Zhenzhao Xu, Hanpu Yao, and Ben Aiken
Statistical and Data Mining Methods - MUSA 500
Professor Eugene Brusilovskiy
Assignment 2

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

Introduction

This investigation examines the spatial relationship between Median House Value of Philadelphia census tracts and several predictor variables. The predictor variables include poverty levels, education, single family homes, and vacancy. In Assignment 1, the relationship between the same dependent variable and predictors was explored using an OLS (Ordinary Least Squares) Regression. The model had an R^2 value of 0.6615 indicating that 66% of the variance of the model was explained by the four predictors, all of which were included when a stepwise regression was completed. Although this R^2 may be strong in the social sciences, the model is called into question since several OLS assumptions were violated such as the linearity of the dependent variable and predictor relationships and, after investigating maps of the variables, the observations may not independent due to some clustering of values in neighborhoods of Philadelphia.

Given the spatial component of this data and OLS assumption violations, an investigation into spatial autocorrelation is warranted. Three spatial investigations, including spatial lag, spatial error, and geographically weighted regression, will be explored. The results, errors, and assumptions of each method will be discussed and compared with the OLS model to determine which method produces the strongest model.

Methods

a) A Description of the Concept of Spatial Autocorrelations

Spatial Autocorrelations examine the clustering, randomness, or dispersion of observations in a geographic context. Clustering is the occurrence of multiple observations with the same or similar values close together geographically. Dispersion is just the opposite: the occurrence when similar observations are spaced apart. Finally, randomness lacks any clear clustering or dispersion of like values. The First Law of Geography as articulated by geographer Waldo Tobler is, *"Everything is related to everything else, but near things are more related than distant things."*[1]

This spatial relationship is often measured using Moran's I. Positive Moran's I values indicate clustering of similar observations. Conversely, negative Moran's I values correspond with dispersion or repulsion of observations. Moran's I values close to 0 suggest little to no autocorrelation of similar

[1] TOBLER, W. R. (1970). "A computer movie simulating urban growth in the Detroit region". Economic Geography, 46(2): 234-240.

observations, or randomness. Values do not exclusively range from -1 to 1; however, values close to 1 have strong positive autocorrelation and values close to -1 have strong negative autocorrelation. Moran's I is determined using the following formula:

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

Here, $\bar{X}$ is the mean of the variable X. X_i and X_j are the variable values at locations i and j respectfully. n is the number of observations. And finally, w_{ij} is a weight index for location i relative to j. This weight index is determined using a weight matrix.

In this investigation a queen weight matrix with one order of contiguity is used for all three spatial models. A weight matrix is a matrix of n x n dimensions where each observation's neighbors are indicated with 1s, and non-neighbors are indicated with 0s. In this case, all census tracts that share a border or a point with the observation are considered queen neighbors. The order of contiguity refers to how many layers of queen neighbors are used; for example, two orders of contiguity would indicate that the immediate queen neighbors and the queen neighbors of the immediate queen neighbors are included. Since just one order is used only the immediate queen neighbors are included. Often, statisticians will explore more than one weight matrix when investigating spatial autocorrelation to determine if a spatial relationship is exclusive to one matrix; however, for this investigation there will only be one matrix used throughout.

In order to determine the significance of spatial autocorrelation, as measured by Moran's I, a significance test is completed using the following hypotheses:

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

In this test, the Moran's I of the existing arrangement is measured. Then the locations of observations are held constant, but the values are randomly shuffled to different locations, and the Moran's I is calculated for this new arrangement. The values are shuffled for a given number of arrangements, or permutations, with the Moran's I determined each time. In this investigation, 999 permutations are taken each time a Moran's I is calculated. Next, the Moran's I of the real arrangement is compared with the other 999 permutations by ranking them in descending order. A pseudo p-value is calculated using the rank, r, and total number of permutations, n_p, below:

$$p - value = \frac{r}{n_p} \quad (2.)$$

Thus, if the real Moran's I is ranked first among all permutations it would have a p-value of $\frac{1}{1,000}$ or 0.001. Since 0.001 is less than 0.05, it is deemed significant, and the null hypothesis is rejected indicating that there is spatial autocorrelation. If, on the other hand, the Moran's I is in the middle of the rankings, say 349 for example, the p-value would be $\frac{349}{1,000}$ or 0.349 meaning the null hypothesis is not rejected and there may not be spatial autocorrelation. It is important to note that this test can get different p-values each time it is completed since the shuffling of values is purely random and the permutations may be different each time even if the same number of permutations is used.

While Moran's I and the significance test explained above quantify and examine the significance for all observations, or the global spatial autocorrelation, the Local Indices of Spatial Autocorrelation (LISA) establishes a measure for each individual observation, or rather a local Moran's I. A similar permutation method is used to determine the significance of this local value; however, instead of shuffling all values, the value of the observation being examined is held constant and all other values are shuffled. Once again, the Moran's I of the observation is taken for each permutation and compared with the real Moran's I for significance.

b) A Review of OLS Regression and Assumptions

The relationship investigated in this report was also explored in Assignment 1 using the Ordinary Least Squares regression method that does not take spatial autocorrelations into account. Refer to Assignment 1 for a detailed description of OLS and explanation of the results for this data set. In short, OLS develops a model that calculates the amount by which the dependent variable, y, changes when a predictor, X_i changes by 1 unit. The equation below is the result of an OLS regression:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \qquad (3.)$$

Here, β_0 is the y-intercept, β_i is the coefficient variable for predictor of X_i and ε is the residual, or the difference between the predicted and observed values for y.

The OLS model relies on several assumptions including linear relationships between the dependent variable and all predictors, no multicollinearity between predictors, randomness and normal distribution of residuals, homoscedastic residuals, and at least 10 observations per predictor.

Additionally, OLS regression assumes that all observations are independent or random, which, in the case of spatial data, means there is no spatial autocorrelation. Given the First Law of Geography, there is often some spatial autocorrelation and therefore spatial data rarely upholds this assumption. The Moran's I of the residuals of an OLS model and the Moran's I significance can be examined to determine if there is spatial autocorrelation, therefore violating this assumption of OLS.

