Prediction of Median House Values in Philadelphia Block Groups

1 Introduction

This project will investigate the relationship between the median housing prices and several neighborhood features in the study area of Philadelphia. The data will be in a census block group unit. There are 4 neighborhood features with a potential influence to the house prices in Philadelphia as the main predictors of this analysis, including the percentage of residents with a bachelor's degree or higher degrees, the percentage of vacant buildings, the percentage of detached single-family houses, and the number of families with incomes below the federal poverty line. And the dependent variable in this project is median house values.

In our previous study, the OLS regression is used to derive an analysis to investigate the relationship between the median house values and the predictors that are mentioned above. There is a problem that the OLS analysis is inappropriate in the cases while analyzing the datasets with a spatial component inside, as OLS just estimates the correlation in a linear regression model without considering the spatial proximities between variables, which can also be explained as the spatial autocorrelation. To address this problem, in this project, we aim to use GeoDa and ArcGIS to run spatial lags, spatial errors and geographically weighted regressions (GWR) to examine whether the methods could perform better than OLS when dealing with datasets containing spatial components. In addition, we also run these regression models in R.

2 Methods

2.1 Concept of Spatial Autocorrelation

The First Law of Geography, according to Waldo Tobler, is "everything is related to everything else, but near things are more related than distant things."[1] It's the fundamental assumptions of all spatial analysis [2] and indicates the positive spatial autocorrelation. Spatial autocorrelation is the presence of systematic spatial variation in a mapped variable (R.P. Haining, 2001). When observations are closer to each other in space and have similar data values, it shows positive spatial autocorrelation. When observations tend to have very contrasting values, it shows negative spatial autocorrelation. But it happens fairly rarely, so we mainly focus on positive spatial autocorrelation.

Spatial autocorrelation may be indexed, quantified by including an autoregressive parameter in a regression model, or filtered from variables (Daniel A. Griffith, 2005). It can be quantified with indices. Moran's I is a method of testing spatial autocorrelation or spatial dependencies. The equation is as follows:

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

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

Where $\bar{X}$ is the mean of the variable X, X_i is the variable value at a particular location i, X_j is the variable value at a particular location j, w_{ij} is a weight indexing location of i relative to j, and n is the number of observations. When $I > 0$, large positive values (close to 1) indicate that there is strong positive autocorrelation; When $I < 0$, large negative values (close to -1) indicate that there is strong negative autocorrelation. Values close to 0 reflects that there is no spatial autocorrelation. Under the

independence hypothesis, Moran's I has a mean that is slightly negative, and the expected value of the Moran's I is equal to -1/ (n - 1), n is the number of observations. Unlike the Pearson correlation coefficient, Moran's I is not always between -1 and 1.

When dealing with spatial analysis, we need to define spatial proximity through a weight matrix firstly. The weight matrix is a table which summarizes all the pairwise spatial relationships in the dataset. There are two kinds of spatial weights: Distance-based Measures and Contiguity-based Measures. Here, we use the Queen matrix, a contiguity-based measure, which defines neighbors as spatial units sharing a common edge or a common vertex. Statisticians sometimes are more likely to try different spatial weight matrixes to make sure that the result is not dependent on the matrix.

We use Moran's I to test whether the spatial dependence exists and whether spatial autocorrelation is significant. The hypotheses are as follows:

H_0: There is no spatial autocorrelation between the median house value and the four predictors.

Ha_1: There is a positive spatial autocorrelation between the median house value and the four predictors.

Ha_2: There is a negative spatial autocorrelation between the median house value and the four predictors.

One favored approach for testing whether spatial autocorrelation is significant from H_0 is the random permutation test, which consists of randomly reassigning the attribute values to a cell in the studied extent following some hypothesized process (such as complete spatial randomness, CSR) and computing the Moran's I value each time thus creating an empirical distribution of I under H_0, then rank all Moran's I values in descending order. The Pseudo P-value is obtained by taking the rank of Moran's I and dividing it by the total number of permutations. If Pseudo P-value is less than the alpha value, which generally is set to 0.05, then we can say that is a significant spatial

autocorrelation and reject the null hypothesis.

To get a better understanding of details of spatial autocorrelation, LISA (Local Indices of Spatial Autocorrelation) analysis should be included. It can demonstrate to what extent the neighboring values of location i of variable X is related to the value at location j. The analysis categorizes the blocks into four types:

- High-High: the deviation of the variable value and the average deviation of neighbor locations from the global mean are both positive, it's positive spatial autocorrelation.
- High-Low: the deviation of the variable value from the global mean is positive, but the average deviation of neighbor locations from the global mean is negative, it's negative spatial autocorrelation.
- Low-Low: the deviation of the variable value and the average deviation of neighbor locations from the global mean are both negative, it's positive spatial autocorrelation.
- Low-High: the deviation of the variable value from the global mean is negative, but the average deviation of neighbor locations from the global mean is positive, it's negative spatial autocorrelation.

2.2 OLS Regression and Assumptions

A multiple Ordinary Least Squares (OLS) regression had been used to find out the linear relationship between median house value with the proportion of residents with at least a bachelor's degree, the proportion of vacant houses in each block group, the proportion of detached houses, and the number of residents living in poverty. There are five assumptions for OLS regression models: **Linear relationship** - There exists a linear relationship between each dependent and explanatory variable; **Independence -** The observations are independently and identically distributed; **Multivariate Normality** - The error term ε is normally distributed conditional on the explanatory

variables, which is important for point estimation; **Homoscedasticity -** The error term ε is homoscedastic, which means the residuals have constant variance at every point in the linear model; **No Multicollinearity -** OLS regression has no multi-collinearity, which means that there should be no strong linear relationship (r>|0.8|) between the independent variables. For detailed explanations, you can refer to the former reporter.

One of the most important assumptions is that observations are independent of each other and the errors are random or independent, but if the data has a spatial component, values of a variable in nearby areas may be related to each other and are not independent. The assumptions don't hold. In this case, we can use Moran's I to test whether the randomness or independence of errors hold. In addition, we can also regress the OLS residuals on nearby residuals, which are residuals at neighboring block groups as defined by the Queen matrix to conduct the test.

A spatial lag is a variable that essentially averages the neighboring values of a location. If the value of the dependent variable at one location is associated with the values of that variable in nearby locations, we can use the spatial lag model to address the spatial autocorrelation in the dependent variable. The spatial lag term $rho(\rho)$ reflects how much a spatial feature is influenced by its neighbors. The ρ is the slope of the fitted line of regression of the OLS residuals with their neighboring residuals. It demonstrates the relationship between the residuals and their neighbors. If ρ is significant, OLS is biased and inconsistent (so don't use it!)

GeoDa is a free and open source software tool that is often used for spatial data analysis. It provides tools to check other assumptions. One assumption is that the data should have heteroscedasticity, which means the variance of the dependent variable does not change as the variance of the predictor changes. GeoDa has three different diagnostics for heteroscedasticity: the **Breusch-Pagan Test**, the **Koenker-Bassett Test**, and the **White Test**. The null hypothesis here is that of there is no heteroscedasticity, while the alternate hypothesis is having heteroscedasticity. In general, if the p-value is less than

0.05, then we can reject the null hypothesis for the alternate hypothesis of heteroscedasticity. Another assumption that GeoDa can test is the Normality of the Residuals. The **Jarque-Bera Test** can be used to test this assumption. The null hypothesis is there is normal distribution in the residuals, while the alternative hypothesis is there is non-normality in residuals. If the p-value is less than 0.05, we can reject the null hypothesis of normality for the alternative hypothesis of non-normality.

2.3 Spatial Lag and Spatial Error Regression

We will be using GeoDa to perform spatial lag and spatial error regressions on the Philadelphia block group dataset, and exploring whether we are able to fix spatial autocorrelation by using these two regressions.

Spatial lag regression is one of the methods that are available to account for correlation between neighboring observations. More specifically, we will see similar observations, or each observation related to each other when spatial autocorrelation exists, consequently, we no longer have independent observations and randomly distributed residuals. Despite OLS regression is not suitable in this case, the issue can be solved by adding an additional lagged variable in the existing OLS regression. The extra lagged variable in spatial lag regression is dedicated to the dependent variable, which is LNMEDHVAL, and the components of this variable are its coefficient (β) and y-lag variable (Wy). The y-lag variable in the spatial lag regression can account for spatial autocorrelation because we will define a weight matrix such as rook, queen, or nearest neighbor based on our needs for each observation, and the values of lagged variable will be calculated from the average values of its neighbors. In this project, we use the queen matrix. As a result, the issues with spatial autocorrelation should be fixed by adding this term. The spatial lag model will have the following terms in the equation:

$$LNMEDHVAL = \rho * W_{LNMEDHVAL} + \beta_0 + \beta_1 * LNNBELPOV + \beta_2 * PCTVACANT + \beta_3 * PCBACHMORE + \beta_4 * PCTSINGLES + \varepsilon \quad (2)$$

In the equation, ρ represents the coefficient of the y-lag variable ($W_{LNMEDHVAL}$), and its value ranges from -1 to 1. The magnitude of the lag parameter can indicate how strong spatial autocorrelation is in the existing OLS regression, since a large value of lag parameter (ρ) is a sign showing that the existing variables in the OLS regression performed badly in accounting for spatial autocorrelation. $W_{LNMEDHVAL}$ is the y-lag variable which is the average value of the neighboring observations. β_0 is the dependent variable intercept, and β_0 will be the predicted value of LNMEDHVAL when all the independent variables are set equal zero. Next, the β coefficients show how will LNMEDHVAL fluctuate when each independent variable increases by one unit accordingly. Finally, ε represents the residuals between the actual LNMEDHVAL and values predicted by our model.

