Homework 2: Using Spatial Lag, Spatial Error, and Geographically Weighted Regression to Predict Median House Values in Philadelphia Block Groups

1. Introduction

One of the most well-liked research areas in the study of real estate value is house price prediction, which is important because it can aid in government and urban planning decision-making. In the previous report, we conducted OLS regression to examine the relationship between median house values and several neighborhood characteristics, including the number of households living in poverty, the percentage of individuals with a bachelor's degree or higher, the percent of vacant houses, and the percent of single house units. However, OLS regression can be inappropriate when dealing with datasets that have a spatial component. Because spatially autocorrelated OLS residuals will lead to systematic under-prediction or over-prediction in certain parts of the study region, furthermore, the significance estimates for the β coefficients in OLS may be incorrect.

Our goal in this research is to investigate any spatial autocorrelation in neighborhood variables and investigate ways to potentially enhance the findings. Using Philadelphia data at the Census block group level, we will use GeoDa and ArcGIS to run spatial lag, spatial error and geographically weighted regression to see whether these methods can explain the spatial autocorrelation that might remain in the OLS residuals., using Philadelphia data at the Census block group level.

2. Methods

It is believed that Waldo Tobler's (1970) conclusion of **the first law of geography**, "Everything is related to everything else, but near things are more related than distant things. That is all spatial stats have relations with each other at near locations. Spatial autocorrelation is similar to correlation but focuses more on the relationships of values within a single variable and variables nearby. A variable is considered having positive spatial autocorrelation if it has related values with variables close in space.

a) A Description of the Concept of Spatial Autocorrelation

Moran's I (1950) is perhaps the most widely used method of testing for spatial autocorrelation or spatial dependencies which can be defined mathematically as below:

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

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

Where:

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

When there is strong positive autocorrelation, Moran's I tend to approach large positive values (close to 1), which means similar values tend to cluster together. On the contract, when there is strong negative autocorrelation, Moran's I tend to approach large negative values (close to -1), which means similar values tend to dispersion. Moran's I's expected value is -1/(n-1) which means that there is no spatial autocorrelation, where n is the total number of observations.

When there are n observations, an n times n table can be generated to summarize all the pairwise spatial relationships in the dataset. This table is called a **weight matrix** which can be used in the estimation of spatial regression and the calculation of spatial autocorrelation as well. Usually, it is better to try several different weight matrices to guarantee that the results are not an artifact of the matrix which is being used. In this project, we will use GeoDa to generate a Queen contiguity 1720 * 1720 weight matrix with data set which contains 1720 block groups. That is all 1720 block groups that cross with census block group A at a point, or a segment is considered to be A's neighbors.

With the **hypothesis** below, Moran's I is used to test whether spatial dependence exists and whether spatial autocorrelation is significant:

- Ho: No spatial autocorrelation, which means there is no spatial autocorrelation between median house values and four neighborhood characteristics.
- H_{a1}: There is a positive spatial autocorrelation between median house values and four neighborhood characteristics.
- H_{a2}: There is a negative spatial autocorrelation between median house values and four neighborhood characteristics.

To compute Moran's I for house price variables, the first step is to randomly shuffle the values of the house price variables, and then to calculate Moran's I for this shuffled map, then repeat the shuffling for another 999 times and calculate Moran's I each time. Next, put the total 1000 Moran's I values in descending order and compare the Moran's I value for the observed house price variable with the Moran's I values for the random permutations.

To better understand spatial autocorrelation, Luc Anselin created the **Local Indices of Spatial Autocorrelation (LISA)**, which can describe the degree to which data at sites close to i are related to i. For each location i, the local Moran's I is computed. In general, the average deviation of a value's nearby values is compared to a value's divergence from the global mean at position i. The block groupings are divided into four categories by the analysis:

- High X_i & High X_j + SA: The deviation of the i^{th} location from the global mean and the average deviation of neighborhood locations j from the global mean are both positive. It suggests that there is positive autocorrelation.
- Low X_i & High X_j – SA: The deviation of the i^{th} location from the global mean is negative while the average deviation of neighborhood locations j from the global mean is positive. It suggests that there is negative autocorrelation.
- Low X_i & Low X_j + SA: The deviation of the i^{th} location from the global mean and the average deviation of neighborhood locations j from the global mean are both negative. It suggests that there is positive autocorrelation.
- High X_i & Low X_j – SA: The deviation of the i^{th} location from the global mean is positive while the average deviation of neighborhood locations j from the global mean is negative. It suggests that there is negative autocorrelation.

* Here SA refers to spatial autocorrelation

Local LISA statistic (i.e., Local Moran's I) could be positive or negative at location i, statistical significance testing hypothesis is generated as following:
H_o: When the local Moran's I is near to 0, there is no (local) spatial autocorrelation at point i. Therefore, there is no correlation between the values of our variable at location i and those of its neighbors at site j.

H_a: There is positive or negative spatial autocorrelation at location I when local Moran's I is not 0. That is, values of our variable at location i are very similar to (+SA) or starkly different from (-SA) from the nearby locations j.

b) A Review of OLS Regression and Assumptions

The OLS regression is often used to examine the relationship between a variable of interest and one or more explanatory variables. It allows us to calculate the amount of which dependent variable changes when a predictor variable changes by one unit (holding all other predictors constant). In previous report, we have conducted multiple regression and assumed that all the predictors we used were linearly correlated with the dependent variable. Then we ran stepwise regression to see the significance of each predictor and checked the problem of multicollinearity. Finally, we use cross-validation to predict the performance of our model.