Furthermore, spatial autocorrelation can also be tested by regressing the residuals of an OLS model on the residuals of neighboring observations, in this case the queen neighboring block groups. If the

residuals of neighbors, or lagged residuals, are similar, then the OLS assumption of independence is violated due to spatial autocorrelation. This is determined by running this regression of residuals on lagged residuals and examining the coefficient of the lagged residuals, often called *rho* and denoted ρ. Sometimes, this coefficient is referred to as *lambda* and denoted λ, which is the case in the GeoDa software used in this investigation. Alternatively, it is referred to as *Slope b* on some scatterplots in this report. Regardless of the name, when this coefficient has a p-value less than 0.5 it is significant and therefore the assumption of independent observations is violated as there is spatial autocorrelation present.

GeoDa is also used to determine the spatial lag and spatial error regressions. GeoDa has nonspatial capabilities as well and can test OLS regression assumptions such as homoscedasticity and normality of errors. Homoscedasticity is tested in GeoDa using the Breusch-Pagan Test, the Koenker-Bassett Test, or the White Test. These tests use the following simple hypotheses:

H_0: No heteroscedasticity
H_a: Heteroscedasticity

If the GeoDa returned p-value from these tests is significant, or less than 0.5, then the null hypothesis is rejected and the OLS residuals are deemed heteroscedastic and the assumption of homoscedasticity is violated.

GeoDa can also test for the normality of errors using the Jarque-Bera test. In this case the hypotheses below are used:

H_0: Residuals have a normal distribution
H_a: Residuals do not have a normal distribution

The null hypothesis is rejected if the p-value is significant, or less than 0.5. When the null hypothesis is rejected, there is non-normality and the assumption that the residuals have a normal distribution is violated.

c) Spatial Lag and Spatial Error Regression

To deal with spatial autocorrelation, spatial lag and spatial error regression models are introduced. In this project, we will be using GeoDa for running spatial lag and spatial error regressions.

Spatial lag model assumes that the value of the dependent variable at one location is associated with the values of that variable in nearby locations, so the model includes the spatial lag of the dependent variable as a predictor.

$$LNMEDHVAL = \rho W_{LNMEDHVAL} + \beta_0 + \beta_1 LNNBELPOV + \beta_2 PCTSINGLES + \beta_3 PCTBACHMOR + \beta_4 PCTVACANT \quad (4.)$$

Here, $W_{LNMEDHVAL}$ is the spatial lag of variable $LNMEDHVAL$, which is the average value of $LNMEDHVAL$ at its neighboring spatial units, and in this project, neighbors are as defined as queen neighbors. ρ is the coefficient of the y-lag variable $W_{LNMEDHVAL}$, and ρ is constrained between -1 and 1. β_0 is the value of y-intercept. β_i is the coefficient variable for predictor of x_i (for example, β_1 is the coefficient variable of $LNNBELPOV$) and it shows the true average change in the dependent variable $LNMEDHVAL$ associated with a 1-unit increase in the predictor x_i (holding all other variables constant). ε is the residual, which is the difference between $\hat{y}$ (predicted y) and y (observed y).

Spatial error model assumes that the residual at one location is associated with residuals at nearby locations, so the model includes the spatial lag of the residuals as a predictor, and the model is regressed using the equation below:

$$LNMEDHVAL = \beta_0 + \beta_1 LNNBELPOV + \beta_2 PCTSINGLES + \beta_3 PCTBACHMOR + \beta_4 PCTVACANT + \lambda W_\varepsilon + u \quad (5.)$$

$$\varepsilon = \lambda W_\varepsilon + u \quad (6.)$$

Here, W_ε is the is the spatial lag of variable ε, which is the average value of ε at its neighboring spatial units. u is random noise. λ is the coefficient of the lag variable W_ε, and λ is constrained between -1 and 1. ε is λ times of spatial lag of residuals (W_ε) plus random noise u. β_0 is the value of y-intercept. β_i is the coefficient variable for predictor of x_i.

Assumptions that are needed for OLS are still needed for both spatial lag and spatial error regression models (except that of spatial independence of observations).

The goal of spatial lag and spatial error regression is to minimize the to take into consideration the fact there may be spatial dependencies in the residuals or in the data, that is, we want the value of the residuals' global Moran's I to be close to 0. Otherwise, when spatial autocorrelation is present, values of a variable in nearby areas are related to each other and are not independent and the independent assumption is violated. Another effect of these models is that there are usually less heteroscedastic in the residuals.

In this project, we will compare the results of spatial lag regression with OLS and the results of spatial error regression with OLS. And several criteria are considered to decide whether the spatial models perform better than OLS. One criterion is Akaike Information Criterion (AIC) / Schwarz Criterion (SC). AIC and SC are measures of the goodness of fit of an estimated statistical model. They are relative measures of the information, and the lower the AIC and SC, the better the fit. Another criterion is called log likelihood. Log likelihood is a method of fitting a statistical model to the data and estimating model parameters. The higher the log likelihood, the better the model fit. The other criterion is likelihood ratio test. It compares the OLS model with the spatial model and uses hypothesis testing to determine which one is better. The hypotheses are listed as follows.

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

If $p < 0.05$, we reject the null hypothesis, and state that the spatial lag (error) model is doing a better job than the OLS model, and vice versa.

Another way of comparing OLS results with spatial lag and spatial error results is by looking at the Moran's I of regression residuals. Significance test and randomly shuffle are applied to determine if there is spatial autocorrelation in the residuals.

H_0: No spatial autocorrelation in residuals
H_a: Positive spatial autocorrelation in residuals
H_b: Negative spatial autocorrelation in residuals

By testing the significance, we can compare the p-value of H_0 and the value of Moran's I of the models. The model with the highest p-value and the lowest absolute Moran's I is the better one. It's worth noting that both or neither of the lag and error models account for spatial autocorrelation is possible.

d) Geographically Weighted Regression

In this part, the Geographically Weighted Regression (GWR), is examined in ArcGIS. GWR is a modelling technique designed to deal with spatial non-stationarity, meaning the modeled relationships are not constant across space. Since the regression models above (OLS, Spatial Lag and Spatial Error Regression) assume that we are dealing with spatial stationarity, which is usually not the case in practice, it is possible for Simpson's paradox to occur. The Simpson's paradox is a statistic phenomenon that a relationship appears in several groups of data but disappears when the groups are combined. The influence of Simpson's paradox can be problematic, so instead of a global regression, a local regression is worth investigating.

