBACHMOR   0.00851381
PCTSINGLES   0.00203342
PCTVACANT   -0.00852940
             Std. Error  z value
(Intercept)  0.20111357  19.3843
LNNBELPOV    0.00629287  -5.4116
PCTBACHMOR   0.00052193  16.3120
PCTSINGLES   0.00051577   3.9425
PCTVACANT    0.00074367 -11.4694
                         Pr(>|z|)
(Intercept) < 0.0000000000000022
LNNBELPOV           0.00000006246
PCTBACHMOR  < 0.0000000000000022
PCTSINGLES          0.00008063502
PCTVACANT   < 0.0000000000000022

Rho: 0.6511, LR test value: 911.51, p-value: < 0.000000000000000222
Asymptotic standard error: 0.01805
    z-value: 36.072, p-value: < 0.000000000000000222
Wald statistic: 1301.2, p-value: < 0.000000000000000222

Log likelihood: -255.74 for lag model
ML residual variance (sigma squared): 0.071948, (sigma: 0.26823)
Number of observations: 1720
Number of parameters estimated: 7
AIC: 525.48, (AIC for lm: 1435)
LM test for residual autocorrelation
test value: 67.737, p-value: 0.00000000000000022204
```

Table 2: Spatial Lag Regression Output

According to table 2, the weighted response variable, LNMEDHVAL, termed “Rho” in the output, is 0.6511, which indicates positive relationships. This means that high median household values tend to cluster and rise with one another. It is associated with a low p-value as well, meaning it is significant, and this significance is shared with the other predictors LNNBELPOV, PCTBACHMOR, PCTSINGLES, and PCTVACANT. Compared to the original OLS regression output, the results have large differences when observing their AICs because the Spatial Lag regression’s is 525.48 versus 1,435, this is a 909.52-unit difference in AIC is 909.52 and shows that the Spatial Lag regression is a better model. Also, observing the sigma squared versus the sigma from the OLS results, the former has lower residual variance than OLS, this means that the Spatial Lag model explains more variation, and is supported by the LM test being statistically significant with the alternative hypothesis being that adding spatial lag was an improvement.

Going deeper into the other predictors, LNNBELPOV's coefficient means that a 1% increase in a block's poverty is associated with a 0.0340% decrease in median home value, PCTBACHMOR's coefficient means that a 1% increase in a block's bachelor's or more degree attainment is associated with a 0.009% increase in median home value, PCTSINGLES's coefficient means that a 1% increase in a block's detached single family home is associated with a 0.002% increase in median home value, and lastly PCTVACANT's coefficient means that a 1% increase in percent vacancies in a block is associated with a 0.009% decrease in its median home value. Compared to the original OLS regression, these results are much smaller predicted changes as LNNBELPOV is –0.079, PCTBACHMOR is 0.021, PCTSINGLES is 0.003, and PCTVACANT is –0.019 for the coefficient values.

Test	Statistic	df	p-value
Breusch-Pagan	210.76	4	0.00000000000000022
Koenker-Bassett	51.411	4	0.0000000001832

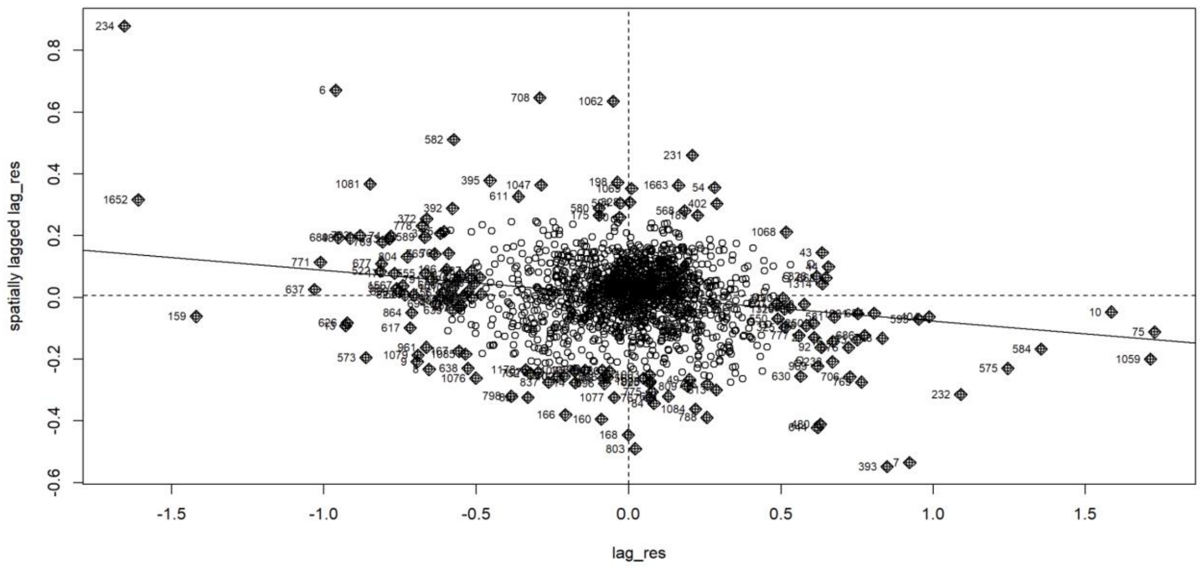
Table 2.1: Heteroskedasticity Test—Spatial Lag Breusch-Pagan

According to table 2.1 above, the Breusch-Pagan and Koenker-Basset modification still indicate that the model is heterscedastic with the small p-values, and to additionally note, the Spatial Lag model's Koenker-Bassett p-value is actually even smaller than the OLS's value, which may mean that the Spatial Lag regression is more heteroscedastic. It is possible this might reflect luxury versus market-value housing markets within a block that make its heterscedasticity a bit more pronounced when accounting for spatial aspects in the lag regression.