This OLS regression is based on assumption below:

- There are linear relationships between dependent variable y and each of the predictors x.
- Residuals are normally distributed.
- Homoscedasticity - the variance of the residuals is constant regardless of the value of each x (or the value of y predicted by the model).
- Observations are independent. That is, there should be no spatial, temporal or other forms of dependence in the data.
- Predictors should not be strongly correlated with each other.
- No fewer than 10 observations per predictor.

The problem with residuals arises when **the data has a spatial component**, that is the assumption that observations/errors are random/independent often doesn't hold. Computing Moran's I of the residuals is a way to test OLS residuals for spatial autocorrelation. An ideal Moran's I should be close to 0. Another way to test OLS residuals for spatial autocorrelation is to regress $\hat{\varepsilon}$ on nearby residuals $W\hat{\varepsilon}$. In this report, nearby residuals refer to residuals at neighboring block groups which are defined by the Queen matrix. Ideally, there should be no relationship between $\hat{\varepsilon}$ and $W\hat{\varepsilon}$, that is the coefficient of $W\hat{\varepsilon}$ denoted by λ (as opposed to β_1) is not significantly different from 0. (λ can range between -1 and 1).

The model incorporates the spatial lag of the dependent variable as a predictor, presuming that the value of the dependent variable at one place is related to the values of that variable in the neighboring locations. The spatial autocorrelation is indicated by the rho(ρ) in this spatial lag model. The slope of the fitted line of regression between the OLS residuals and their surrounding residuals is known as ρ. It illustrates how the residuals and their neighbors are related.

We can **test additional regression hypotheses in GeoDa**, the program we're using to do your OLS regression. The first is the homoscedasticity assumption, which is related to the assumption of error independence. We assumed that the regression residuals should be random noise. However, the variance in the residuals may change with values of another variable, that is heteroscedasticity. We can first save OLS residuals and anticipated values and perform a residual-by-predicted scatter plot in GeoDa to test for heteroscedasticity. Then, GeoDa offers the Breusch-Pagan Test, the Koenker-Bassett Test, and the White Test as three distinct diagnostics for heteroscedasticity. The Breusch-Pagan Test and the Koenker-Bassett Test will be used. The null hypothesis here is that of homoscedasticity (No heteroscedasticity). If the p-value is less than 0.05, then we can reject the null hypothesis for the alternate hypothesis of heteroscedasticity. The assumption of the normality of the residual/error is another one. The Null Hypothesis that the residuals are from a normal distribution is tested using the GeoDa Jarque-Bera test. If p 0.05, the alternative hypothesis of non-normality should be accepted instead of the null hypothesis of normality.

c) Spatial Lag and Spatial Error Regression

We will cover two spatial regression models, **spatial lag and spatial error regressions**, available in GeoDa. The geographic regression models incorporate factors, known as rho (ρ) in the spatial lag model and lambda (λ) in the spatial error model, which account for spatial autocorrelation in addition to the independent variables entered the OLS model.

First, in **spatial lag model**. Assumes that a dependent variable's value at one place will be related to its value at surrounding locations as defined by the weights matrix W. This indicates that the dependent variable's spatial lag is a predictor in the model. Here, the y-lag variable Wy's coefficient is ρ(rho), just as the variable X1's coefficient is β1.

$$OLS: \quad y = \underbrace{\beta_0}_{\substack{Intercept \\ (Constant)}} + \underbrace{\beta_1 X_1 + \cdots + \beta_n X_n}_{Predictors} + \underbrace{\varepsilon}_{Residuals}$$

$$Spatial\ Lag: \quad y = \underbrace{\rho W y}_{Lag\ of\ y} + \underbrace{\beta_0}_{\substack{Intercept \\ (Constant)}} + \underbrace{\beta_1 X_1 + \cdots + \beta_n X_n}_{Predictors} + \underbrace{\varepsilon}_{Residuals}$$

In this case, spatial lag equation is written as following:

$$LNMEDHVAL = \rho W_{LNMEDHVAL} + \beta_0 + \beta_1 PCBACHMORE + \beta_2 PCTVACANT \\ + \beta_3 LNNBELPOV + \beta_4 PCTSINGLES + \varepsilon$$

Next, in **spatial error model**. Takes for granted that the residual at one place is connected to residuals at other locations, as defined by the weights matrix W. We have a two-step (stage) regression practically: First, our OLS model, which regresses Y on the predictors, is first performed. Second, we decompose the residuals (ε) into two parts: one with a spatial pattern (λWε), and one that is just random noise. This is done by regressing the residuals on the nearest neighbor residuals, which removes the spatial information from the OLS residuals (u).