Another regression can be used to account for spatial autocorrelation is the spatial error model, which adds an additional spatially lagged residuals term to the OLS regression to help account spatial autocorrelation. An OLS regression will be executed first, then the residuals will be transformed before included in the spatial error regression. The spatial error regression accounts clustering in residuals by first creating a weights matrix for the residuals using either rook, queen or specified distance neighboring methods, thus spatial dependency in the residuals will be accounted by the lagged residuals ($\lambda W\varepsilon$). More specifically, the original residuals ε from the OLS regression are decomposed into lagged residuals and the random noise (μ). The equation for spatial error model is:

1) $LNMEDHVAL = \beta_0 + \beta_1 * LNNBELPOV + \beta_2 * PCTVACANT + \beta_3 * PCBACHMORE + \beta_4 * PCTSINGLES + \varepsilon$ (3)
2) $\varepsilon = \lambda W\varepsilon + \mu$ (4)
3) *spatial error regression equation*:
$LNMEDHVAL = \beta_0 + \beta_1 * LNNBELPOV + \beta_2 * PCTVACANT + \beta_3 * PCBACHMORE + \beta_4 * PCTSINGLES + \lambda W\varepsilon + \mu$ (5)

Similar to the spatial lag regression, the β coefficients show how will LNMEDHVAL

change when each independent variable increases by one unit accordingly. λ represents the spatial autoregressive coefficient, which can be varied between -1 and 1. $W\varepsilon$ is the lagged residuals. Finally, μ is the random noise.

Since both spatial lag regression and spatial error regression are developed on OLS regression, some assumptions of OLS regression are still required, like Linearly Relationship, Multivariate Normality, Homoscedasticity and No Multicollinearity. The details have been described above. But the assumption that spatial independence of observations is not needed.

The reason we will perform spatial lag regression and spatial error regression on the Philadelphia block group dataset is that we suspect that spatial dependencies may exists in the residuals or in the observations. Therefore, we are trying to test whether these regressions can help remove the spatial autocorrelation in the residuals of the model. We would like to see that the resulting residuals are no longer spatially correlated and have less heteroskedasticity.

Once we have the output from the spatial lag regression and spatial error regression, we will compare each regression model with that of the original OLS regression to conclude whether spatial regressions perform better than the OLS regression. The criteria used to determine the performance of the spatial regressions are Akaike Information Criterion (AIC), Log Likelihood, and results from Likelihood Ratio Test. AIC is used to analyze the goodness of fit of a model, from which we will be able to see the precision and complexity of the model and the tradeoff between precision and complexity by comparing the AIC values for two models. The lower the AIC value a model gets, the more desirable the model is. Thus, we expect to see the spatial regressions to have lower AIC values than the OLS regression.

Log likelihood is also another valuable piece of information we need to determine whether spatial regressions are necessary. It is related to the maximum likelihood

method of fitting a model to the data. A higher value means that the parameter in the model makes the result more likely to occur than a model with lower log likelihood. Note that this criterion can only be used for comparing nested models, which means we can only compare the Log Likelihood of OLS with spatial lag or between OLS and spatial error. After comparison, we conclude which model is better by choosing the model with higher log likelihood, thus we expect both spatial regression models to have higher log likelihood than OLS regression.

The third criteria is the Likelihood Ratio Test, and we will use this test to compare the OLS model with each of the spatial model. The null hypothesis of likelihood ratio test in this case is that the spatial lag (error) model is not a better specification than the OLS model, while the alternative states the opposite than the null. When p-value is less than 0.05, we can reject the null hypothesis of the spatial model is not better.

Additionally, Moran's I of the regression residuals is also a good choice to decide whether the spatial models are better than OLS regression. Since we suspect that there is spatial autocorrelation in the data, we expect the OLS regression to have a large positive Moran's I value, which signifies clustering in the residuals. Since the spatial regressions can account for spatial autocorrelation, we can expect these two models to return very low Moran's I value because a low Moran's I value represents that there is not clustering in the residuals.

2.4 Geographically Weighted Regression

In order to derive the Geographically Weighted Regression (GWR) analysis for this project, the ArcGIS software is used. The Simpson's paradox, which is a statistical phenomenon, states that, while the population is divided into sub-populations, the link between the two variables of a population occurrence reverses or vanishes. For example, for the analysis of the relationship between the burglaries and the median house values in Philadelphia, from the perspective of all block groups, the burglaries and median

house value are negatively correlated, however, if the data is sorted according to different regions of the city such as west, north, south Philadelphia, the relationship between the median house values and burglaries is different within various regions of the city. From a mathematics perspective, the GWR regression model is applied according to the following equation based on each observation.

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i \quad (6)$$

In which, the subscript i indicates the relationship between y which is the dependent variable, and x_k which is the predictors and k is a value in a range from 1 to m, m is the total number of census blocks. And the GWR regression model represents the relationship between the predictors and dependent variable, which is specific to that location. Additionally, the local regression of GWR is a regression for each census block group.

To run the local regression of GWR for every census tract, multiple locations or observations are required. In the dataset, other observations were used for the running of GWR regression. Also, the observations close to the location i are derived with greater weights, the weights of observations are different in different locations due to the effects of parameters in various locations. Therefore, the estimations of parameters of location i are affected relatively more significantly by the observations close to the location i.

Furthermore, the nearby locations of i could be weighed by various bandwidths such as the fixed and adaptive bandwidths. The fixed bandwidth (h) suggests that in each census tract, after the regression of observations in the fixed bandwidth distance, the amount of observations are different around each location, however, both the area and the fixed bandwidth distance would be constant. Also, the adaptive bandwidth (h) suggests the amount of observations would stay fixed, however, the area of bandwidths would be different. Moreover, the weighing function would assign the observations that

are closer to the bandwidth distance with a higher weight, and zero weights are assigned to the observations that outside the bandwidth distance. As the weighing function, the fixed and adaptive bandwidth are appropriate for different types of observations respectively. For situations where the distribution of observations is rather steady in terms of factors such as the number of neighbors and size across the space, the fixed bandwidth kernel would be more suitable. For situations where distribution of observations changes throughout space with events concentrated or polygons unevenly shaped or sized, the adaptive bandwidth kernel is more appropriate. In Philadelphia, due to the irregular distribution of census block groups and the fact that they are relatively smaller and more concentrated in the center areas of the city than they are on the city's periphery, adaptive bandwidth will be used in this project.

Before running GWR regression, we run OLS regression firstly in this model to confirm that the model is reasonable with relationships that are worth to be investigated. With these relationships in the model, the GWR could be processed for the analysis of spatial non-stationarity and spatial autocorrelation. In addition, various OLS regression assumptions still hold in GWR regression, such as homoscedasticity, non-multicollinearity, residual normality. The multicollinearity problem in the model is most likely to occur when the value for an independent variable cluster spatially in a substantial way. In addition, if GWR has two or more variables that have similar patterns of clusters at all locations in a certain part of Philadelphia, it also exhibits the multicollinearity problem. In this case, the Condition Number from *'Cond. Number'* in the attribute table could be used to find out when the results are unstable. It indicates the instability of the results due to local multicollinearity.

Because GWR is comparatively more sophisticated in the current study, p-values are not included in the GWR results in this project. A regression is applied to every observation in GWR, deriving a great mount of regressions with parameters that require to be estimate. Also, the type I and type II errors might exist in the OLS regression. A type I error, also known as false positive, would occur when a false null hypothesis is

invalidated. The alpha level of the hypothesis test represents the possibility of a type I error. A type II error, also known as ‘false negative’, arises when the model fails to invalidate a misleading null hypothesis, and the probability of the occurrence of it could be decreased by increasing the power. They are also existing in GWR, the possibility of the occurrence of the errors tends to increase greatly as the running of a great number of regressions in GWR. Therefore, in this case, the p-values are inaccurate for rejection of the null hypothesis.

3 Results

3.1 Spatial Autocorrelation Results

We have calculated the global Moran’s I values for the dependent variable **LNMEDHVAL** and plotted a scatter plot to show the spatial lag among the levels of the dependent variable in Figure 1. The global Moran’s I value is 0.794, which means that **LNMEDHVAL** has a very high positive spatial autocorrelation. The similar median house values tend to cluster together. Then, we conducted the random permutations test. The k is 999 times. The Moran’s I value for the permutations is 0.7936, which is lightly lower than the original value. The p-value is 0.001, which means there is (at most) a one-in-a-thousand chance of observing a Moran’s I of 0.794 if in fact there is no spatial autocorrelation present. The dependent variable **LNMEDHVAL** is statistically significantly spatial autocorrelated, so we can reject the null hypothesis of no spatial autocorrelation.