Model	AIC	SC	Log Likelihood	Likelihood Ratio	p-value
OLS	1434.9867	1467.6871	-711.4933	911.5100	0.00000000000000022
Spatial Lag	525.4800	563.6305	-255.7400	911.5100	0.00000000000000022

Table 2.2 OLS vs. Spatial Lag AIC, SC, Log Likelihood, Likelihood Ratio

As mentioned previously, the AIC and SC differences between the OLS and Spatial Lag regressions is quite enormous, and that the latter is a much better model fit for the variables because it is preferable to have a lower AIC and SC values. Elsewhere, "maximum likelihood picks the values of the model parameters that make the data 'more likely' than any other values of the parameters would make them" (Dattalo, 2009), and the higher value is more desirable, so the Spatial Lag model being less negative is preferred. Now, when viewing the likelihood ratio and its associated p-value, it signals which of the two models performs better than the other, and in this case it is the Spatial Lag model because the small p-value guides us to reject the null hypothesis, which is that the Spatial Lag regression does not perform better than the OLS regression in favor of the alternative hypothesis, which states that the Spatial Lag regression is doing a better job than the OLS regression.



Graph 2.3: Spatial Lag Residuals Moran’s I Scatter Plot

Per graph 2.3, after including spatial lag, it looks like the residuals are more balanced and cloudy compared to the OLS’s Moran’s I scatter plot’s, which had a very positive spread. Accounting for spatial lag corrected the residuals to a more random form, but there are more distant points in the high-low and low-high quadrants, which show the isolated outlier blocks likely in transitioning zones of high-highs and low-lows. With all this being said, it can be concluded that when comparing OLS versus Spatial Lag regressions, that Spatial Lag outperforms by a large margin.

```
Call:
errorsarlm(formula = LNMEDHVAL ~ LNNBELPOV + PCTBACHMOR + PCTSINGLES +
    PCTVACANT, data = regress_data, listw = queenlist)

Residuals:
      Min        1Q    Median
-1.926477 -0.115408  0.014889
       3Q       Max
 0.133852  1.948664

Type: error
Coefficients: (asymptotic standard errors)
               Estimate
(Intercept) 10.90643423
LNNBELPOV   -0.03453408
PCTBACHMOR   0.00981293
PCTSINGLES   0.00267792
PCTVACANT   -0.00578308
             Std. Error  z value
(Intercept)  0.05346779 203.9814
LNNBELPOV    0.00708933  -4.8713
PCTBACHMOR   0.00072896  13.4615
PCTSINGLES   0.00062083   4.3134
PCTVACANT    0.00088670  -6.5220
                          Pr(>|z|)
(Intercept) < 0.0000000000000022
LNNBELPOV       0.00000110881774
PCTBACHMOR  < 0.0000000000000022
PCTSINGLES      0.00001607388517
PCTVACANT       0.00000000006937

Lambda: 0.81492, LR test value: 677.61, p-value: < 0.000000000000000222
Asymptotic standard error: 0.016373
    z-value: 49.772, p-value: < 0.000000000000000222
Wald statistic: 2477.2, p-value: < 0.000000000000000222

Log likelihood: -372.6904 for error model
ML residual variance (sigma squared): 0.076551, (sigma: 0.27668)
Number of observations: 1720
Number of parameters estimated: 7
AIC: 759.38, (AIC for lm: 1435)
```

Table 3: Spatial Error Regression Output

According to table 3, the spatial error variable for LNMEDHVAL, termed "Lambda" in the output, is 0.81492, which indicates that block neighbors significantly affect median home values. This $p < 0.001$ statistical significance is shared with the other predictors LNNBELPOV, PCTBACHMOR, PCTSINGLES, and PCTVACANT. Compared to the original OLS regression output, the results have large differences when observing their AICs because the Spatial Error regression's is 759.38 versus 1,435, this is a 675.62-unit difference in AIC and shows that the Spatial Error regression is a better model. Also, observing the sigma squared versus the sigma from the OLS results, the former has lower residual variance than OLS, this means that the Spatial Error model explains more variation.