$$Spatial\ Error: \begin{cases} y = \underbrace{\beta_0}_{\substack{Intercept \\ (Constant)}} + \underbrace{\beta_1 X_1 + \cdots + \beta_n X_n}_{Predictors} + \varepsilon \\ \varepsilon = \underbrace{\lambda W \varepsilon}_{Spatially\ Lagged\ Residuals} + \underbrace{\mu}_{Random\ Noise} \end{cases}$$

$$y = \underbrace{\beta_0}_{\substack{Intercept \\ (Constant)}} + \underbrace{\beta_1 X_1 + \cdots + \beta_n X_n}_{Predictors} + \underbrace{\lambda W \varepsilon}_{Spatially\ Lagged\ Residuals} + \underbrace{\mu}_{Random\ Noise}$$

In this case, spatial error equation is written as following:

$$\begin{cases} y = \beta_0 + \beta_1 PCBACHMORE + \beta_2 PCTVACANT + \beta_3 LNNBELPOV + \beta_4 PCTSINGLES + \varepsilon \\ \varepsilon = \lambda W_{\varepsilon} + \mu \end{cases}$$

$$\downarrow$$

$$y = \beta_0 + \beta_1 PCBACHMORE + \beta_2 PCTVACANT + \beta_3 LNNBELPOV + \beta_4 PCTSINGLES + \lambda W_{\varepsilon} + \mu$$

Except for the condition of spatial independence of observation, both the spatial lag and the spatial error regression models require **the assumptions that are needed for OLS** indicated in section b. Regression with spatial lag and spatial error **aims to account for the possibility of spatial dependencies in the residuals or the data**. The residuals produced by spatial lad and spatial error approaches may no longer be spatially autocorrelated or less heteroscedastic.

However, we still presumptively assume that all the predictors are linearly connected to the DV, that the residuals are normal, and that multicollinearity should not exist.

To determine if spatial models outperform OLS, we will **compare the results of spatial lag regression with OLS and the results of spatial error regression with OLS**. The following three criteria make up the benchmark for comparison:

- Akaike Information Criterion (AIC) /Schwarz Criterion (SC). The goodness of fit of an estimated statistical model is measured by the AIC and SC. They can be used to illustrate the trade-off between the precision and complexity of the model and are a relative measure of the information lost when a certain model is employed to explain reality. The better the fit in GeoDa, the lower the AIC and SC.
- Log Likelihood is associated with the maximum likelihood method of fitting a statistical model to the data and estimating model parameters. Maximum likelihood chooses the model parameter values that make the data "more likely" than they would be given any other parameter value. The higher the log-likelihood, the better the model fit. Log likelihood is often exclusively used to compare nested models. Because the spatial regression model may be reduced to OLS when the spatial factor is dropped, OLS is a particular case of the spatial lag and spatial error models. However, because spatial lag and spatial error are not special cases of one another, we are unable to compare them using the log likelihood ratio.
- Likelihood Ratio Test compares the OLS model with the spatial model. The null hypothesis is that the OLS model is not a better specification than the spatial lag or spatial error model. If P 0.05, the null hypothesis can be successfully rejected, then it can be concluded that the spatial lag or spatial error model performs better than the OLS model.

Another way to compare the OLS results with spatial lag and spatial error results is by looking at Moran's I of regression residuals. The closer Moran's I to zero, the less spatial autocorrelation, the better the model.

d) Geographically Weighted Regression

We will do **Geographically Weighted Regression (GWR)** analyses in ArcGIS. The OLS regression, spatial lag or error regression all assume that dealing with spatial stationarity, that is the assumption that modeled relationships are constant across space. Although required, this presumption is probably untrue in reality. Recognizing that the problem is **spatial non-stationarity**, we can say that we have several **local regressions** rather than a single global regression for each point. This is Geographically Weighted Regression (GWR). According to Simpson's paradox, a trend may be present in numerous groups of data yet vanish or change direction when the groups are combined. The GWR result could differ or even reverse from the global regression result.

For each observation ***i*** (***i*** = 1…n), **the GWR model's equation** performs OLS regression on ***i*** and nearby observations ***x*** (suppose there are a total of ***m*** observations). The closer the observations to ***i***, the higher the weight:

$$y_i = \beta_{i0} + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \ldots + \beta_{im}x_{im} + \varepsilon_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik}\, x_{ik} + \varepsilon_i$$

In this report, equation can be written (***i***=1...n):

$$LNMEDHVAL = \beta_{i0} + \beta_{i1}PCBACHMOREi + \beta_{i2}PCTVACANTi + \beta_{i3}LNNBELPOVi + \beta_{i4}PCTSINGLESi + \varepsilon_i$$

Where,

- β_{i1} = when PCBACHMORE increases 1%, the LNMEDHVAL would increase β_{i1} unit, MEDHVAL increase by (e^β_{i1}-1)*100%, holding other variables constant.
- β_{i2} = when PCTVACANT increases 1%, the LNMEDHVAL would increase βi2 unit, MEDHVAL increase by (e^β_{i2}-1)*100%, holding other variables constant.
- β_{i3} = when LNNBELPOV100 increases 1 unit(NBELPOV100 increases 1 household), the LNMEDHVAL would increase β_{i3} unit, MEDHVAL increase by (1.01^β_{i3}-1)*100%, holding other variables constant.
- β_{i4} = when PCTSINGLES increases 1%, the LNMEDHVAL would increase β_{i4} dollars, MEDHVAL increase by (e^β_{i4}-1)*100%, holding other variables constant.
- ε_i = the residuals

To perform a regression for each location:

- First it requires multiple observations (locations) to run a regression, not just a single observation (location) ***i***.
- GWR does the regression using additional observations from the dataset, giving more weight to observations near to location ***i***.
- The weight of an observation varies with location ***i***.
- The assessment of the location i's parameters is more influenced by observations that are closer to ***i***.