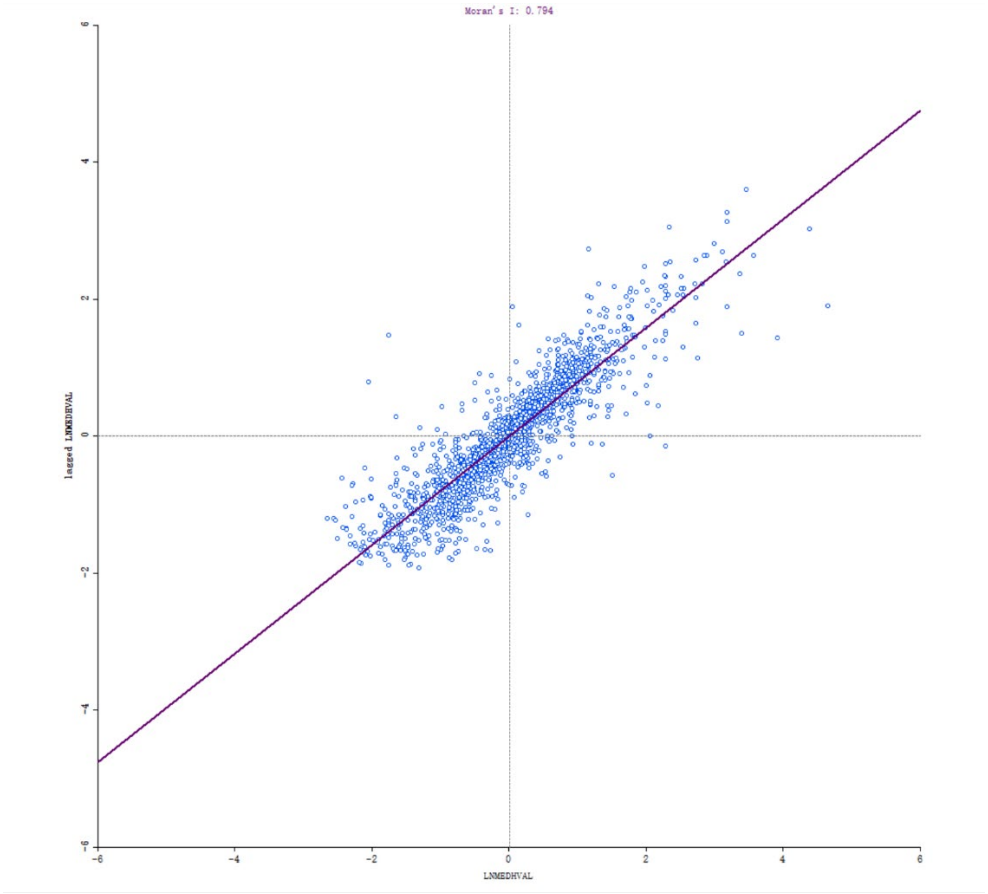

Figure 1. Global Moran's I Value of Median House Value

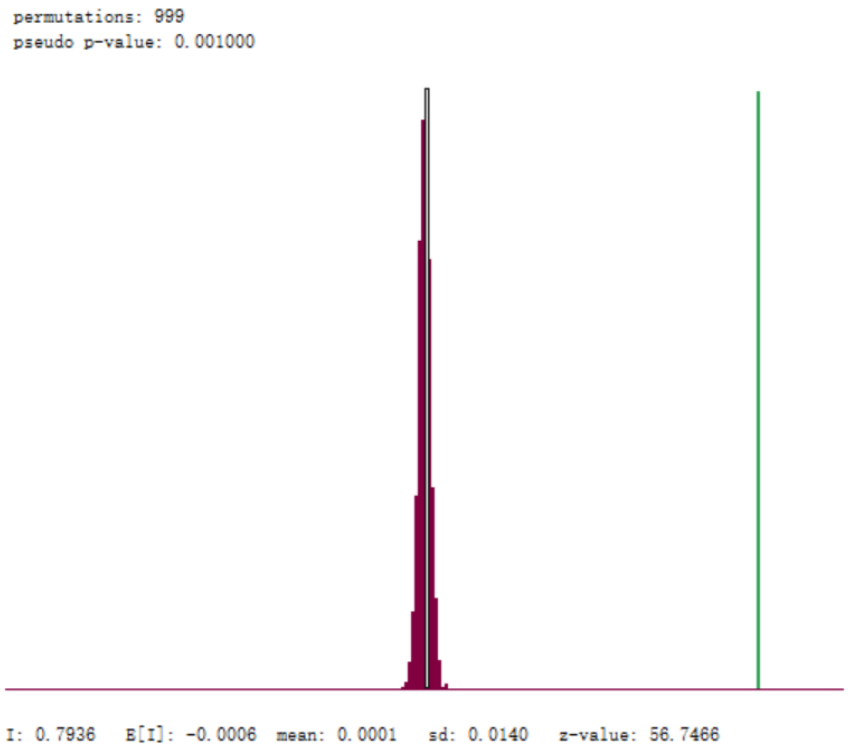

Figure 2. Histogram of Moran's I Value for Random 999 Permutations Test

For Local Moran's I results, we plotted the Significance Map and LISA Cluster Map obtained by running the Local Moran's I in Figure 3 & 4. We can find that most areas of Philadelphia are not significant, the significant areas are mainly concentrated in the central, northeast and northwest of Philadelphia, and half of the south. It's clear that the areas of Kensington, Port Richmond, Oxford Circle and Lawncrest are not significant. The low-low cluster areas are mainly located in the central, southwest and parts of west Philadelphia. There are only a few low-high cluster areas, which are distributed in the south, northeast and west Philadelphia. The high-high areas accounts for a large part and are mainly clustered in the northeast, northwest, and south Philadelphia. The high-low areas, amount is as same as the low-high areas', locate in the west and south.

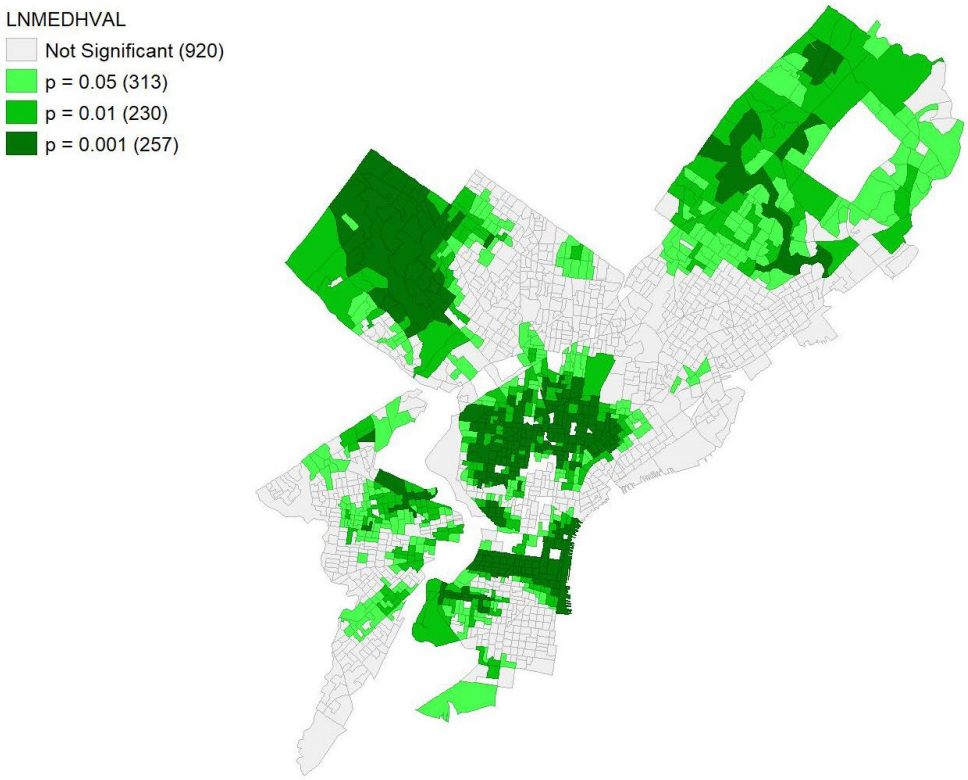

Figure 3. Significance Map of Local Moran's I

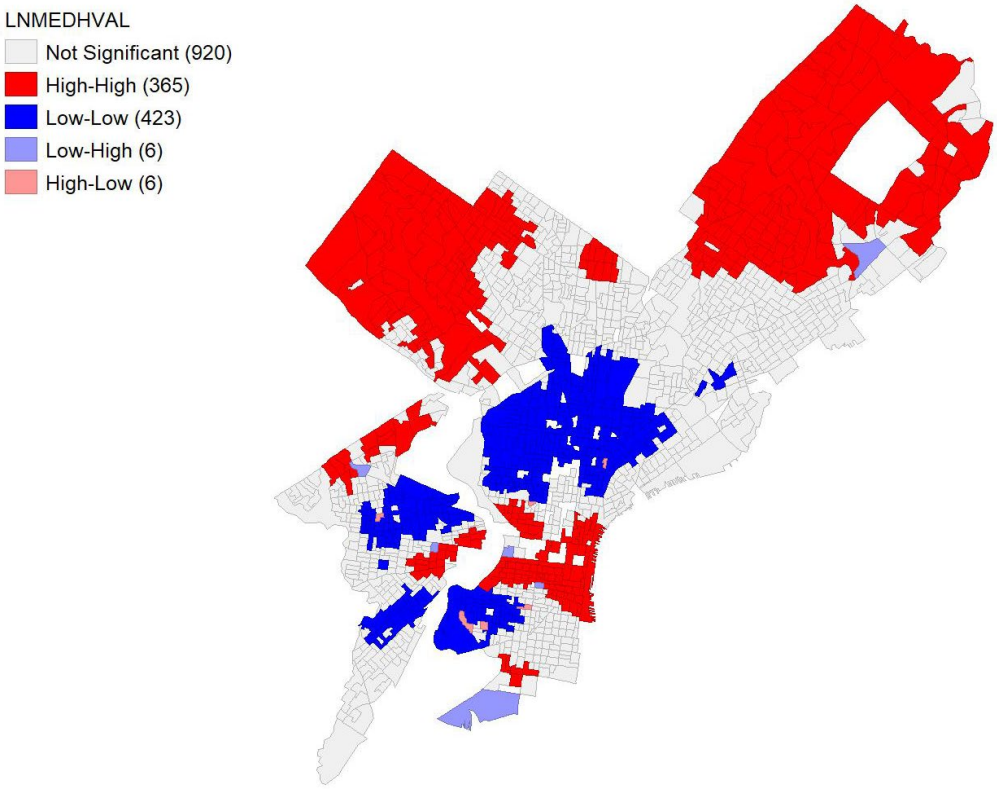

Figure 4. LISA Cluster Map

3.2 OLS Regression Results

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
PCTVACANT	-0.0191563	0.000977851	-19.5902	0.00000
PCTSINGLES	0.00297695	0.000703155	4.23371	0.00002