For each observation, or location, $i = 1 \ldots \mathrm{n}$, the equation of the GWR model is:

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i \qquad (7.)$$

$+ \varepsilon_i$ #7.

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i \qquad (7.)$$

where y_i indicates the dependent variable of location i, β_{i0} is the interception, β_{ik} is the coefficient related to the weight between location i and k, ε_i is the error term.
The estimator of the model takes the form:

$$\beta\,(u) = (X^T W(u) X)^{-1} X^T W(u) y \qquad (8.)$$

where $W(u)$ is the square matrix of weights relative to the position of u in the study area, $X^T W(u) X$ is the geographically weight matrix and y is the vector of the values of the dependent variable.

$$W(u) = \begin{bmatrix} w_1(u) & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & w_n(u) \end{bmatrix} \quad (9.)$$

To run the local regression for each observation, GWR needs other observations in the dataset, and the observations nearer to location i are given larger weight. As a result, observations further from i have less influence on the estimation of the parameter of location i, consistent with the First Law of Geography.

There are two ways to define how which observations to consider: fixed bandwidth or adaptive bandwidth. Fixed bandwidth holds the bandwidth distance unchanged, but the number of observations vary according to each location. Alternatively, the number of observations surrounding location i remain fixed but the bandwidth distance varies in adaptive bandwidth. We will use adaptive bandwidth in this regression because the distribution of observations varies across space.

GWR relies on the same assumptions as OLS, such as normality of residuals, homoscedasticity, and no multicollinearity. In GWR, multicollinearity manifests as spatial clustering of an explanatory variable, so the variable is redundant because it tells the same story at every location. The condition number in the attribute table is an indicator of local multicollinearity. If it's larger than 30 or is Null, the result is unstable due to local multicollinearity.

Different from OLS, the p-value is not a part of GWR result. In OLS, the statistical significance is tested using a t-test and the associated p-value is an indicator of significance of each parameter. However, in GWR, there is a set of parameters for every regression point. The number of tests will be exponentially larger than in OLS if a t-test is used to determine whether the parameters are locally significant.

Results

a) Spatial Autocorrelation

Spatial Autocorrelation was first examined with the global Moran's I using the queen matrix in GeoDa. In Assignment 1 it was determined that the dependent variable is normal after a log transformation, therefore in this investigation the Median House Value will once again be log transformed. Figure 1 displays the scatterplot of lagged LNMEDHVAL and LNMEDHVAL revealing that the Moran's I is 0.794. Since this value is close to 1, it appears that there is strong positive spatial autocorrelation in the dataset; however, significance must be checked.

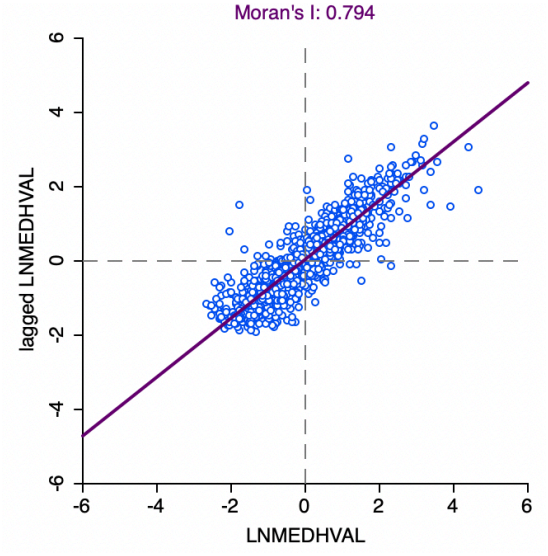

Figure 1 – Scatterplot of Lagged Dependent Variable

Figure 2 displays the results of the Moran's I significance test. The locations are shuffled for 999 permutations and the histogram shows the resulting Moran's I values. The histogram shows that the actual Moran's I value is the highest, or ranked first, of the 1,000 total values making the p-value 0.001, which is less than 0.05 indicating that this Moran's I value is in fact significant. Therefore, the null hypothesis is rejected and there is significant positive spatial autocorrelation for the dependent variable LNMEDHVAL.

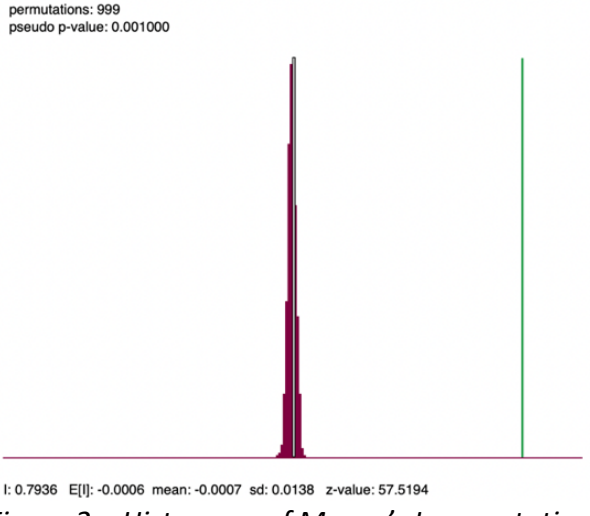

Figure 2 – Histogram of Moran's I permutations

Next, the Local Moran's I, or LISA, is explored in Figures 3 and 4. First, Figure 3 shows the Moran's I significance of each census tract and Figure 4 shows the clusters of these values. Much of the city does not have a significant Local Moran's I, indicating that the Moran's I and the dependent variable is inconsistent in nearby census tracts and there is not spatial autocorrelation locally.