The radius of the circular area is the **bandwidth** when the weighting function generates points in a circle centered on the target location. There are two types of bandwidth depending on whether or not the distance between regression point ***i*** and the observations stays constant. Around each point ***i*** the number of observations will vary, but the fixed bandwidth distance ***h*** (and the area) will not change. The ***distance$_{ij}$*** is the separation between data point ***j*** and regression point ***i***. Around each point ***i*** the number of observations will vary, but the fixed bandwidth distance ***h*** (and the area) will not change. While adaptive bandwidth ***h*** means that number of observations will remain fixed but the area will not be the same. Results are significantly impacted by the weights' underlying assumptions. When the distribution of observations is largely consistent over space, a fixed bandwidth kernel will be more suited (e.g., number of neighbors, size). When the distribution varies throughout space, an adaptive bandwidth kernel is acceptable (i.e., events are clustered, or polygons are heterogeneously shaped or sized). After choosing a kernel type, optimization removes part of the uncertainty, but robustness checks are still required.

A lot of **OLS assumptions mentioned above still hold in GWR**:

- Normality of residuals.
- Homoscedasticity.
- No multicollinearity (GWR requires at least 300 observations). Results are unreliable in global regression models like OLS when two or more variables display multicollinearity. To account for each feature in the dataset, GWR creates a local regression equation. There will most likely be an issue if the value of an explanatory variable spatially clusters significantly. Additionally, multicollinearity issues may arise when more than two variables exhibit comparable clustering patterns in each area.
- Dependent variable may not normal, but it is acceptable when residuals close to normal.

In a global model such as OLS, it is usual to test whether the parameter estimates are significantly different from zero. This can be accomplished with a t-test, the t-statistics and their associated p-values are usually provided on the output. However, **p-values are not part of the GWR output**. There could be hundreds, or thousands of tests needed to determine whether parameters are locally significant considering that there is one set of parameters and one set of standard errors associated with each regression point. Recall the type I error concept: if the 0.05 significance level is utilized, we would anticipate 5 out of 100 tests to be significant, but they are not! (Also, type II error might lead us to anticipate that certain outcomes that are significant are actually not significant.) For a model with four predictors and 2000 regression points, there would be 10,000 significance tests (five per point, one for the intercept and four for the predictors), and we would anticipate 500 significant tests

3. Results

a) Spatial Autocorrelation

First, we calculated the global Moran's I values for the dependent variable LNMEDHVWL and presented a scatter plot of it. In Figure 1&2, we see that Moran's I for LNMEDHVAL value is 0.794, meaning that LNMEDHVAL has a high spatial autocorrelation.

We also conducted the random permutations test on the result. It shows the Moran's I for the Median House Value is much higher than the Moran's I for all 999 random permutations. Again, we can say that the LNMEDHVAL is significantly spatial autocorrelated, and we can reject the null hypothesis that there is no spatial autocorrelation of the dependent variable.

Figure 1: Geoda Moran's I scatter plot & permutation plot

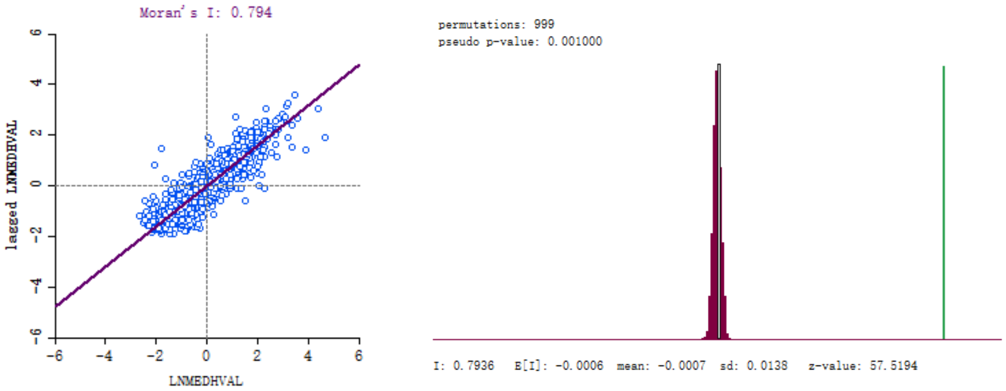
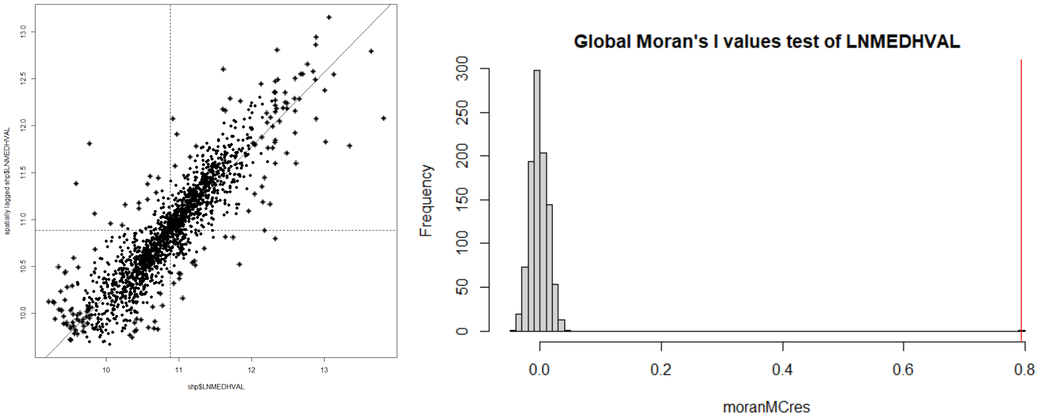