Going deeper into the other predictors, LNNBELPOV's coefficient means that a 1% increase in a block's poverty is associated with a 0.0350% decrease in median home value, PCTBACHMOR's coefficient means that a 1% increase in a block's bachelor's or more degree attainment is associated with a 0.01% increase in median home value, PCTSINGLES's coefficient means that a 1% increase in a block's detached single family home is associated with a 0.003% increase in median home value, and lastly PCTVACANT's coefficient means that a 1% increase in percent vacancies in a block is associated with a 0.006% decrease in its median home value. Compared to the original OLS regression, these results are much smaller predicted changes as LNNBELPOV is –0.079, PCTBACHMOR is 0.021, PCTSINGLES is 0.003, and PCTVACANT is –0.019 for the coefficient values.

Test	Statistic	df	p-value
Breusch-Pagan	210.76	4	0.00000000000000022
Koenker-Bassett	51.411	4	0.0000000001832

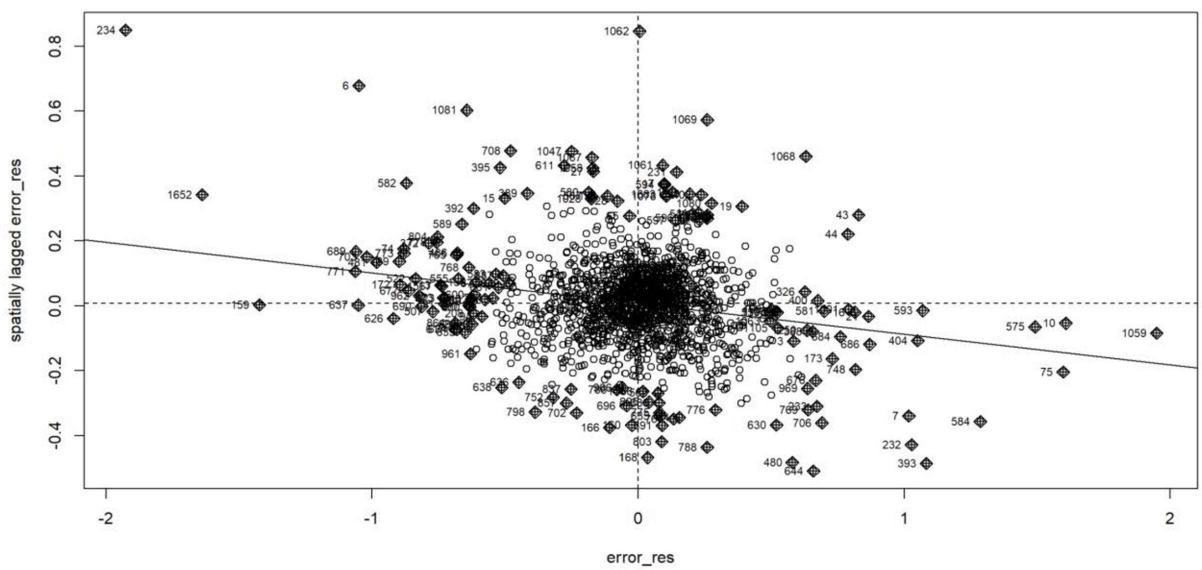
Table 3.1: Heteroskedasticity Test—Spatial Error Breusch-Pagan

According to table 3.1 above, the Breusch-Pagan and Koenker-Basset modification still indicate that the model is heterscedastic with the small p-values, and to additionally note, the Spatial Error model's Koenker-Bassett p-value is actually even smaller than the OLS's value, which may mean that the Spatial Error regression is more heteroscedastic. It may be the case that some unobservable error makes its heterscedasticity a bit more pronounced when accounting for spatial error aspects in the regression.

Model	AIC	SC	Log Likelihood	Likelihood Ratio	p-value
OLS	1434.9867	1467.6871	-711.4933	677.6100	0.00000000000000022
Spatial Error	759.3807	797.5313	-372.6904	677.6100	0.00000000000000022

Table 3.2: OLS vs. Spatial Error AIC, SC, Log Likelihood, Likelihood Ratio

As mentioned previously, the AIC and SC differences between the OLS and Spatial Error regressions is also large like between OLS and Spatial Lag, and that Spatial Error is a much better model fit for the variables because it is preferable to have a lower AIC and SC value. Elsewhere, the higher value log likelihood is more desirable, so the Spatial Error model being less negative is preferred. Now, when viewing the likelihood ratio and its associated p-value, it signals which of the two models performs better than the other, and in this case it is the Spatial Error model because the small p-value guides us to reject the null hypothesis, which is that the Spatial Error regression does not perform better than the OLS regression, in favor of the alternative hypothesis, which states that the Spatial Error regression is doing a better job than the OLS regression.



Graph 3.3: Spatial Error Residuals Moran's I Scatter Plot

In graph 3.3, after including spatial errors, it looks like the residuals are cloudy compared to the OLS's Moran's I scatter plot's, which had a very positive spread, and similar to the plot for Spatial Lag regression. It looks like accounting for spatial error corrected the residuals to a more random form with similar distant points in the high-low and low-high quadrants as graph 2.3, again, which show the isolated outlier blocks likely in transitioning zones of high-highs and low-lows. So, it can be concluded that when comparing OLS versus Spatial Error regressions, that Spatial Error outperforms as well.