However, in several parts of the city the Moran's I and the dependent variable are clustered. In Figure

4, the dark red and dark blue areas indicate positive spatial autocorrelation and therefore relatively homogenous areas. Dark red census tracts have high LNMEDHVAL and their queen neighbors are similarly high. Dark blue tracts have low LNMEDVAL and their queen neighbors are similarly low. These dark red areas are expensive parts of the city, in Center City, Northeast Philadelphia, and Chestnut Hill. The dark blue tracts are cheaper areas including West Philadelphia, North Philadelphia, and the Greys Ferry and Kingsessing areas.

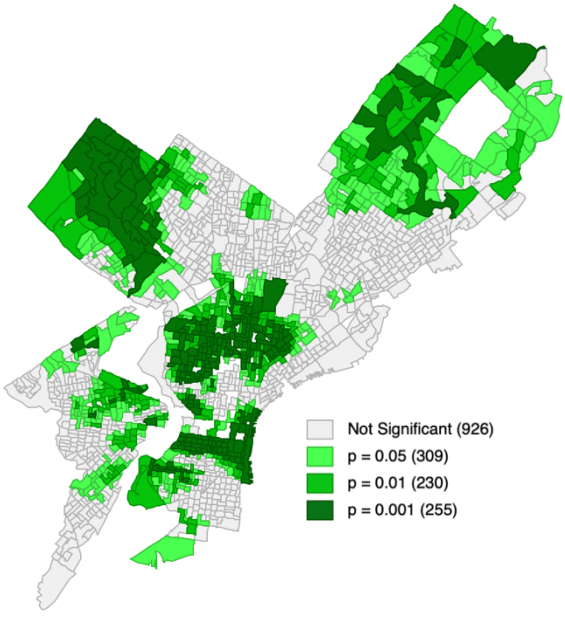

Figure 3 – Local Moran's I Significance Map

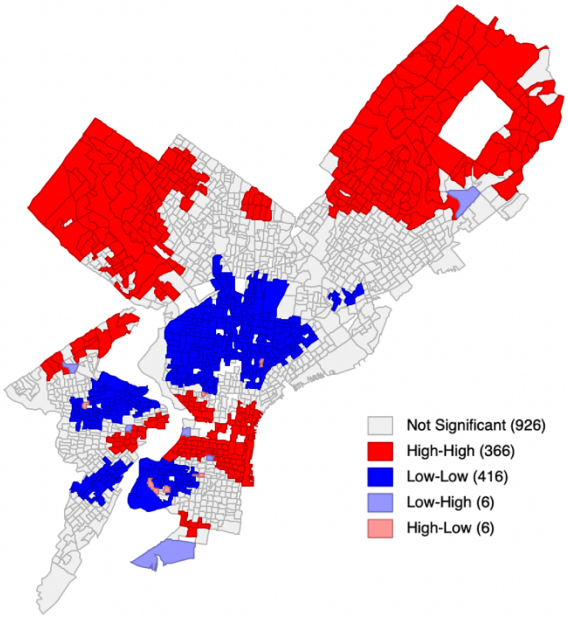

Figure 4 – Local Moran's I Cluster Map

On the other hand, the light colors (light red and light blue) have negative spatial autocorrelation and are abnormal or outliers for their surrounding census tracts. Light red tracts have high observed LNMEDVAL, but their queen neighbors are low; in other words, these tracts are abnormally expensive for their area. And light blue tracts have low observed LNMEDVAL, but their queen neighbors are high – these tracts are abnormally inexpensive for their area.

b) A Review of OLS Regression and Assumptions: Results

Table 1 below shows the OLS output from GeoDa. The p-probability values indicate that all four predictors are significant. Additionally, the R^2 value, or the coefficient of determination, is 0.6623, and the adjusted R^2, or $R^2{}_{adj}$, is 0.6615 indicating that 66% of the variance in LNMEDHVAL is explained by the model. For a more detailed explanation of OLS results, refer to Assignment 1.

Table 1 – OLS Regression Output

SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION

Data set:	Regression Data		
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
LNNBELPOV	-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

SPECIFICATION ROBUST TEST

TEST	DF	VALUE	PROB
White	14	111.3224	0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE

FOR WEIGHT MATRIX: queen-hw2-1
(row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.3131	22.3763	0.00000
Lagrange Multiplier (lag)	1	930.5854	0.00000
Robust LM (lag)	1	441.1036	0.00000
Lagrange Multiplier (error)	1	491.0070	0.00000
Robust LM (error)	1	1.5252	0.21684
Lagrange Multiplier (SARMA)	2	932.1106	0.00000

The heteroscedasticity tests displayed in Table 1 include the Breusch-Pagan test, the Koenker-Bassett test, and the White test. In all three cases the p-value is significant since it less than 0.05, meaning the null hypothesis is rejected and there is a problem with heteroscedasticity, violating a key OLS assumption.

The Jarque-Bera test in Table 1 also has a probability less than 0.05 indicating that the null hypothesis of normality of residuals is rejected for the alternative hypothesis. This non-normality of residuals is problematic, violating another OLS assumption.

Figure 5 displays the OLS residuals and weighted residuals on a scatterplot. The weighted residuals are the average of the OLS residuals of each tracts queen neighbors. If each observation were truly independent and not spatially correlated then there would not be a relationship in this scatterplot;

however, there does appear to be a relationship between these variables. That relationship is characterized by Slope b, which is also called rho, ρ, or lambda, λ. Here Slope b is 0.733 meaning that as the weighted residual changes by 1 unit, the residual changes by 0.733 units. Since the p-value of Slope b is 0, it is significant and therefore indicative of significant spatial autocorrelation.