Figure 2: R Moran's I scatter plot & permutation plot



From the maps, we can see most of Philadelphia is not significant. They are the south area, the east area, the northwest area, and the northeast area. The low-low areas mainly cluster in north Philadelphia, and part of the south and west Philly. The high-high areas mainly cluster in northwest, northeast, and part of south Philly. The high-low areas and low-high areas are distributed more separately, most of them appear in south Philadelphia.

Comparing the significance map with the cluster map, in Figure 3&4, the high-high areas are more significant in northwest and south Philly, meaning that in these areas, the dependent variable has higher spatial autocorrelation. The low-low areas are more significant in north Philly, also meaning that it has high spatial autocorrelation in that area.

Figure 3: Geoda Significance Map and Cluster Map

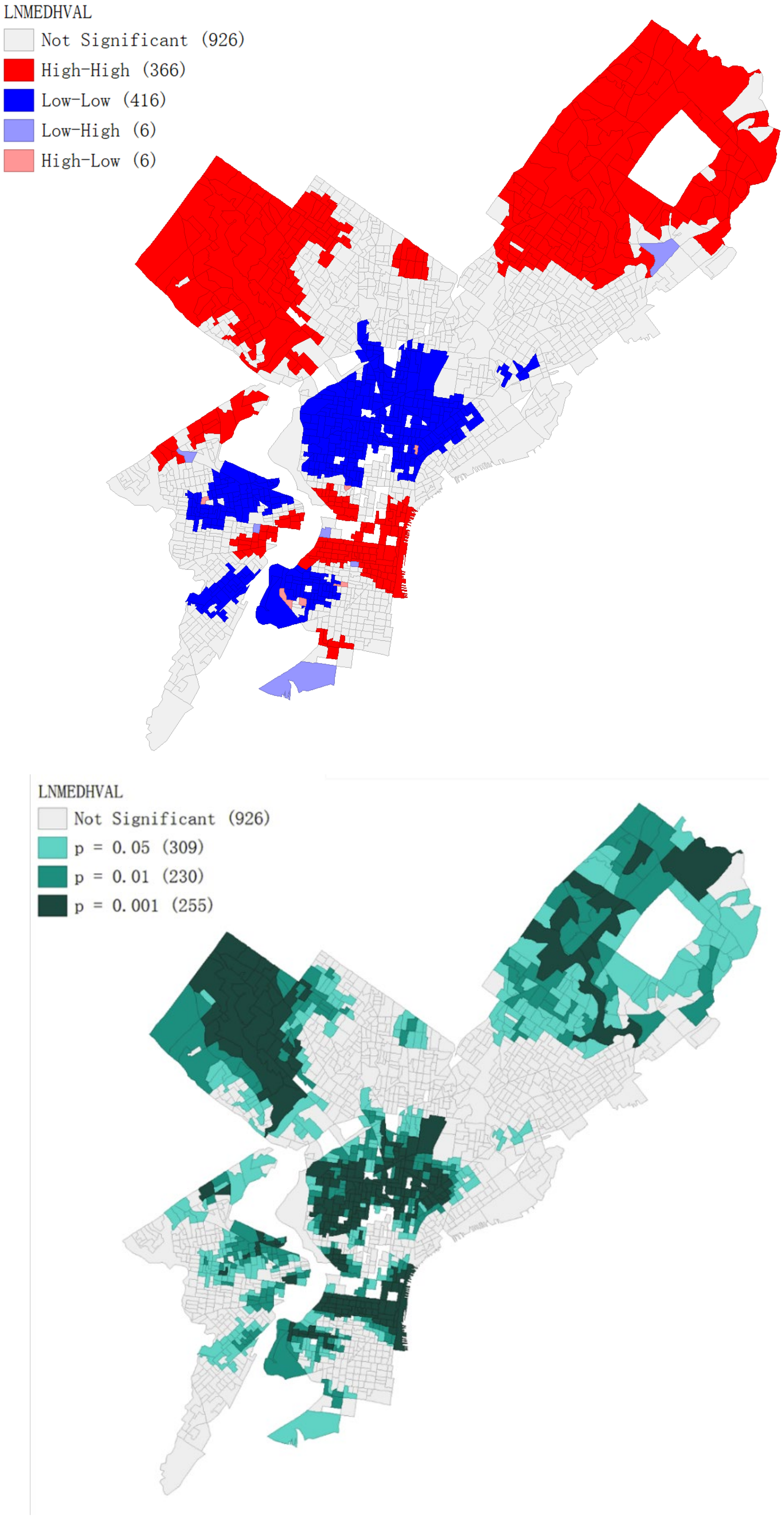
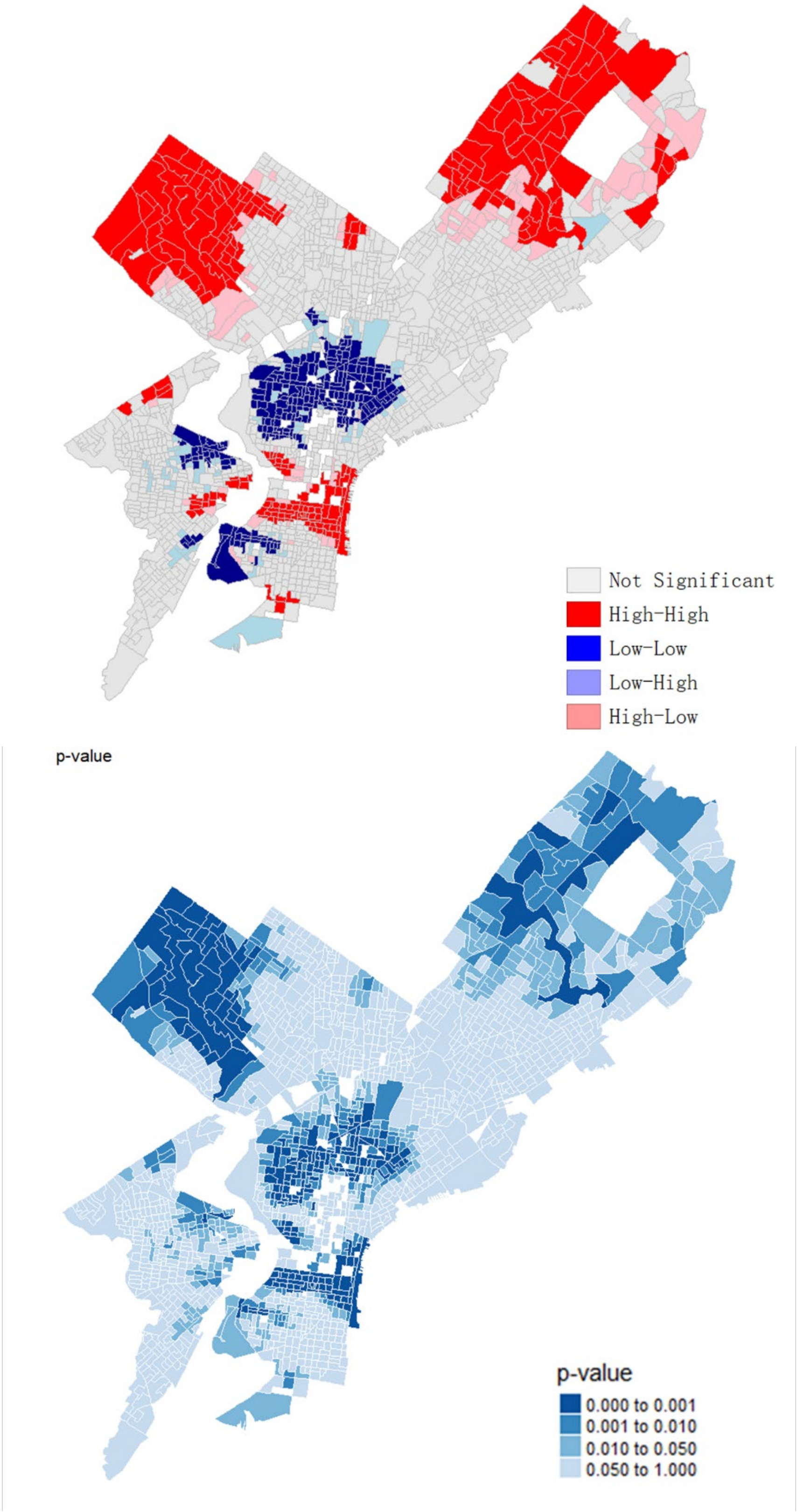


Figure 4: R Significance Map and Cluster Map



b) A Review of OLS Regression and Assumptions:

Table 1.1: GEODA OLS Regression Result

```
--------------------------------------------------------------------------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
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
--------------------------------------------------------------------------
```