Model	**AIC**	**SC**
Spatial Lag	525.4800	563.6305
Spatial Error	759.3807	797.5313

Table 3.4: Spatial Lag vs. Spatial Error AIC

Now, when comparing the Spatial Lag and Spatial Error regressions with one another in table 3.4, it looks like Spatial Lag dominates Spatial Error as a better performing model for the variables with a 233.9007-unit and 233.9008-unit differences between the two AICs and SCs, respectively. This shows that the median house values in different blocks are more internally related to the model, meaning that house values are more strongly influenced by other house values that neighbor them, hence, Spatial Lag regression pairing spatial weights with the dependent variable $\rho W y$. If it were the other case, that would mean the house values are more strongly influenced by some unobserved external variable, hence, Spatial Error regression pairing spatial weights with the errors $\lambda W \mu$.

3.4 Geographically Weighted Regression Results

```
Call:
gwr(formula = LNMEDHVAL ~ LNNBELPOV + PCTBACHMOR + PCTSINGLES +
    PCTVACANT, data = regress_spatial, gweight = gwr.Gauss, adapt = bandwidth,
    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
```

Table 4: GWR with Adaptive Bandwidth

```
Call:
gwr(formula = LNMEDHVAL ~ LNNBELPOV + PCTBACHMOR + PCTSINGLES +
    PCTVACANT, data = regress_spatial, bandwidth = bandwidth_fixed,
    gweight = gwr.Gauss, hatmatrix = TRUE, se.fit = TRUE)
Kernel function: gwr.Gauss
Fixed bandwidth: 2863.492
Summary of GWR coefficient estimates at data points:
                   Min.    1st Qu.     Median    3rd Qu.       Max.  Global
X.Intercept.  9.9111181 10.7329171 10.9397426 11.1639961 14.1200780 11.1138
LNNBELPOV    -0.4449896 -0.0737744 -0.0433084 -0.0171174  0.1491700 -0.0789
PCTBACHMOR   -0.0860913  0.0118750  0.0168149  0.0213553  0.0306653  0.0209
PCTSINGLES   -0.0238330 -0.0073895 -0.0025702  0.0040499  0.0189995  0.0030
PCTVACANT    -0.0469926 -0.0137374 -0.0088796 -0.0038447  0.0778856 -0.0192
Number of data points: 1720
Effective number of parameters (residual: 2traceS - traceS'S): 346.718
Effective degrees of freedom (residual: 2traceS - traceS'S): 1373.282
Sigma (residual: 2traceS - traceS'S): 0.2785229
Effective number of parameters (model: traceS): 255.6033
Effective degrees of freedom (model: traceS): 1464.397
Sigma (model: traceS): 0.2697189
Sigma (ML): 0.2488723
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 700.3524
AIC (GWR p. 96, eq. 4.22): 352.3471
Residual sum of squares: 106.5324
Quasi-global R2: 0.843808
```

Table 5: GWR with Fixed Bandwidth