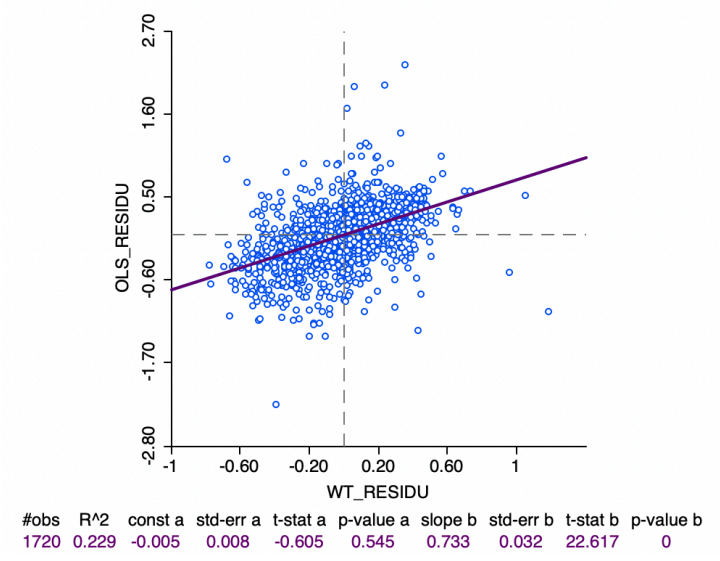

Figure 5 – Scatterplot of OLS Residuals and Weighted Residuals

Figures 6 and 7 below display the Moran's I statistics of the OLS regression residuals. The scatterplot of Figure 6 shows the Moran's I is 0.313, indicating that there is low positive spatial autocorrelation. Figure 7 shows the histogram of Moran's I values for with 999 permutations and indicates that the Moran's I of 0.313 is in fact significant since the p-value is below 0.05. This is problematic for OLS regression, once again violating the assumption that observations are independent and not spatially autocorrelated.

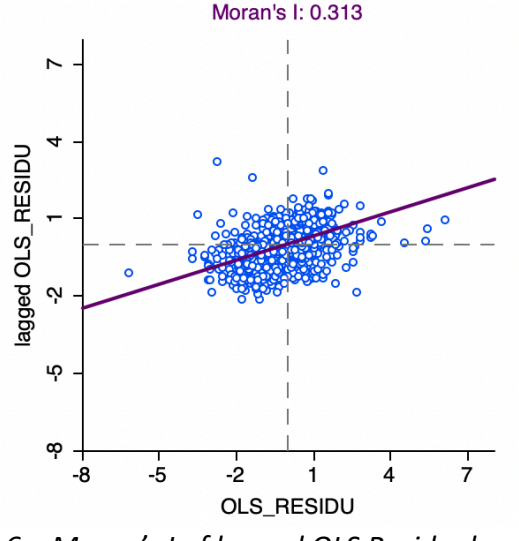

Figure 6 – Moran's I of lagged OLS Residuals

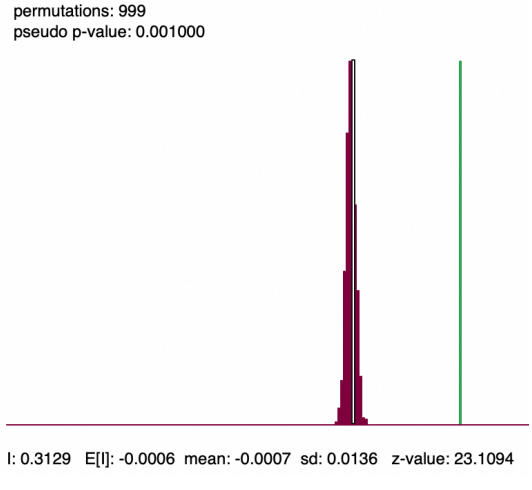

Figure 7 -Histogram of Lagged OLS Residual Moran's I

c) Spatial Lag and Spatial Error Regression Results

As is shown in Table 2, W_LNMEDHVAL is the spatial lag of the dependent variable LNMEDHVAL. And it has a probability of less than 0.00001, so it means that it is of great significance. The median house value in an area is associated with median house value in surrounding areas. The coefficient of it is around +0.651, which means that there is a positive relationship between median house value and the value in surrounding areas.

Table 2 – Spatial Lag Regression Output

SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION

Data set:	Regression Data		
Spatial Weight:	queen-hw2-1		
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):	0.651107		
R-squared:	0.818603	Log likelihood:	-255.562
Sq. Correlation:	-	Akaike info criterion:	523.123
Sigma-square:	0.0719325	Schwarz criterion:	555.824
S.E of regression:	0.268202		

Variable	Coefficient	Std.Error	z-value	Probability
W_LNMEDHVAL	0.651107	0.0180482	36.076	0.00000
CONSTANT	3.89835	0.20109	19.3861	0.00000
LNNBELPOV	-0.0340632	0.00629222	-5.41355	0.00000
PCTBACHMOR	0.00851569	0.00052192	16.3161	0.00000
PCTSINGLES	0.00202905	0.00051571	3.93448	0.00008
PCTVACANT	-0.00852676	0.00074357	-11.4673	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

DIAGNOSTICS FOR SPATIAL DEPENDENCE

SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX: queen-hw2-1

TEST	DF	VALUE	PROB
Likelihood Ratio Test	1	911.8633	0.00000

Other variables LNNBELPOV, PCTBACHMOR, PCTSINGLES and PCTVACANT are all very significant ($p < 0.01$). This is consistent with OLS results, which shows that the lagged dependent variable W_LNMEDHVAL does not affect the significance of other variables.

According to the Breusch-Pagan test, our spatial lag model has a small Breusch-Pagan test probability of 0.000, which means that there is still heteroscedastic issue of the residuals. Introduction of new variables may be useful to address this problem.

Table 3 – Criteria for comparing the models

Model	Akaike Info Criterion	Schwarz Criterion	Log Likelihood	Likelihood Ratio Test Prob.
OLS	1432.99	1460.24	-711.493	/
Spatial Lag	523.123	555.824	-255.562	<0.0001

According to Table 3, the AIC and SC of spatial lag regression is smaller (the smaller the better), the log likelihood of spatial lag regression is bigger (the bigger the better), and the likelihood ratio test probability is less than 0.0001 (spatial lag model is better). Based on these criteria above, spatial lag regression model is better than the OLS model.