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 : Regression Data
(row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.3129	22.3664	0.00000

Table 1. OLS Regression Results

We regressed the y, which is natural log of Median House Value (LNMEDHVAL) on the natural log of number of households with income below 100% poverty level (LNBELPOV100), percentage of residents in Block Group with at least a bachelor's degree (PCBACHMOR), percentage of housing units that are detached single family (PCTSINGLES), and the percentage of vacant housing units (PCTVACANT). The output of our regression is shown in Table 1. The R-Squared value is 0.66, which means that about 66% of the variance of the dependent variable LNMEDHVAL can be explained by the predictors. All the four predictors are significant since they have a p-value less than 0.05. The results from the three tests of Breusch-Pagan, Koenker-Bassett and White show that their p-values are all close to zero, which indicates that we can reject the null hypothesis that there is no heteroscedasticity. They are consistent with each other. We used the Jarque-Bera test to check the normality of residuals, whose p-value is almost equal to zero, meaning that we can reject the null hypothesis that there is normal distribution in the residuals. So, there is a problem of Non-normality.

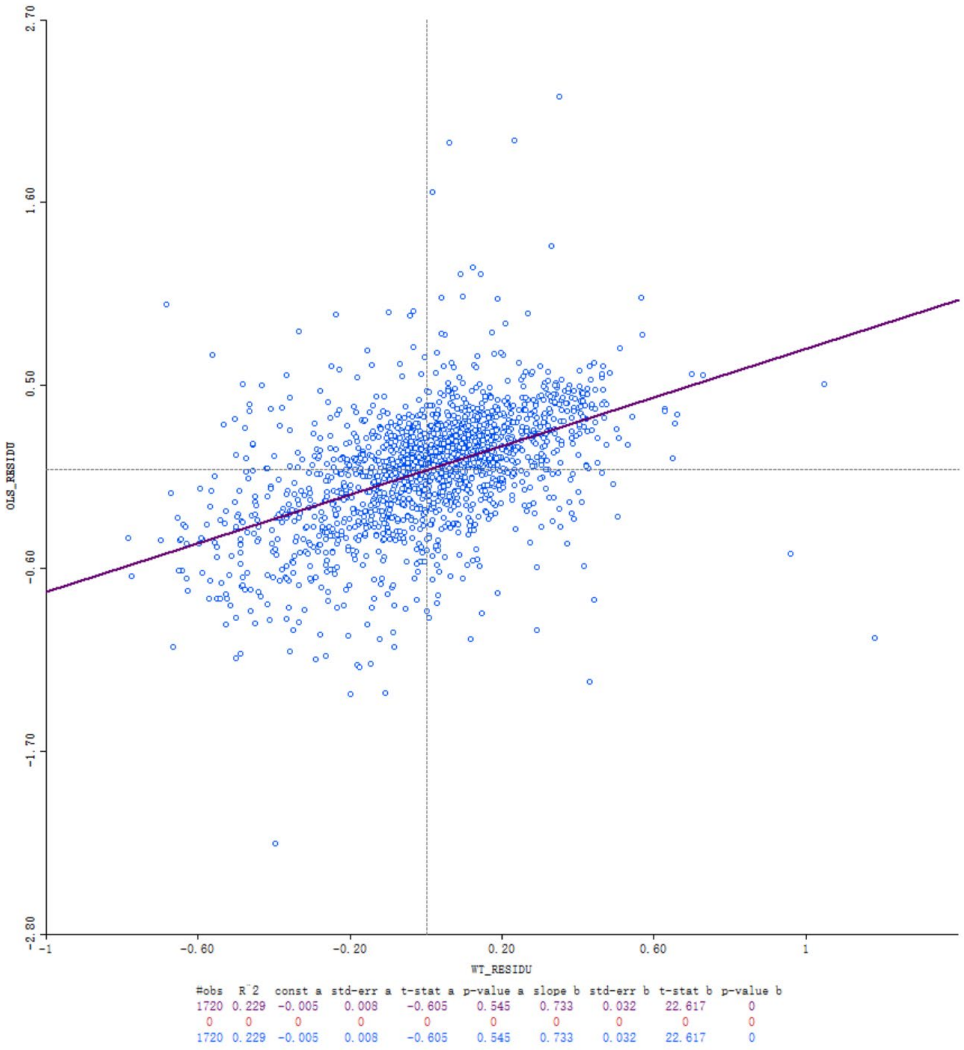

Figure 5. OLS Residuals VS WT Residuals

Figure 5 shows the relationship between the OLS residuals and weighted residuals. The weighted residuals are calculated through averaging the residuals of the queen neighbors at each unit. From Figure 5, we can find that as the weighted residuals of each unit increase, the residuals increase, so there is a positive relationship between each unit and its neighbors. The best fit line has a slope of 0.733 and the p-value is close to zero, meaning that the correlation between them is highly significant and also represents the highly spatial correlation.

From Figure 6, we can know that the Moran's I of OLS residuals is 0.313, which indicates the positive spatial autocorrelation. The pseudo-p-value of the 999 random

permutation tests for the OLS residuals is 0.001, which is highly significant as it's far less than 0.05. So we can reject the null hypothesis that there is no spatial autocorrelation. This is problematic because a vital assumption of OLS regression is that observations are independent of each other, but when spatial autocorrelation is present, the assumption would not hold.

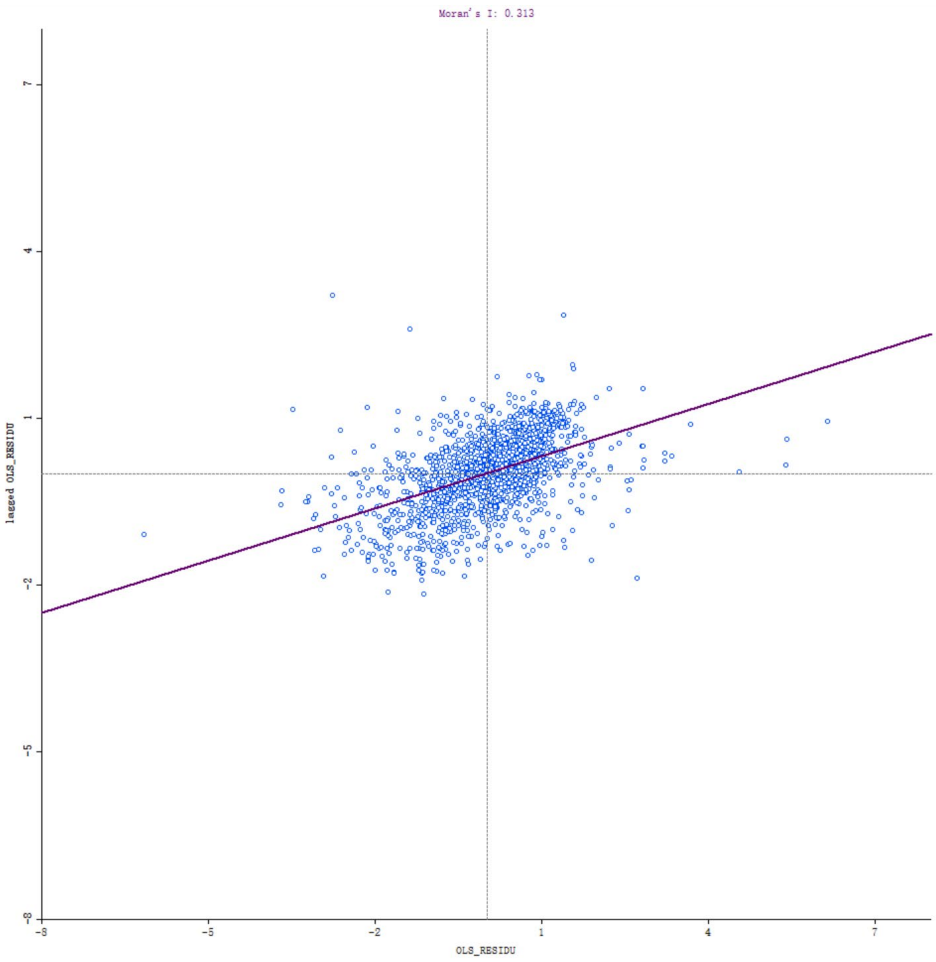

Figure 6. Moran's I for OLS residuals of median house values

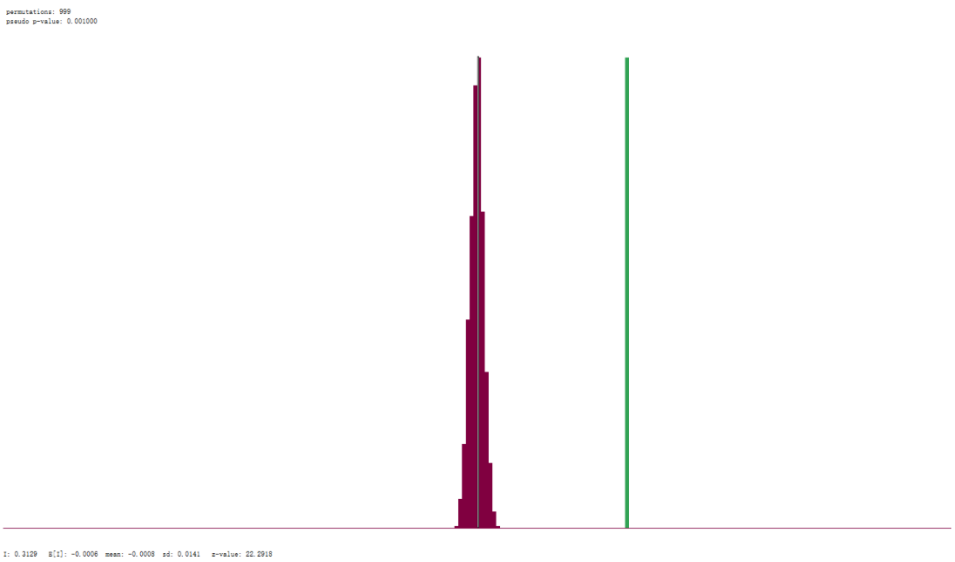

Figure 7. Histogram of Moran's I for random permutation test for OLS residuals

3.3 Spatial Lag and Spatial Error Regression Results

SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set : Regression Data
Spatial Weight : Regression Data
Dependent Variable : LNMEDHVAL Number of Observations: 1720
Mean dependent var : 10.882 Number of Variables : 6
S.D. dependent var : 0.62972 Degrees of Freedom : 1714
Lag coeff. (Rho) : 0.651097