Table 4 shows the GWR results with adaptive bandwidths and table 5 shows the GWR results with fixed bandwidths. Before the results, the bandwidth differences must be known; an adaptive bandwidth accounts for more variation with Philadelphia's blocks, in other words, it adjusts for the distance among the numerous block observations, this means that Philadelphia's density is accounted for from suburban edges to urban cores; a fixed bandwidth does not account for that variation in Philadelphia, it assumes that the blocks are evenly spaced at a "fixed" distance, hence, no accounting for density.

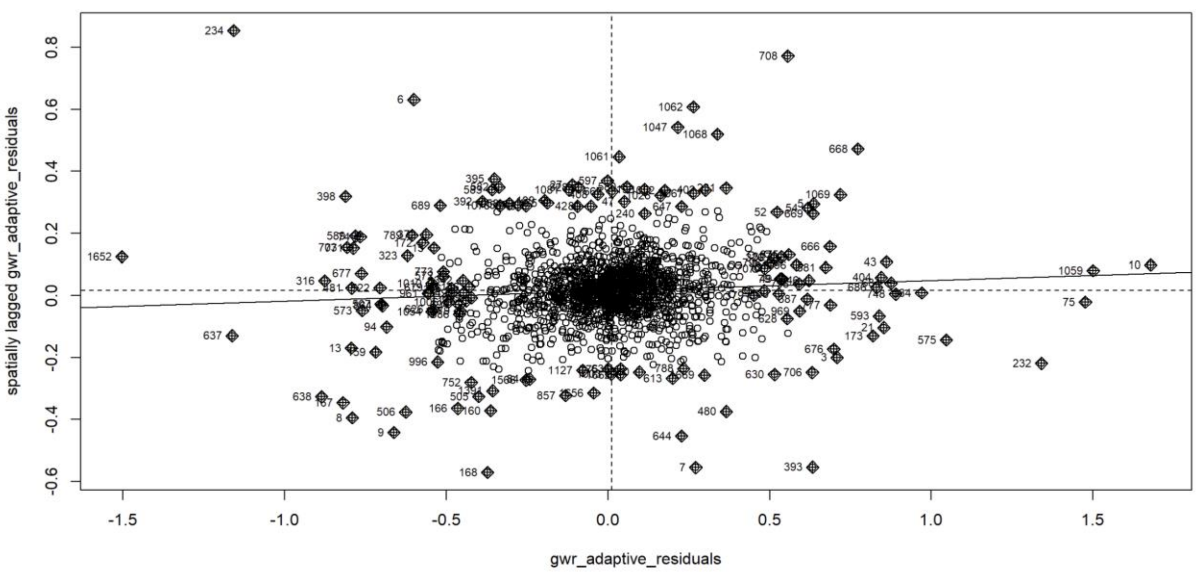
Taking a look at the predictors, because GWR accounts for spatial non-stationarity, the results provide summary statistics for each explanatory variable's coefficients. With that in mind, we look to the median to provide the most general picture of a block group accounting for space, with Adaptive and Fixed GWRs mentioned respectively: LNNBELPOV at –0.040 and –0.043 indicates poverty generally decreases median house value, PCTBACHMOR at 0.015 and 0.017 indicates higher percentages of bachelor's or more degrees increases median house value, PCTSINGLES at –0.002 and –0.003 indicates that higher percentages of single detached family homes actually decrease median house value (this is slightly surprising, but the relatively high range is unsurprising, so it may vary in some blocks' housing markets), PCTVACANT at –0.009 and –0.009 indicates that higher percentages of vacancies are associated with a decrease in median house value, and this statistic has a pretty large range like PCTSINGLES, indicating varying block housing markets. Also, LNNBELPOV having a max value in the positives may allude to blocks that are undergoing gentrification.

Clearly, the Adaptive GWR has greater explanatory power in comparison. When looking at the quasi-global R2, the Adaptive GWR explains 84.79% whereas the Fixed GWR explains 84.38% of the variance in the data. Other performance metrics are the sigma values at 0.266 and 0.270, with the Adaptive GWR possessing the smaller one, which means that it has smaller variance in its residuals, and this is the same case with the residual sum of squares being 103.7, less than its counterpart's 106.5. The AIC value is discussed below.

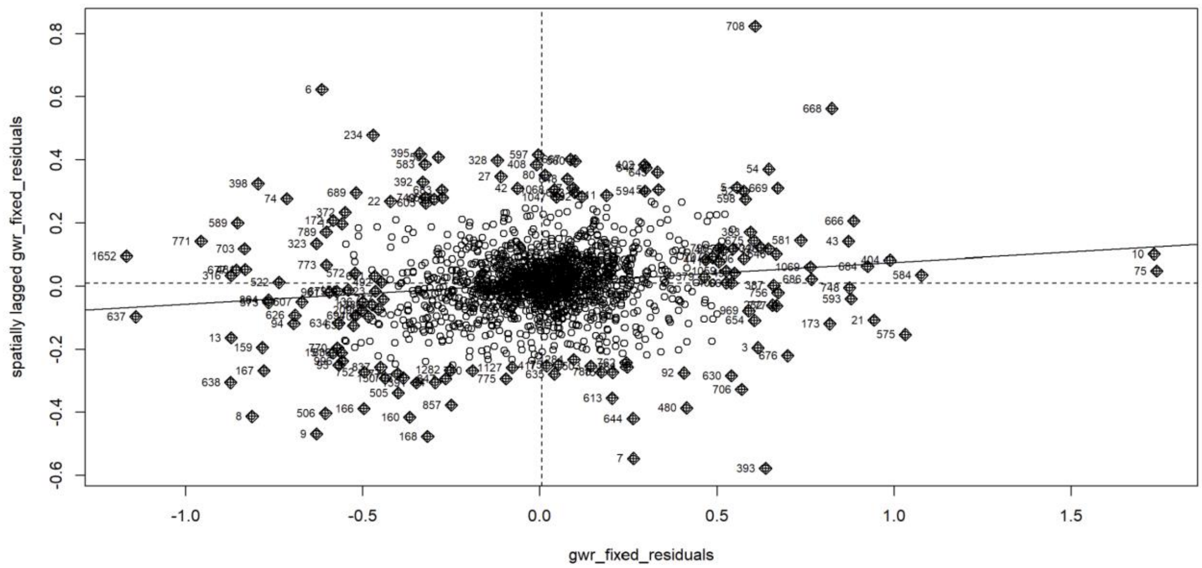
Model	AIC
Adaptive GWR	308.7123
Fixed GWR	352.3471
OLS	1434.9867
Spatial Lag	525.4800
Spatial Error	759.3807

Table 6: Model AIC Comparisons

According to table 6, among the model AICs the GWR with adaptive bandwidths performed much better than the rest of the models, and by a significant margin to the OLS, Spatial Lag, and Spatial Error models, although the gap is less pronounced between the GWR with the adaptive bandwidths versus the GWR with the fixed bandwidths.