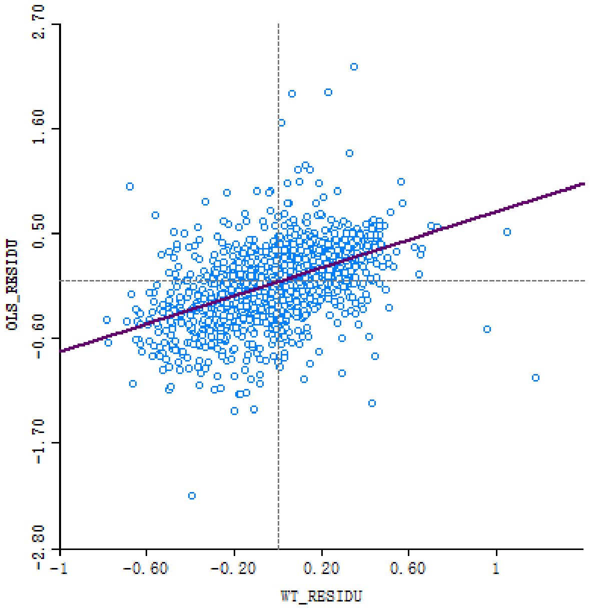
Table 1.2: R OLS Regression Result

```
--------------------------------------------------------------------------
lm(formula = LNMEDHVAL ~ LNNBELPOV + PCTVACANT + PCTBACHMOR +
PCTSINGLES, data = shp@data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.25817 -0.20391  0.03822  0.21743  2.24345

Coefficients:
--------------------------------------------------------------------------
              Estimate Std. Error t value             Pr(>|t|)
(Intercept) 11.1137781  0.0465318 238.843 < 0.0000000000000002 ***
LNNBELPOV   -0.0789035  0.0084567  -9.330 < 0.0000000000000002 ***
PCTVACANT   -0.0191563  0.0009779 -19.590 < 0.0000000000000002 ***
PCTBACHMOR   0.0209095  0.0005432  38.494 < 0.0000000000000002 ***
PCTSINGLES   0.0029770  0.0007032   4.234            0.0000242 ***
--------------------------------------------------------------------------
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3665 on 1715 degrees of freedom
Multiple R-squared:  0.6623,     Adjusted R-squared:  0.6615
F-statistic: 840.9 on 4 and 1715 DF,  p-value: < 0.00000000000000022
log Lik                                 -711.4933 (df=6)
studentized Breusch-Pagan test          p-value  0.00000001102
White's test result                     P-value: 0
Jarque Bera Test X-squared = 778.96, df = 2, p-value < 0.00000000000000022
--------------------------------------------------------------------------
```

We regressed the LNMEDHVAL on PCTSINGLES, PCTBECHMOR, PCTVACANT, and LNNBELPOV100. The p-value for all these four predictors is less than 0.05, meaning that they are all statistically significant. From the adjusted R-squared, we see that 66% of the variance in LNMEDHVAL has been explained by the model.

From the results of the Breusch-Pagan test, the studentized Breusch-Pagan test, and the Koenker-Bassett test, the p-values are all less than 0.05, so we can reject the null hypothesis for the alternate hypothesis of heteroscedasticity. The results from the 3 tests are consistent with each other, indicating a problem of heteroscedasticity. The normality of residuals is tested from the Jarque-Bera test, the p-value of it is close to zero, so we can reject the Null Hypothesis of normality for the alternative hypothesis of non-normality, which is also problematic.

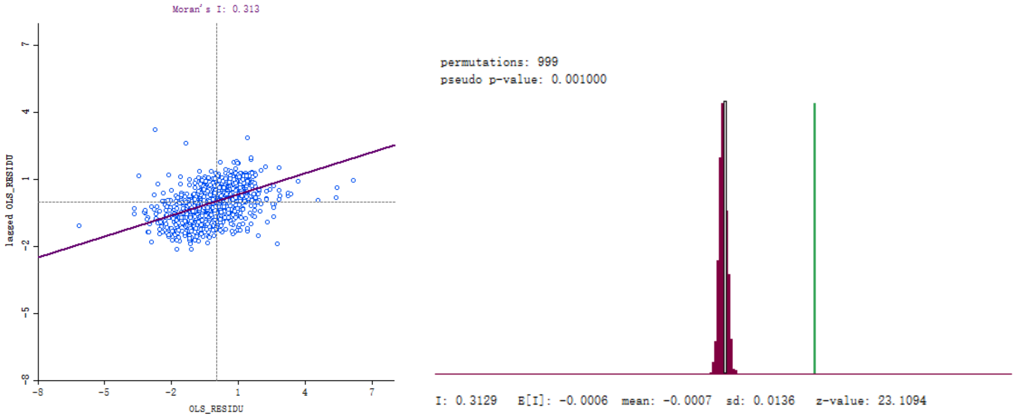
Figure 5: the scatterplot of **OLS_RESIDU** by **WT_RESIDU**



```
#obs  R^2   const a  std-err a  t-stat a  p-value a  slope b  std-err b  t-stat b  p-value b
1720  0.229  -0.005   0.008     -0.605     0.545     0.733     0.032     22.617      0
 0     0       0       0          0          0         0         0          0         0
1720  0.229  -0.005   0.008     -0.605     0.545     0.733     0.032     22.617      0

              Chow test for sel/unsel regression subsets: can't compute
```