R-squared : 0.818564 Log likelihood : -255.74
Sq. Correlation : - Akaike info criterion : 523.48
Sigma-square : 0.071948 Schwarz criterion : 556.18
S.E of regression : 0.268231

Variable	Coefficient	Std.Error	z-value	Probability
W_LNMEDHVAL	0.651097	0.0180501	36.0716	0.00000
CONSTANT	3.89845	0.201114	19.3843	0.00000
PCTVACANT	-0.0085294	0.000743667	-11.4694	0.00000
PCTSINGLES	0.00203342	0.00051577	3.9425	0.00008
PCTBACHMOR	0.00851381	0.000521935	16.312	0.00000
LNNBELPOV	-0.0340547	0.00629287	-5.41163	0.00000

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	4	220.3884	0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : Regression Data

TEST	DF	VALUE	PROB
Likelihood Ratio Test	1	911.5067	0.00000

OBS	LNMEDHVAL	PREDICTED	RESIDUAL	PRED ERROR
1	12.324	12.34227	-0.07748	-0.01841
2	12.111	12.44037	-0.35901	-0.32915
3	12.324	12.18353	0.33527	0.14033

Table 2 Spatial Lag Regression Results

We run the spatial lag model for LNMEDHVAL with the lagged variable (W_LNMEDHVAL) and the four original predictors. The lagged variable is obtained from the Queen weight matrix, which means the neighbors of the dependent variable in up, down, left, right, and diagonal directions are accounted in the model. The results

are shown in Table 2, we can find that the lagged variable has a coefficient (ρ) that is equal to 0.651097, it can be interpreted as the spatial autocorrelation between nearby observations and the one we are focused is very strong because ρ is constrained between -1 and 1. And the p-value is less than 0.05, which indicates that the variable is significant in the model.

The remain independent variable (LNNBELPOV, PCTBACHMORE, PCTSINGLES, PCTVACANT) are also significant because the p-values are less the 0.05 significance level. In both spatial lag regression and OLS regression, these four independent variables have extremely low p-values, which means these variables are considered significant. However, there is still heteroscedasticity in the model even with the lagged variable added because Table 2 shows an extreme low p-value of Breusch-Pagan test. The low p-value indicates that we can reject the null hypothesis that there is no problem of heteroscedasticity.

To check whether the spatial lag model is better than the original OLS model, we have run the following tests. The first test is Akaike Information Criterion (AIC), and the spatial lag model returned a result equal to 523.48, and the OLS model returned an AIC value equals to 1432.99. We can conclude that the spatial lag model is performing better because its AIC is nearly three times as low as that of the OLS model. The log likelihood of the spatial lag model is -255.74, and that of the OLS model is -711.493, and spatial lag model is performing better than OLS model. Finally, the probability of the likelihood ratio test is extremely small, and we can reject the null hypothesis that the OLS model is better, and conclude the spatial lag model is better than OLS model. Therefore, we can confidently say that spatial lag model is better than OLS model based on the three test results.

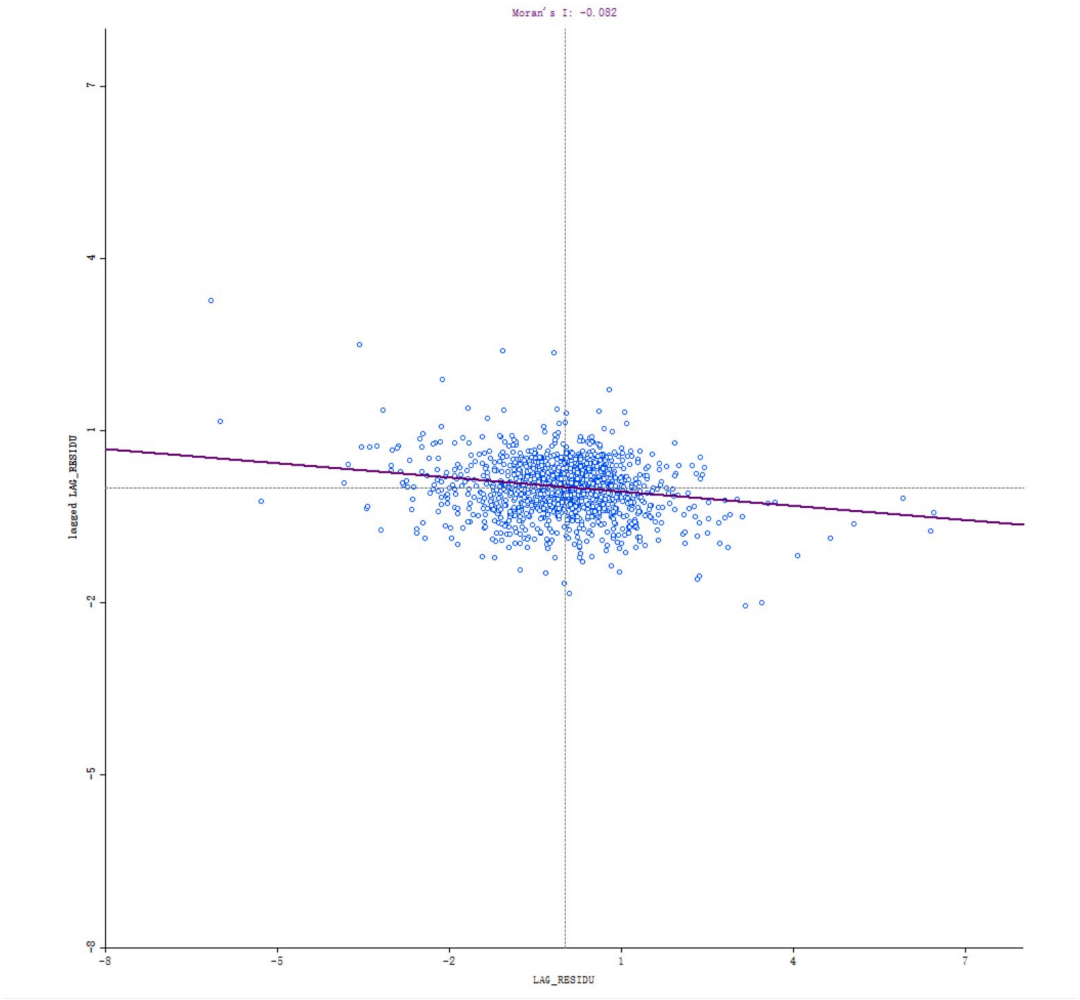

Figure 8. Moran's I for Spatial Lag Residuals of Median House Values

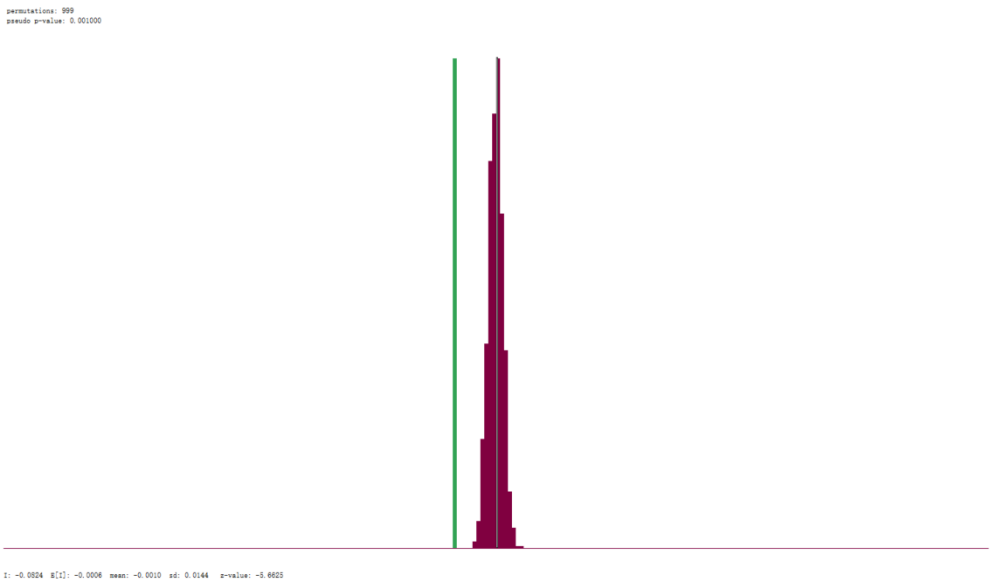

Figure 9. Histogram of Moran's I for Random Permutation Test for Spatial Error Residuals

The Moran's I scatterplot of the spatial lag regression residuals shows the Moran's I value equals to -0.082, which is smaller than the OLS residual Moran's I value (0.313). A large Moran's I value close to 1 or a small Moran's I value close to -1 both indicate spatial autocorrelation, and Moran's I value of OLS regression shows there is spatial autocorrelation that nearby observations tend to be correlated with each other. Based on the histogram in Figure 9, the Moran's I value doesn't equal to the expected Moran's I value, the gap between the calculated value and expected value illustrates that spatial autocorrelation exists, but the model is still performing better than the OLS model because the smaller Moran's I value. Overall, the spatial lag regression is performing better in accounting spatial autocorrelation, than OLS regression, and it is a more appropriate model to predict LNMEDHVAL.

SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set : Regression Data
Spatial Weight : Regression Data
Dependent Variable : LNMEDHVAL Number of Observations: 1720
Mean dependent var : 10.882000 Number of Variables : 5
S.D. dependent var : 0.629720 Degrees of Freedom : 1715
Lag coeff. (Lambda) : 0.814918

R-squared : 0.806957 R-squared (BUSE) : -
Sq. Correlation : - Log likelihood : -372.690368
Sigma-square : 0.0765508 Akaike info criterion : 755.381
S.E of regression : 0.276678 Schwarz criterion : 782.631

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	10.9064	0.0534678	203.981	0.00000
PCTBACHMOR	0.00981293	0.000728964	13.4615	0.00000
PCTVACANT	-0.00578308	0.000886701	-6.52201	0.00000
PCTSINGLES	0.00267792	0.000620832	4.31343	0.00002
LNNBELPOV	-0.0345341	0.00708933	-4.87127	0.00000
LAMBDA	0.814918	0.016373	49.7719	0.00000

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	4	210.9923	0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : Regression Data

TEST	DF	VALUE	PROB
Likelihood Ratio Test	1	677.6059	0.00000

OBS	LNMEDHVAL	PREDICTED	RESIDUAL	PRED ERROR
1	12.324	11.49305	0.12929	0.83081
2	12.111	11.56699	-0.15100	0.54423
3	12.324	11.36926	0.58402	0.95460
4	11.64	11.34593	-0.23457	0.29452

Table 3. Spatial Error Regression Results

The components of spatial error regression are similar to OLS regression, except there is an extra term, $W\varepsilon$ (the lagged residuals) . Consequently, in the spatial error regression summary table we have a coefficient called LAMBDA, which is the coefficient for the lagged residuals. As shown in Table 3, LAMBDA equals to 0.814918 and it is significant because the null hypothesis stating that this term is non-significant is rejected by the extreme low probability (close to zero). Therefore, we can conclude that

each residual of the OLS model is related to its neighboring residuals defined by the queen weights matrix, and it is necessary to include the lagged residuals term to account for spatial autocorrelation. The high value of LAMBDA illustrates that the OLS regression doesn't do a great job to account for the spatial component, and lambda term proves that there is spatial autocorrelation.

The rest of the independent variables in the spatial error model, LNNBELPOV, PCTBACHMORE, PCTSINGLES, and PCTVACANT are considered significant in the model because these variables are critical to predict the median household income, which is the same conclusion as the OLS regression model.

To evaluate the performance of spatial error model, the first test we need check is the Breusch-Pagan test. The null hypothesis for this test is that there is homoscedasticity. However, the test returned an extreme low p-value (~0.00000), the null hypothesis is rejected and there is heteroscedasticity in the model still. The next test we can check is the Akaike Information Criterion (AIC) test, the spatial error model returned an AIC value 755.381, and the OLS model returned 1432.99. Similar results are observed for log likelihood, the spatial error model returned -372.6903 for log likelihood and the OLS model returned -711.493. The spatial error model had a smaller AIC value and a larger log likelihood than OLS model, thus, the spatial error model works better at this point. To prove this statement, a likelihood ratio test is needed, we observed an extreme low p-value (~0.00000), which favored the alternative hypothesis that the spatial error model is better.

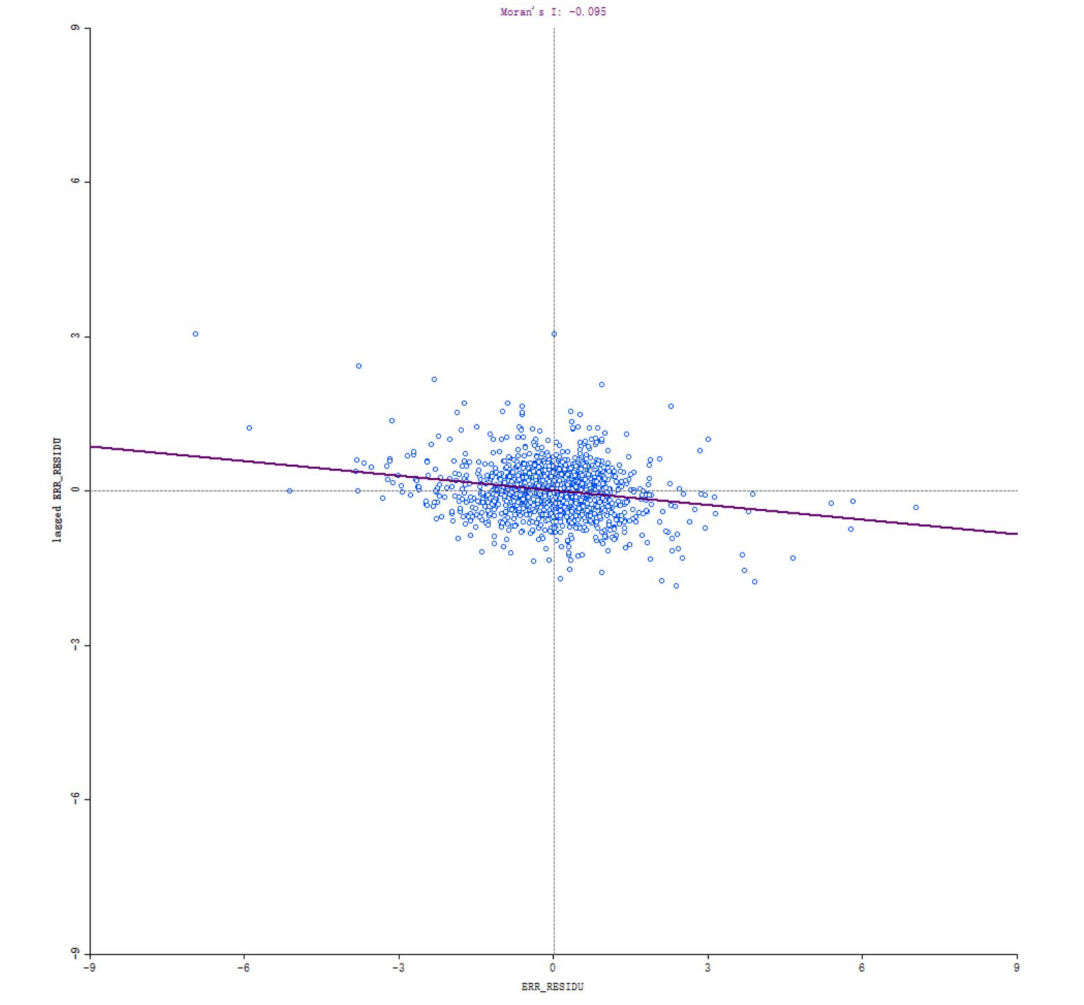

Figure 10. Moran's I for Spatial Error Residuals of Median House Values

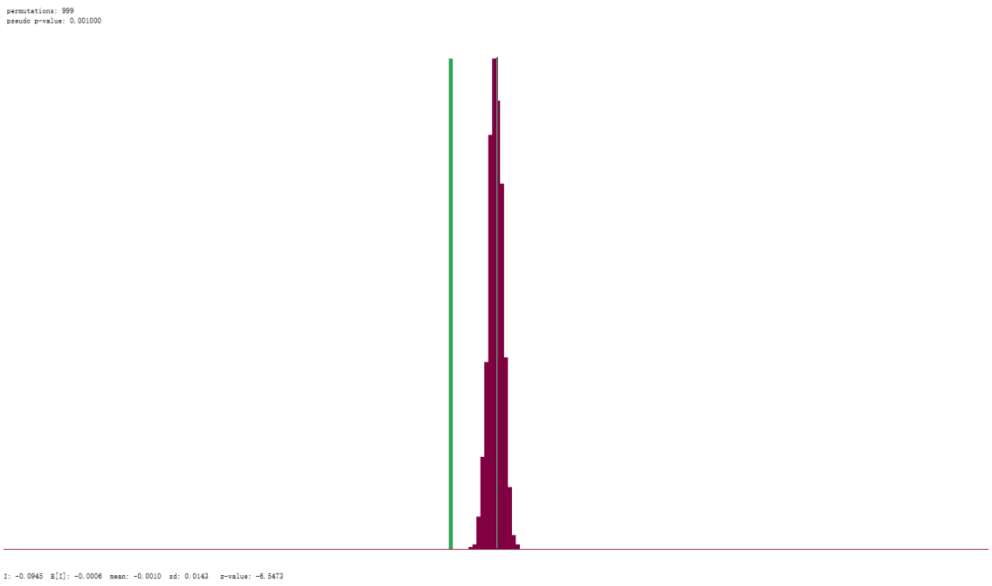

Figure 11. Histogram of Moran's I for Random Permutation Test for Spatial Error Residuals

The Moran's I scatterplot of spatial error model showed a Moran's I value equals to -0.095, which is what we expected because the closer Moran's I value is to 0, the less spatial autocorrelation within the model. On the other hand, the Moran's I value of OLS regression is 0.313, which illustrates that there are some degrees of clustering. Based on the histogram of Moran's I values for the 999 permutations, the Moran's I value doesn't equal to the expected Moran's I value, the gap between the calculated value and expected value illustrates that spatial autocorrelation exists, but we can say that the model is performing better than the OLS model because the smaller Moran's I value shown in the plot. In conclusion, the spatial error model is better than the OLS model because the tests we performed above all favored the spatial error model than OLS model.