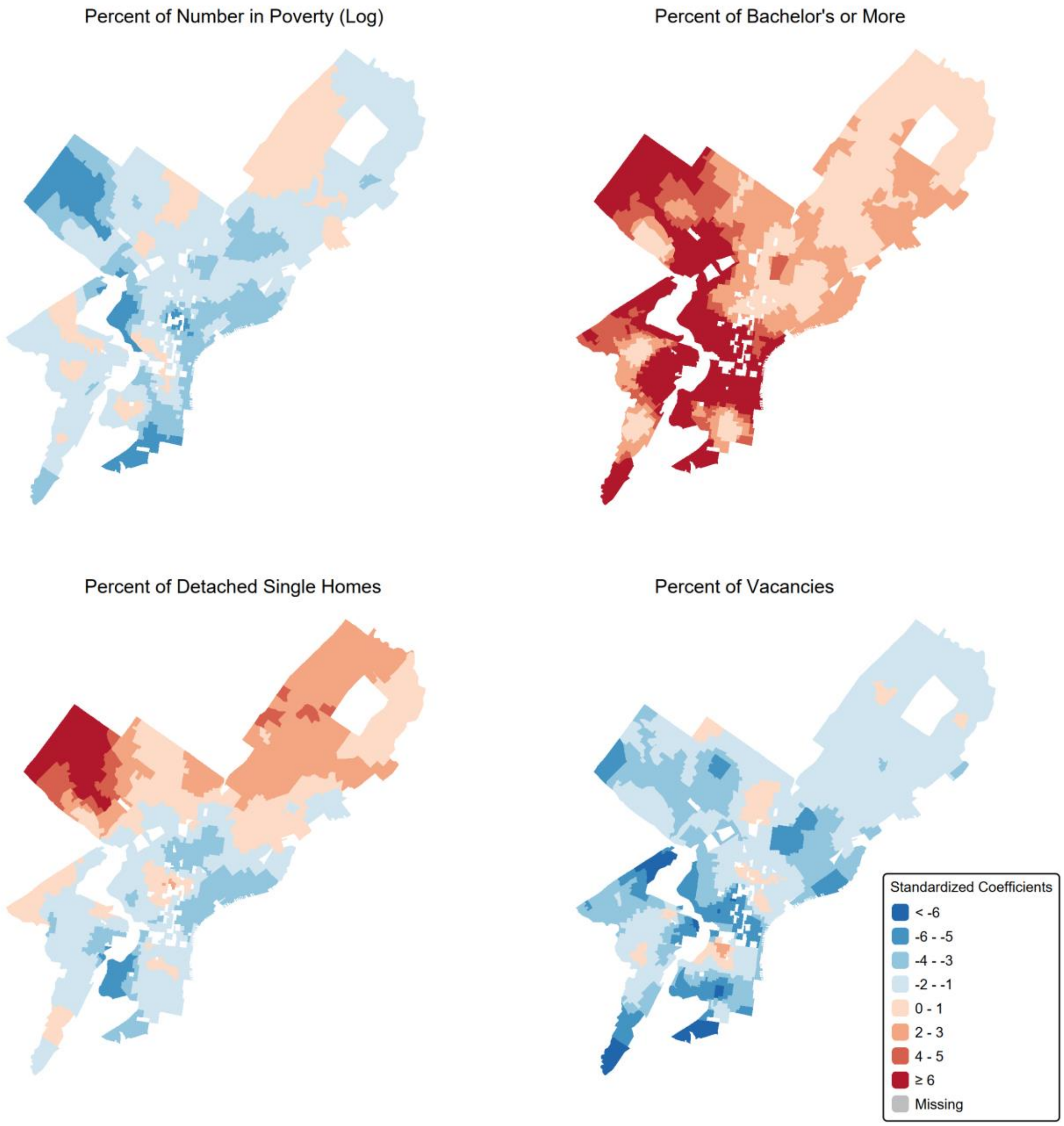


Graph 4.1: GWR Adaptive Bandwidth Residuals Plot



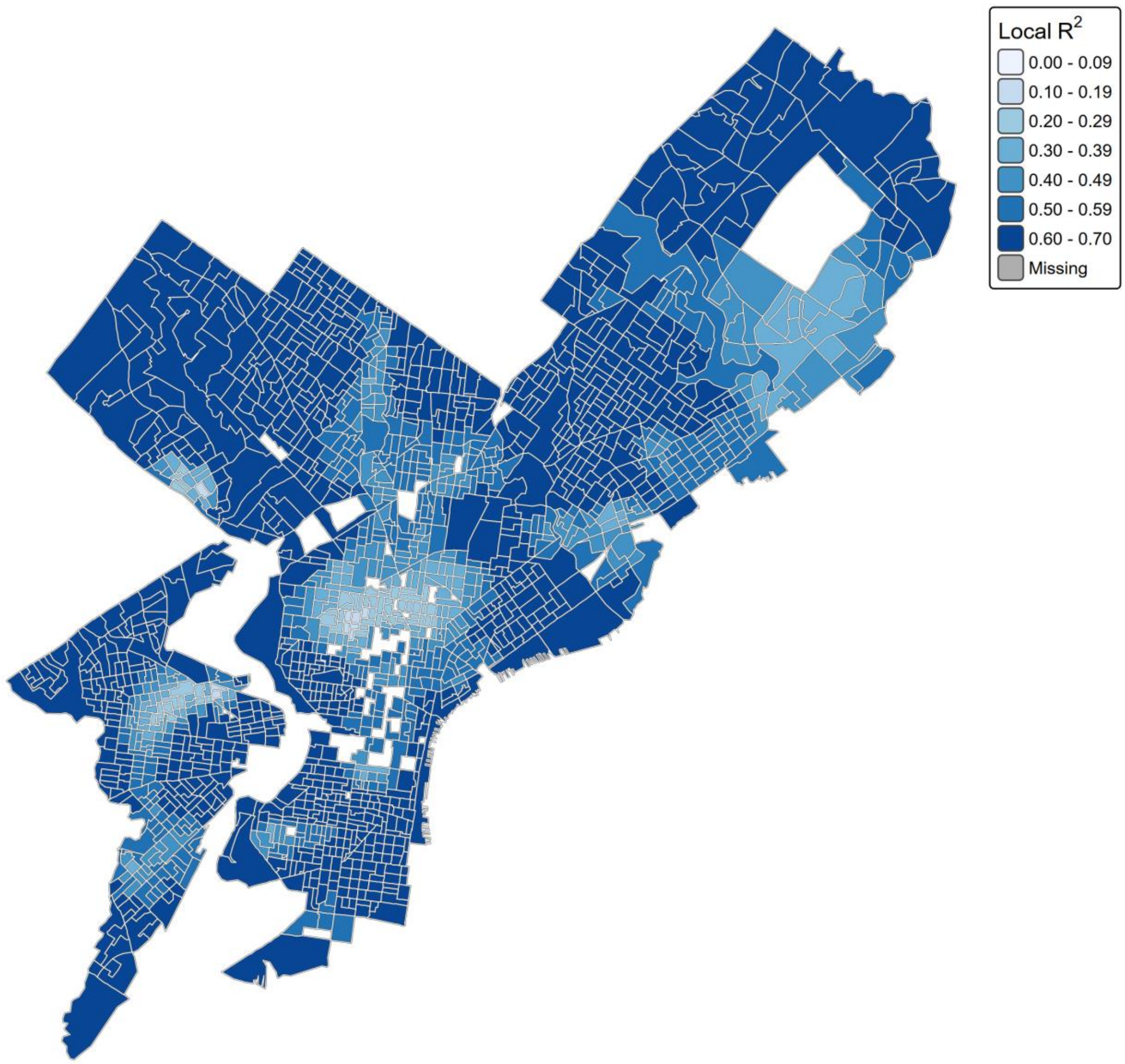
Graph 5.1: GWR Fixed Bandwidth Residuals Plot

According to graph 4.1 for the Adaptive GWR and graph 5.1 for the Fixed GWR, it looks like both GWRs have more residual randomization compared to OLS in graph 4, Spatial Lag in graph 2.3, and Spatial Error in graph 3.3. It also looks like Spatial Lag and Spatial Error seemed like they overcorrected comparatively as they have a negative slope going through the high-low and low-high quadrants. Obviously, OLS's residuals remain the most prominently slanted in the high-high and low-low quadrants. And when comparing both the GWR graphs, even if somewhat marginal, it looks as if the GWR with adaptive bandwidths is the closest to a horizontal line than the GWR with fixed bandwidths. This coincides with the AIC that essentially declares that the Adaptive GWR is the best fit model compared to all the others in this paper that were conducted. This makes sense, because Philadelphia's blocks are not uniformly drawn.



Map 3: Standardized Coefficient Maps—LNNBELPOV, PCTBACHMOR, PCTSINGLES, PCTVACANT

Map 3 showcases four different predictors and their coefficients divided by the standard error. When looking at the number of those in poverty, it looks like the case tends to be that as the variable increases that it decreases the median house values in their areas. This is with the exception of certain areas in pockets of West, Northeast, and portions of Central Philadelphia, where the increase in poverty increases median house value—this could mean a few things, like the areas are undergoing gentrification or that the market demand for housing exceeds the housing that is built. Moving to the percent of bachelor's degrees or more, it looks like higher education attainment only has a positive relationship with median house value (for the adaptive bandwidths), and the strongest regions are in West Philadelphia where the university anchor institutions are, Central Philadelphia, Northwest Philadelphia, and especially surrounding the Schuylkill River which may indicate a higher price tag for a waterfront property as well as general areas where students or graduates tend to reside. With the percentage of detached single-family homes, it looks like they have an increasing impact with median household value in the Philadelphia suburbs as opposed to the denser portions of Philadelphia as the blocks move inland, this may be because the denser cores of Philadelphia highly prefer efficient land use / housing for its citizens in such populous areas (i.e. multi-family, mid- to high-density residential units). Percent vacancies seem to have a similar visual pattern to the number of individuals in poverty in that it has an overall effect of decreasing median house value as the percentage of vacancies rise. This may also be linked to gentrification as vacant lots tend to be lower cost and generally sought after by property developers and property investors who buy up underutilized parcels for future reinvestment (Moselle, 2025) (Sarkis, 2018).