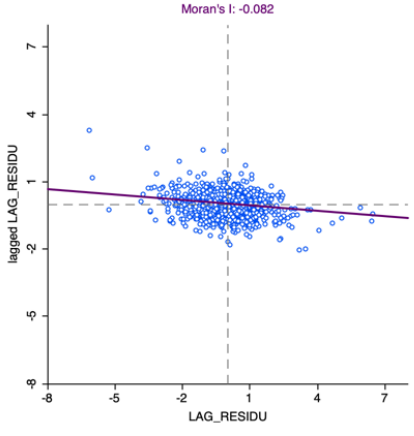

Figure 8 – Moran's I of lagged SL Residuals

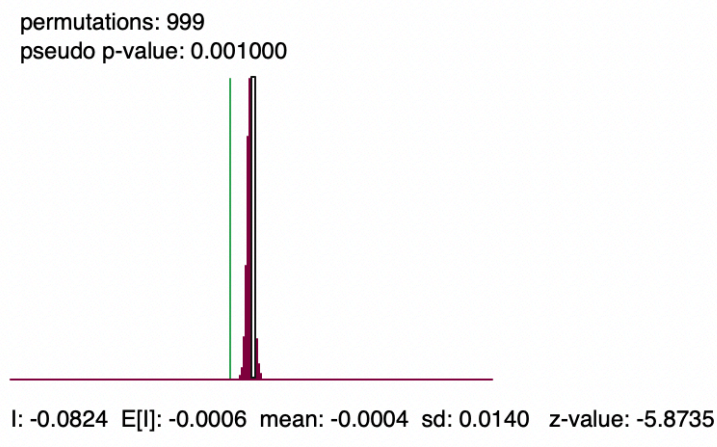

Figure 9 -Histogram of Lagged SL Residual Moran's I

Figures 8 and 9 above display the Moran's I statistics of the spatial lag regression residuals. Figure 8 shows the Moran's I is -0.082, indicating that the Moran's I is negative and very close to 0. Figure 9 shows the histogram of Moran's I values for with 999 permutations and indicates that the Moran's I of -0.082 is significant. This shows that there is a slightly negative spatial autocorrelation in the residuals of the model. However, compared to the OLS regression (Moran's I = 0.313), the residuals have less spatial autocorrelation.

With all these criteria considered, the spatial lag model is better than the OLS model.

The results of the spatial error regression are listed in Table 4.

Table 4 – Spatial Error Regression Output

SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION

```
Data set:            Regression Data
Spatial Weight:      queen-hw2-1
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):0.814872

R-squared:           0.806997          R-squared (BUSE):          -
Sq. Correlation:     -                 Log likelihood:            -372.492533
Sigma-square:        0.0765348         Akaike info criterion:     754.985
S.E of regression:   0.276649          Schwarz criterion:         782.235
```

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	10.9062	0.0534556	204.023	0.00000
LNNBELPOV	-0.0345369	0.00708851	-4.87224	0.00000
PCTBACHMOR	0.00982427	0.000728944	13.4774	0.00000
PCTSINGLES	0.00266586	0.000620803	4.29421	0.00002
PCTVACANT	-0.00577991	0.000886626	-6.519	0.00000
LAMBDA	0.814872	0.0163744	49.765	0.00000

REGRESSION DIAGNOSTICS

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	4	211.1640	0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE

SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX: queen-hw2-1

TEST	DF	VALUE	PROB
Likelihood Ratio Test	1	678.0016	0.00000

As is shown in Table 4, LAMBDA is the coefficient of the lagged residual. And it has a probability of less than 0.00001, so it means that it is of great significance. The coefficient of it is around +0.814, which means that there is a strong positive relationship between median house value and the spatially lagged residuals in surrounding areas.

Other variables LNNBELPOV, PCTBACHMOR, PCTSINGLES and PCTVACANT are all very significant ($p < 0.01$). This is consistent with OLS and spatial lag regression results, which shows that the lagged residual does not affect the significance of other variables.

According to the Breusch-Pagan test, our spatial error model has a small Breusch-Pagan test probability of 0.000, which means that there is still heteroscedastic issue of the residuals.

Table 5 – Criteria for comparing the models

Model	Akaike Info Criterion	Schwarz Criterion	Log Likelihood	Likelihood Ratio Test Prob.
OLS	1432.99	1460.24	-711.493	/
Spatial Error	754.985	782.235	-372.492	<0.0001

According to Table 5, the AIC and SC of spatial lag regression is smaller (the smaller the better), the log likelihood of spatial lag regression is bigger (the bigger the better), and the likelihood ratio test probability is less than 0.0001 (spatial error model is better). Based on these criteria above, spatial

error regression model is also better than the OLS model.

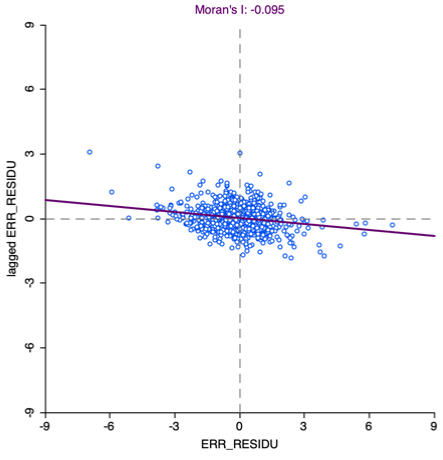

Figure 10 – Moran's I of lagged SE Residuals

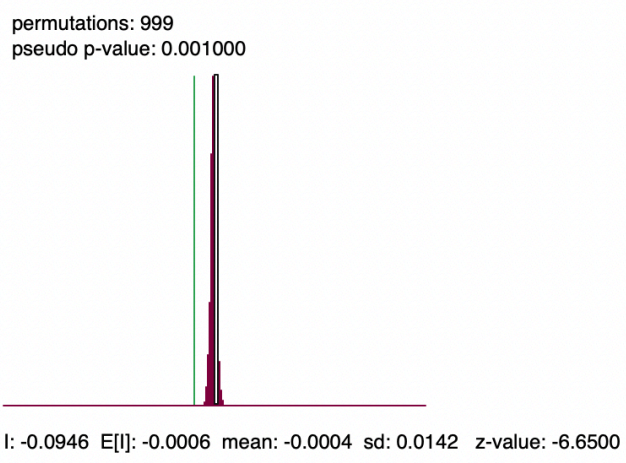

Figure 11 -Histogram of Lagged SE Residual Moran's I

Figures 10 and 11 above display the Moran's I statistics of the spatial lag regression residuals. Figure 10 shows the Moran's I is -0.095, indicating that the Moran's I is negative and very close to 0. Figure 11 shows the histogram of Moran's I values for with 999 permutations and indicates that the Moran's I of -0.095 is significant. This shows that there is also a slightly negative spatial autocorrelation in the residuals of the model. However, compared to the OLS regression (Moran's I = 0.313), the residuals have less spatial autocorrelation.