From comparing OLS model with spatial lag model and spatial error model, we discovered that both models performed better than OLS model, but it is still unclear which model is the best to use for predicting median household income. One method we can use to determine which model is better is check their AIC values. The spatial lag model had an AIC value equal to 523.48 and spatial error model has an AIC value equal to 755.381. For Schwarz criterion, the spatial lag model returned 556.18 and spatial error model returned 782.631. We can conclude that the spatial lag model is better model to use since it is a model that has better quality based on the AIC and Schwarz criterion values. We can't use the likelihood-ratio test for this because the models are not nested.

3.4 Geographically Weighted Regression Results

GWR_supp

OID	VARNAME	VARIABLE	DEFINITION
0	Neighbors	166	
1	ResidualSquares	126.275971	
2	EffectiveNumber	171.047974	
3	Sigma	0.285523	
4	AICc	668.91665	
5	R2	0.814861	
6	R2Adjusted	0.794536	
7	Dependent Field	0	LNMEDHVAL
8	Explanatory Field	1	PCTBACHMOR
9	Explanatory Field	2	PCTSINGLES
10	Explanatory Field	3	PCTVACANT
11	Explanatory Field	4	LNNBELPOV

Table 4. Results of GWR Regression

According to Table 4 above, the AIC value of GWR is approximately 668.91. According to figure 4 in the previous sections, the AIC value for OLS regression is 1432.99, AIC value for spatial lag and spatial error are 523.48 and 755.38 respectively. Therefore, it is suggested that the GWR could work better in fitting the model than the OLS regression and spatial error model. And the spatial lag model could work better than GWR. Also, based on Table 4 representing, the R-squared value of the regression is approximately 0.81, and the adjusted R-squared value is nearly 0.79. Both of them are larger than the R-squared value of the OLS regression which is about 0.67. The difference in R-square values between the two regressions reveals that the GWR might better fit the model by representing 12% more variance in the dependent variable than the OLS regression.

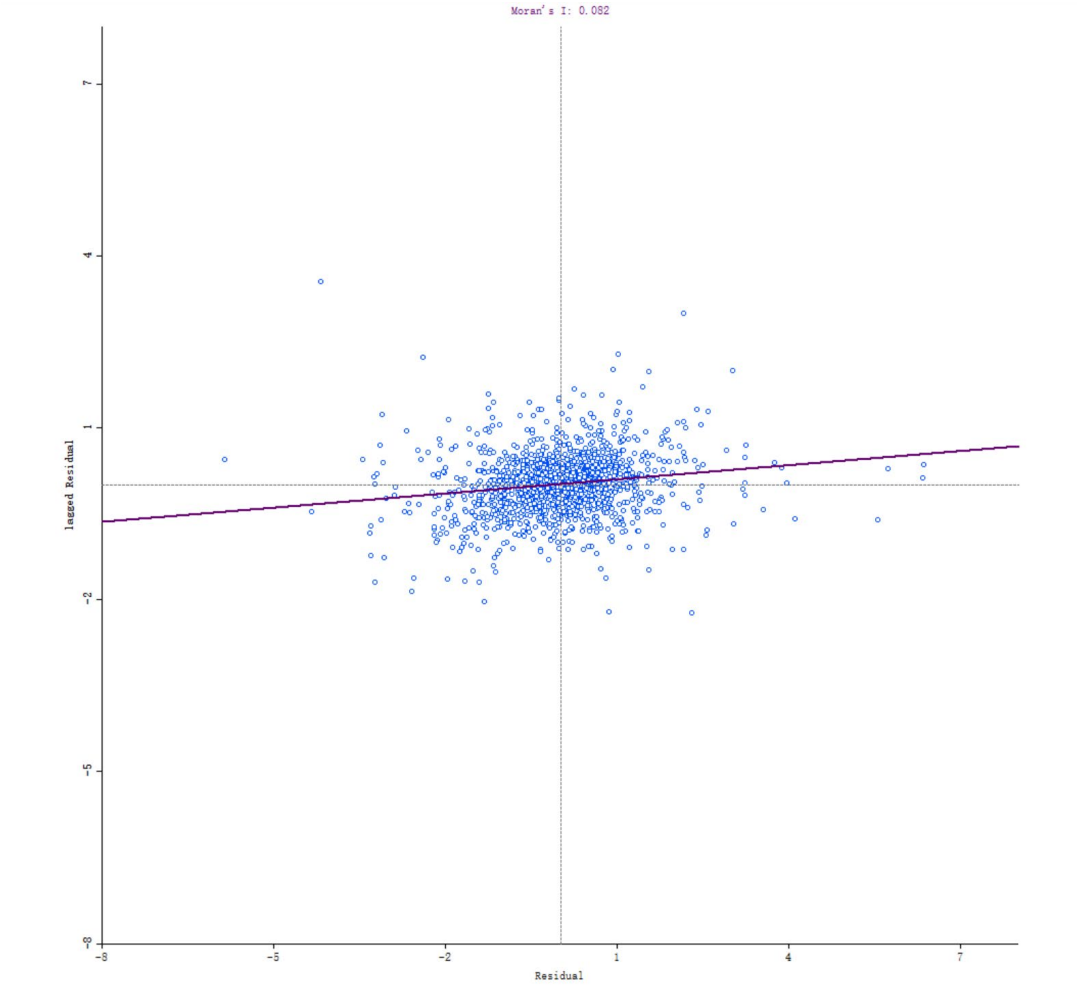

Figure 12. Moran's I for GWR residuals of Median House Values

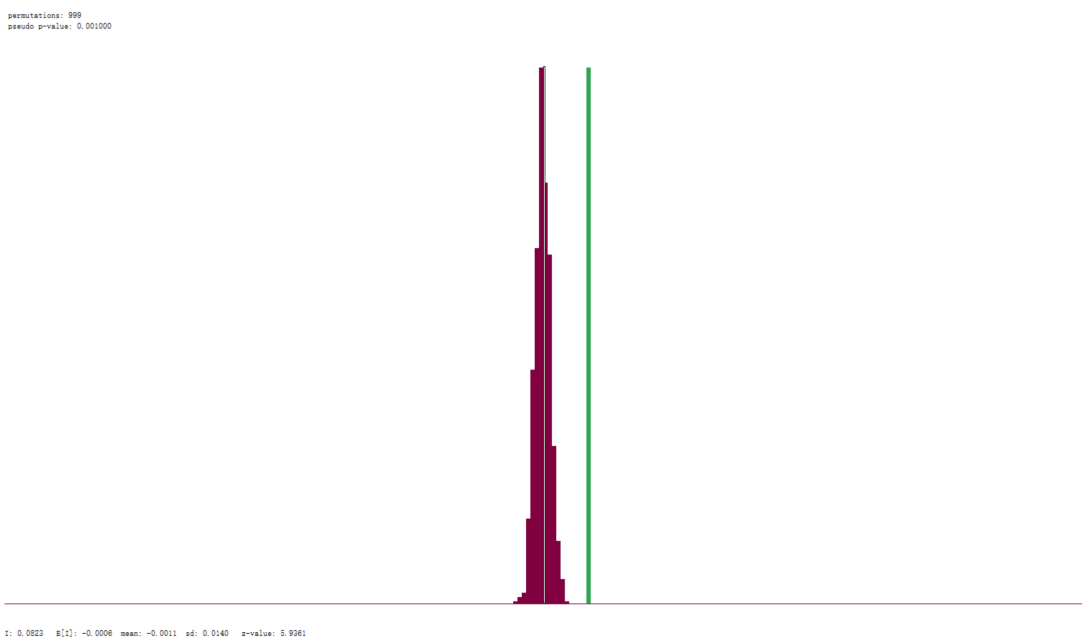

Figure 13. Histogram of Moran's I for Random Permutation Test for GWR Residuals.

According to Figure 12, the Moran’s I of GWR is approximately 0.082. It is less than the Moran’s I of OLS regression (0.313), and it is closer to zero value, indicating that the GWR model is better in describing the spatial variances of the LNMEDHVAL. The Moran’s I for spatial error residuals and spatial lag residuals are -0.095 and -0.082 respectively. The spatial lag model is the least spatial autocorrelation in residuals. And the spatial lag model as it has the largest value of Moran’s I that closest to - 0.0006 (the expected value of Moran’s I as shown in the lower left of the Figure 13). Furthermore, according to Figure 13, which demonstrates the Moran's I of 999 random permutation test on GWR residuals, the GWR residuals have less spatial autocorrelation than the OLS residuals, and the GWR residual and Moran's I are close to the expected Moran's I random distribution. Nevertheless, the p-value is 0.001, showing that the GWR residuals are not random.