Map 4: Local R^2 Map

According to map 4's legend, the lightest blues indicate the weakest local fit for each block to the darkest blues indicate the strongest local fit for each block. With that in mind, it can be said that this model performs quite poorly in portions of Northeast, North, and West Philadelphia. Now, referring toward the beginning at map 2 of the LISA clusters, these happen to be regions that harbor low-low clusters when it comes to median house value—this model is underperforming for poorer areas and poorer Philadelphians with values generally from 0 to 0.39 and performing very well for wealthier Philadelphians with values generally from 0.50 to 0.70; this is a large disparity when the former block groups can only be explained around 39% max of the time versus the majorly 60% to 70% of the time for wealthier block groups with this model.

4. Discussion

This report conducted a comparative analysis of four regression model in predicting median house values (MEDHVAL) across Philadelphia Block Groups: Ordinary Least Squares (OLS), Spatial Lag, Spatial Error, and Geographically Weighted Regression (GWR). While OLS regression provides valuable baseline insights, the Moran's I ≈ 0.31 which violates the model assumptions of independence of residuals. As a result, the next steps were to run the Spatial Lag Model, Spatial Error Model, and Geographically Weighted Regression (GWR). Both Spatial Lag and spatial error corrected the residuals to a more random form, but Spatial Lag outperformed Spatial Error due to spatial correlation. Adaptive GWR outperformed all other models, achieving the highest R^2 (≈ 0.8484), the lowest AIC (≈ 308.7123), and sigma values at 0.266, which means that it has smaller variance in its residuals. The Adaptive GWR also provided insight into how each predictor's influence varied across neighborhoods, revealing spatial patterns that global models were not able to detect.

Despite the GWR model performing best for the data among the four approaches tested, it is not without limitations. The GWR model is sensitive to multicollinearity, particularly when explanatory variables exhibit spatial clustering, a factor not fully accounted for in this analysis. The variables used from the previous paper indicated multicollinearity between the variables representing higher educational attainment (PCTBACHMOR) and poverty rates (NBELPOV100). Additionally, the percentage of single-family homes (PCTSINGLES) was not normally distributed, and the GWR residuals themselves also deviated from normality, as shown in the histogram.

Weighted (or spatially lagged) residuals are produced by averaging the residuals of each spatial unit's neighbors according to a spatial weights matrix. They are typically used in diagnostic analyses, such as Moran's I scatterplots, to identify spatial autocorrelation in model residuals. In contrast, residuals from a spatial lag model represent the model's actual errors (Observed Predicted Values – Actual Predicted Values) after explicitly accounting for the influence of neighboring outcome values through a spatially lagged dependent variable. Therefore, it is important to use precise terminology: "spatially lagged residuals" refer to the spatially weighted averages of residuals from another model, not the residuals generated by a spatial lag model.

While ArcGIS provides tools for conducting Geographically Weighted Regression (GWR), it presents several notable limitations. ArcGIS has limited flexibility in customizing key model parameters such as bandwidth selection, kernel type, and weighting functions. Also, ArcGIS does not produce comprehensive diagnostic outputs, including local standard errors, residuals, or detailed measures of model fit. Without these diagnostics, it is impossible to identify issues such as local multicollinearity or replicate analyses to meet research transparency standards.

5. Citations

Anselin, L. (2020). Visualizing Spatial Autocorrelation. Global spatial autocorrelation (1). https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html

Dattalo, Patrick, 'Statistical Alternatives and Supplements to Random Assignment', Strategies to Approximate Random Sampling and Assignment, Pocket Guides to Social Work Research Methods (New York, 2009; online edn, Oxford Academic, 1 Feb. 2010), https://doi.org/10.1093/acprof:oso/9780195378351.003.0005

MetricGate. (n.d.). Koenker–Bassett Test. https://metricgate.com/docs/koenker-bassett-test/

Miller, H. J. (2004). Tobler's First Law and Spatial Analysis. Annals of the Association of American Geographers, 94(2), 284–289. http://www.jstor.org/stable/3693985

Moselle, A. (2025, October 31). What Philly's corporate investors mean for renters and homebuyers. WHYY; WHYY. https://whyy.org/articles/philadelphia-single-family-homes-corporate-landlords/

Price, G. (2002). (Tech.). Heteroscedasticity: Testing and Correcting in SPSS. Retrieved from https://geocities.restorativland.org/Heartland/4205/SPSS/HeteroscedasticityTestingAndCorrectingInSPSS1.pdf

Sarkis, R. (2018, August 27). Why now is the time to buy vacant land in Philadelphia. Bizjournals; Philadelphia Business Journal. https://www.bizjournals.com/philadelphia/news/2018/08/27/cre-data-reonomy-why-buy-philly-vacant-land-bank.html

Shatz, I. (2024). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. Behav Res 56, 826–845. https://doi.org/10.3758/s13428-023-02072-x

Statistics How To. (n.d.-a). Jarque Bera Test. https://www.statisticshowto.com/jarque-bera-test/

Statistics How To. (n.d.-b). White Test: Definition, Examples. https://www.statisticshowto.com/white-test/