With all these criteria above considered, the spatial error model is also better than the OLS model.

Table 6 – Criteria for comparing the models

Model	Akaike Info Criterion	Schwarz Criterion
Spatial Lag	523.123	555.824
Spatial Error	754.985	782.235

Comparing the results of spatial lag and spatial error model, as in shown in the table below, we can see that the AIC and SC are smaller for the spatial lag model, indicating that the spatial lag model is a better fit than spatial error. Other criteria like likelihood ratio test should not been used in this circumstance, because the models are not nested.

d) Geographically Weighted Regression Results

Table 6 shows the results of the GWR regression. The overall R^2 is 0.81486 indicating that 81% of the variance of the model was explained by the four predictors. This is higher than 0.6623, the R^2 of the OLS regression, so the GWR model does a better job of explaining the variance in the dependent variable. The Akaike Information Criteria of GWR is 668.9166, lower than 1432.99 of OLS and 754.985 of Spatial Error, but higher than 754.985 of Spatial Lag, indicating that GWR fits better than OLS and Spatial Error but worse than Spatial Lag.

Table 6 – Supplementary Table of the GWR regression

OID_	VARNAME	VARIABLE	DEFINITION
0	Neighbors	166	
1	Residual Squares	126.2759715	
2	Effective Number	171.0479745	
3	Sigma	0.285523183	
4	AIC	668.9166503	
5	R^2	0.814861033	
6	R^2 Adjusted	0.794535997	
7	Dependent Field	0	LNMEDHVAL
8	Explanatory Field	1	LNNBELPOV
9	Explanatory Field	2	PCTBACHMOR
10	Explanatory Field	3	PCTSINGLES
11	Explanatory Field	4	PCTVACANT

Figure 12 displays the Moran's I result of GWR residuals. It shows the global Moran's I is 0.082, indicating that there is almost no spatial autocorrelation in GWR residuals. Furthermore, the points are randomly scattered around the origin, indicating that there is also no spatial autocorrelation in GWR residuals locally either.

Compared to the global Moran's I of 0.313 of OLS residuals, there is less spatial autocorrelation in GWR residuals. The global Moran's I are -0.082 of Spatial Lag residuals and -0.095 of Spatial Error Residuals, which are seemingly the same as GWR. However, both of their scatterplots show there are stronger linear relationships between the residuals and lagged residuals, indicating more local autocorrelation in Spatial Lag and Spatial Error residuals than GWR residuals.

Figure 13 shows the histogram of Moran's I values with 999 permutations and indicates that the Moran's I of 0.082 is in fact significant since the pseudo p-value is below 0.05.