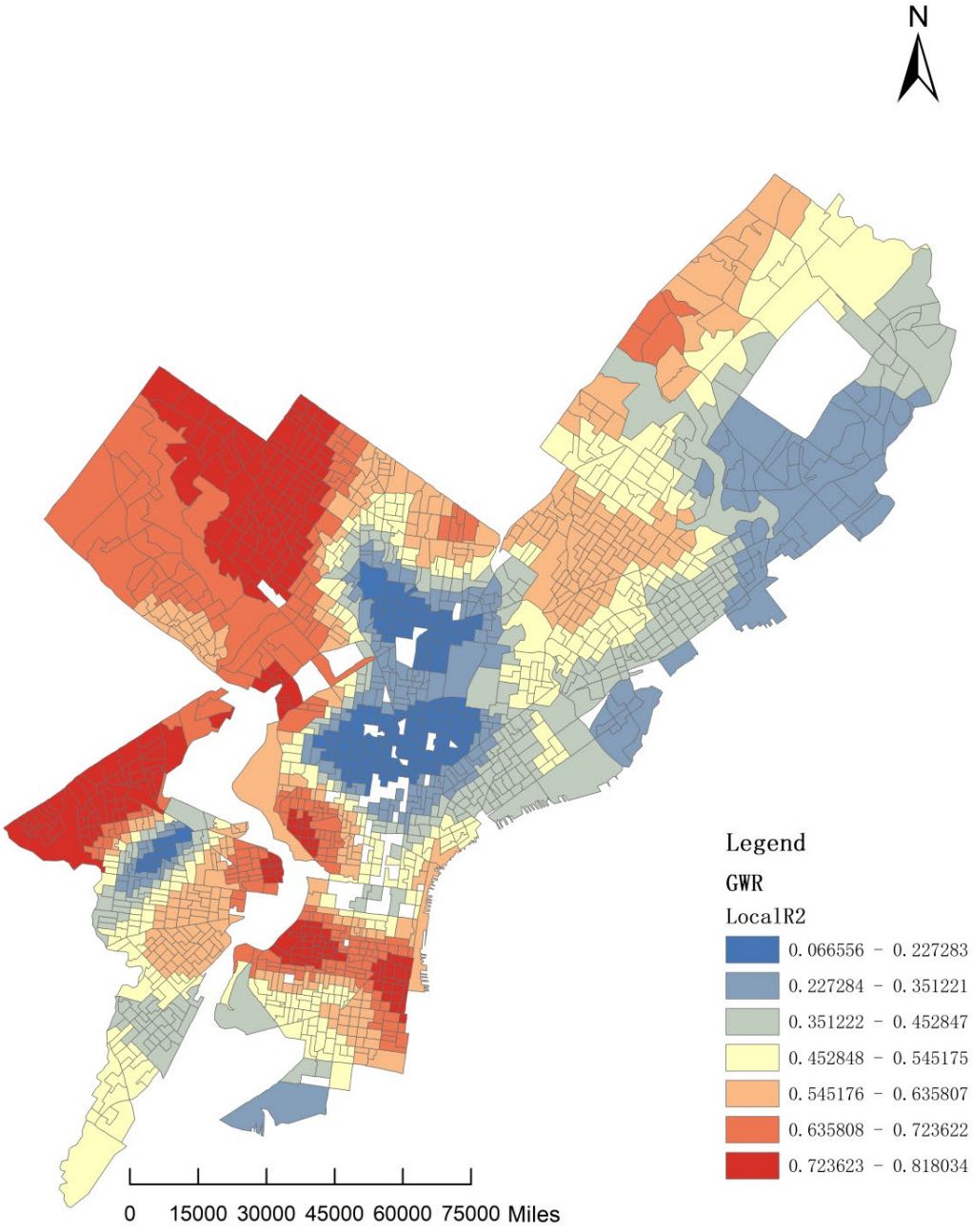

Figure 14. The geographical distribution of the local R-Squared values of GWR

Figure 14 represents the geographical distribution of the local R-Squared values of GWR in Philadelphia. The local R-Squared values of the northwest and southeast Philadelphia are approximately in a range from 0.72 to 0.81 respectively filled with a red color in the map, which are the highest among the whole city. It shows that the predictors account for roughly 72% to 82% of the variance of the median home value.

Also, for the central area of Philadelphia and a small area in the southwest of it are with relatively the lowest local R-Squared value ranging from 0.06 to 0.22 with a dark blue color, representing that the selected predictors of this project can't explain the situation of the median house values of these areas, there are 6% to 22% of the variance of the median house value described. Furthermore, small percentage of areas in north and south east of the city are also relatively with low local R-Squared values in a color of light blue.

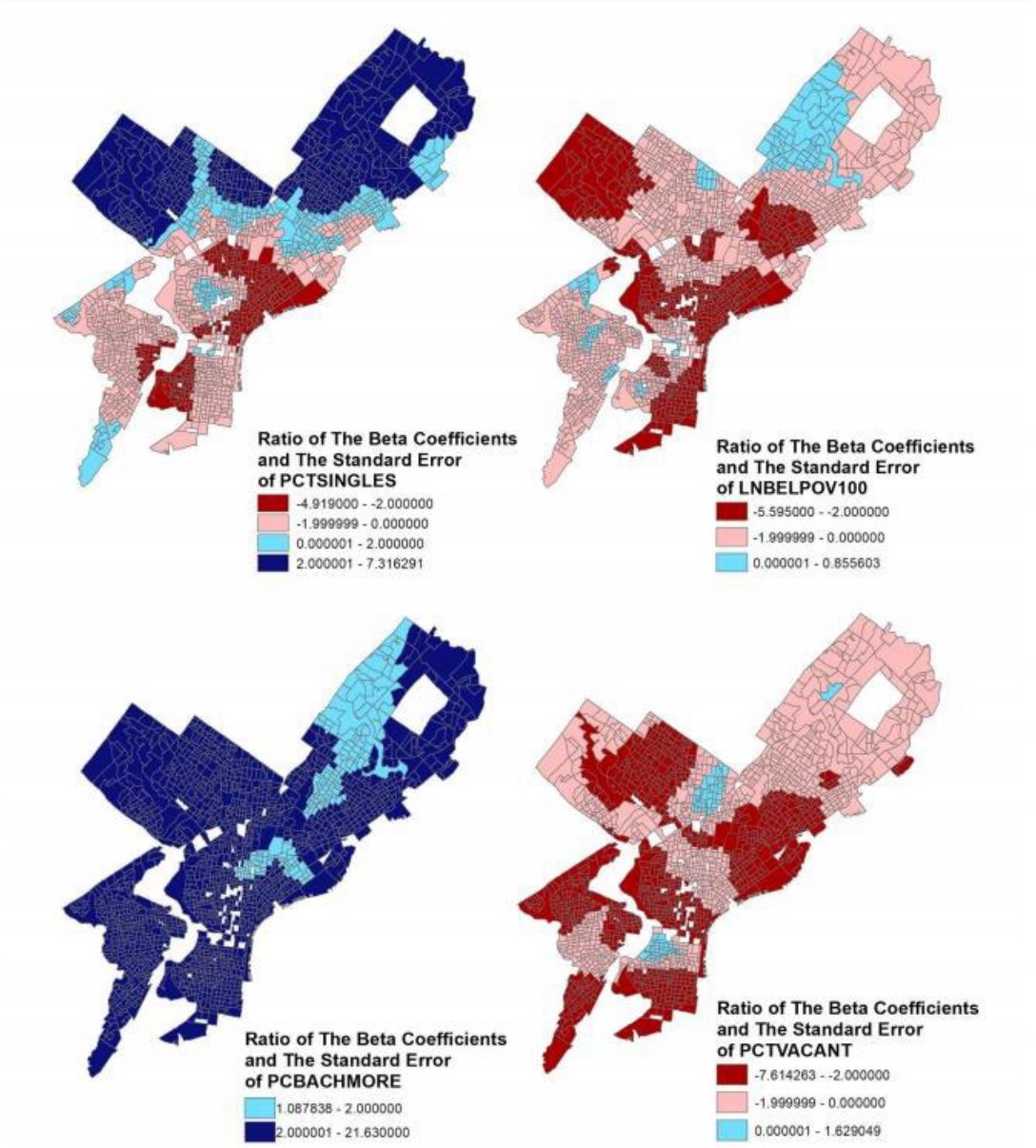

Figure 15. The local regression results of GWR model

The figure 15 represents the local regression results of the GWR model. A beta coefficient is a parameter that could measure the volatility of the individual median house value. In this case, according to figure 15, the predictors have a larger positive impact on median housing values in north and central part of Philadelphia, whereas they have a greater negative affect on median house values in South Philadelphia and a small part of Northwest Philadelphia. Also, in most cases, the predictors of percentage of vacant houses and poverty number could affect the house values negatively. And the percentage of people with higher education and the number of single houses could affect the house values positively.

4 Discussion

In this assignment, we tried to account for spatial autocorrelation in the dataset since the results produced by the OLS model showed clear evidence of clustering. To fix the clustering issue, we fitted three different models, spatial lag regression, spatial error regression, and geographically weighted regression. Based on the test results we discussed, we discovered that all three models performed better than the OLS model, and this result met our expectation of that the new models should accounting spatial autocorrelation when predicting median house values. Moreover, spatial lag model is the best model among all because the lowest AIC values, a closer Moran's I value to the expected value than the rest of the models. However, the models are not perfect, the assumptions of less heteroscedastic was not met as the Breusch-Pagan test of both spatial lag and spatial error showed heteroscedasticity in the model. In addition, since the p-values are still less than 0.05, there still exists spatial autocorrelation associated with the residuals for all three models. The three models are still can't fully solve the spatial autocorrelation issue.

5 References

[1] Tobler, Waldo R. "A computer movie simulating urban growth in the Detroit region." Economic geography 46.sup1 (1970): 234-240.

[2] Miller, Harvey J. "Tobler's first law and spatial analysis." Annals of the association of American geographers 94.2 (2004): 284-289.