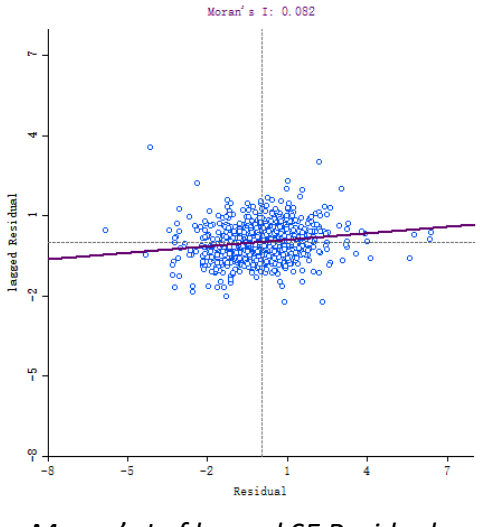

Figure 12 – Moran's I of lagged SE Residuals

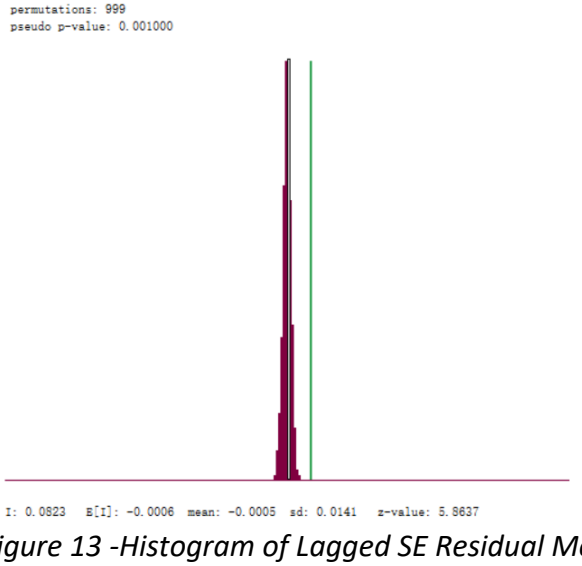

Figure 13 -Histogram of Lagged SE Residual Moran's I

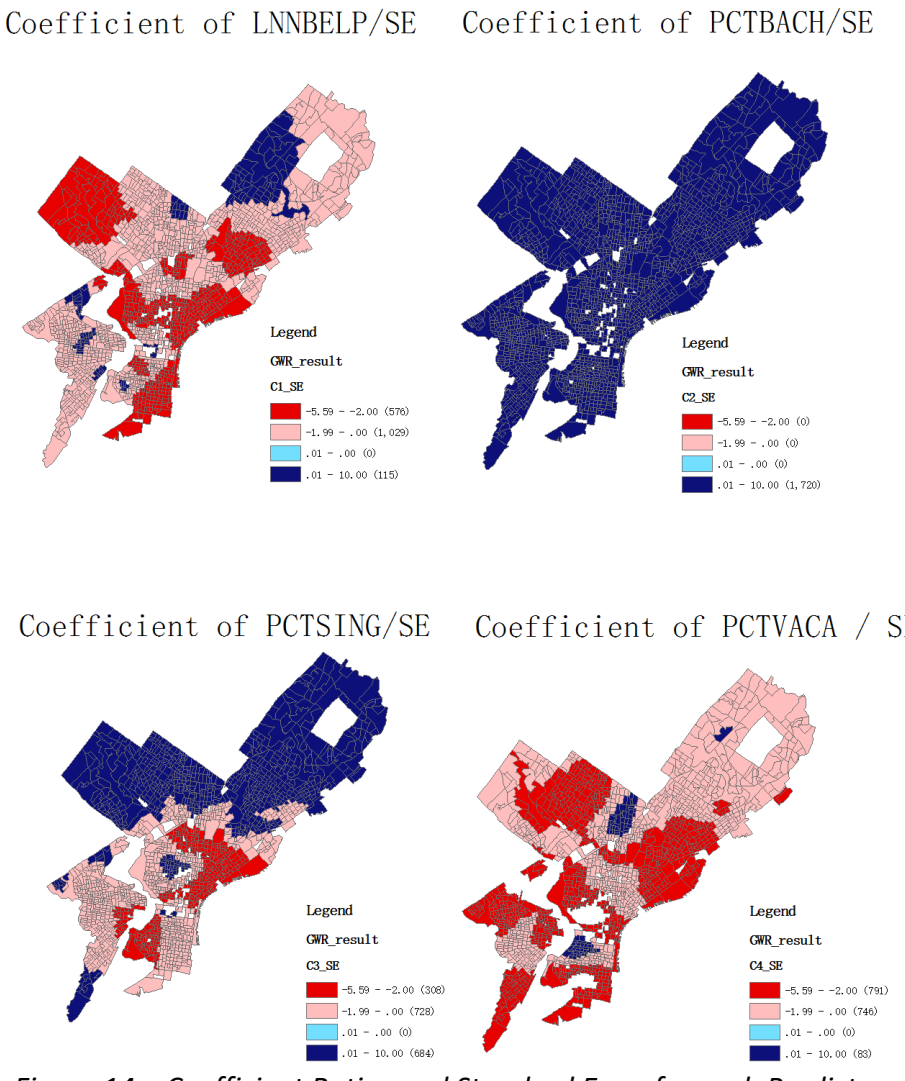

Figure 14 – Coefficient Ratios and Standard Error for each Predictor

Figure 14 displays the local regression results. These maps show the ratio of each beta coefficient and its standard error. The break and color scheme indicate the following:

- Dark red: Ratio $\leq$ -2: A negative relationship with the dependent variable that's possibly significant.
- Pink: -2 < Ratio $\leq$ 0: A negative relationship with the dependent variable that's likely not significant.
- Light blue: -0 < Ratio < 2: A positive relationship with the dependent variable that's likely not significant.
- Dark blue: Ratio $\geq$ 2: A positive relationship with the dependent variable that's possibly significant.

The map on the upper left shows that between LNMEDHVAL and LNNBELP, there is possibly a significant and positive relationship in some areas in the north while a significant and negative relationship in some places in northwest, middle north and southeast.

On the other hand, the map on the upper right shows there is possibly a significant and positive relationship between LNMEDHVAL and PCTBACHMOR at every location.

Next, map on the lower left shows that between LNMEDHVAL and PCTSING, there is possibly a significant and positive relationship in roughly 1/3 of the areas in the north and a significant and negative relationship in some places in the middle of the city.

Finally, map on the lower right shows that between LNMEDHVAL and PCTVACA, there is significant and negative relationship in only a few clustered places, while there is a significant and negative relationship in relatively more places.

The figure below shows the map of local R^2 values. There is strong fit in northwest and west part, and lousy fit in other places. Besides, the map shows that similar R^2 values are clustered, which seems a highly positive spatial autocorrelation in R^2.

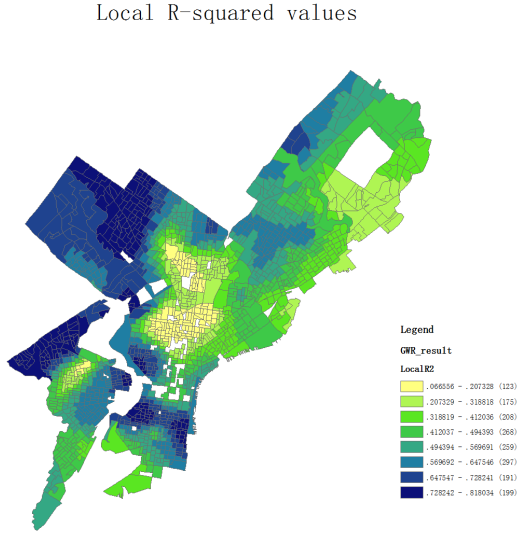

Figure 15 – Map of Local R-squared values

Discussion

In this report, we built upon the OLS regression from Assignment 1 using three spatial regression methods (Spatial Lag, Spatial Error, and GWR) to predict median house values with poverty levels, education, single family homes, and vacancy. The table below shows the comparation of these four models.

Table 8 – The overall comparison of four regression models

	OLS	**Spatial Lag**	**Spatial Error**	**GWR**
R^2	0.6623	0.818603	0.806997	0.81486 (global)
AIC	1432.99	523.123	754.985	668.91
Moran's I of lagged residuals	0.313	-0.082	-0.095	0.082

With a relatively higher R^2, lower AIC and Moran's I of lagged residuals, Spatial Lag model is the best. However, there are still some limitations of Spatial Lag since assumptions are not all met.

The assumption of normality of residuals is violated. According to the Jarque-Bera test, the p-value of it is less than 0.0001, which implies that there are systematic errors in Spatial Lag model. And the Breusch-Pagan test also suggests that there is heteroskedasticity problem in the model. The assumption of spatial stationarity is likely violated. In Figure 8, the scatterplot of Moran's I of lagged SE Residuals, we observed some outlier points in high-low and low-high area far away from cluster in origin point. This means in some area there is strong negative spatial autocorrelation in residuals.