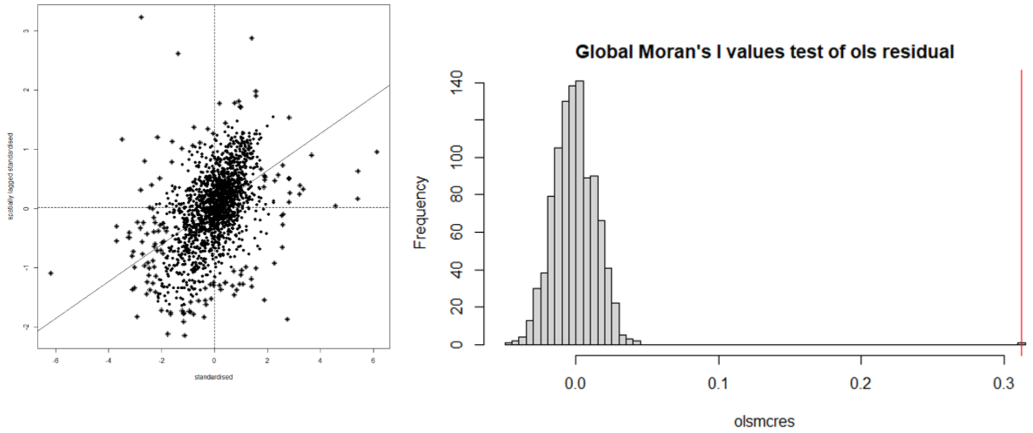
The Figure 5 shows the relationship between the OLS residuals and their weighted residuals. As the WT_RESIDUAL goes up, the OLS residual goes up. The best fit line has a slope of 0.733, and its p-value of it is close to 0, meaning that there is a significant spatial autocorrelation between the residuals and their neighbors.

Figure 6: Geoda Moran's I scatter plot & permutation plot



```
permutations: 999
pseudo p-value: 0.001000

I: 0.3129   E[I]: -0.0006   mean: -0.0007   sd: 0.0136   z-value: 23.1094
```

Figure 7: R Moran’s I scatter plot & permutation plot



The Moran’s I for the OLS residuals is 0.313, which is higher than 0.05, indicating a significant spatial autocorrelation. The pseudo-p-value from the results of the 999 permutations is 0.001, which is much lower than 0.05, meaning that the chance of observing a Moran’s I of 0.313 is very low if there is no spatial autocorrelation present. So, we can reject the null hypothesis that there is no spatial autocorrelation. This is problematic because it means our β coefficients and significance values in the regression may be wrong.

c) Spatial Lag Regression Results

Table 2.1: GEODA Spatial Lag Regression Results

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

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

---------------------------------------------------------------------------
     Variable       Coefficient      Std.Error       z-value    Probability
----------------------------------------------------------------------------
       W_LNMEDHVAL      0.651097      0.0180501         36.0716     0.00000
          CONSTANT       3.89846       0.201114         19.3843     0.00000
         LNNBELPOV    -0.0340547     0.00629287        -5.41163     0.00000
        PCTBACHMOR    0.00851381    0.000521935          16.312     0.00000
         PCTVACANT    -0.0085294    0.000743667        -11.4694     0.00000
        PCTSINGLES    0.00203342     0.00051577          3.9425     0.00008
---------------------------------------------------------------------------
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 : Qweight
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1       911.5067     0.00000
```

```
--------------------------------------------------------------------------
```

Table 2.2: R Spatial Lag Regression Results

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

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

Type: lag 
Coefficients: (asymptotic standard errors) 
--------------------------------------------------------------------------
               Estimate  Std. Error  z value              Pr(>|z|)    
(Intercept)  3.89845505  0.20111357  19.3843 < 0.00000000000000022 
LNNBELPOV   -0.03405466  0.00629287  -5.4116         0.00000006246 
PCTVACANT   -0.00852940  0.00074367 -11.4694 < 0.00000000000000022 
PCTBACHMOR   0.00851381  0.00052193  16.3120 < 0.00000000000000022 
PCTSINGLES   0.00203342  0.00051577   3.9425         0.00008063502 
--------------------------------------------------------------------------

Log-likelihood: -255.74 for lag model
AIC: 525.48, (AIC for lm: 1435)
Likelihood ratio = 911.51, df = 1, p-value < 0.00000000000000022
Breusch-Pagan test     BP = 210.76, df = 4, p-value < 0.00000000000000022
studentized Breusch-Pagan test  BP = 51.411, df = 4, p-value = 0.0000000001832
Jarque Bera Test  X-squared = 2756.9, df = 2, p-value < 0.00000000000000022

--------------------------------------------------------------------------
```

We did the spatial lag regression for the LNMEDHVAL and all other four predictors. The p-value for the W_LNMEDHVAL is close to 0, far less than 0.05, indicating the variable W_LNMEDHVAL is significant and the median house value in one area is associated with that of the surrounding areas.

For the remaining four predictors, their p-values of them are all less than 0.05. Comparing the OLS model and the spatial lag model, the predictors have p-values less than 0.05 in both models. The standard error of the four variables in the spatial lag model is higher than that of the OLS model, meaning that the samples from the spatial lag model distribute more dispersed. Considering the p-value of the Breusch-Pagan test is less than 0.05, so our spatial lag regression residuals are still heteroscedastic.

We use the AIC/SC, the Log-Likelihood, and the Likelihood Ratio Test to compare the two models. The AIC and the SC results of the OLS model are much higher than that of the spatial lag model, indicating that the spatial lag model is a better fit than OLS. The Log-Likelihood value of the spatial lag model is higher than that of the OLS model, again indicating the spatial lag is a better one. In the Likelihood Ratio test, the p-value is less than 0.05, so we can reject the null hypothesis and state the spatial lag model is better than the OLS model.

Figure 8: Geoda Moran’s I scatter plot & permutation plot

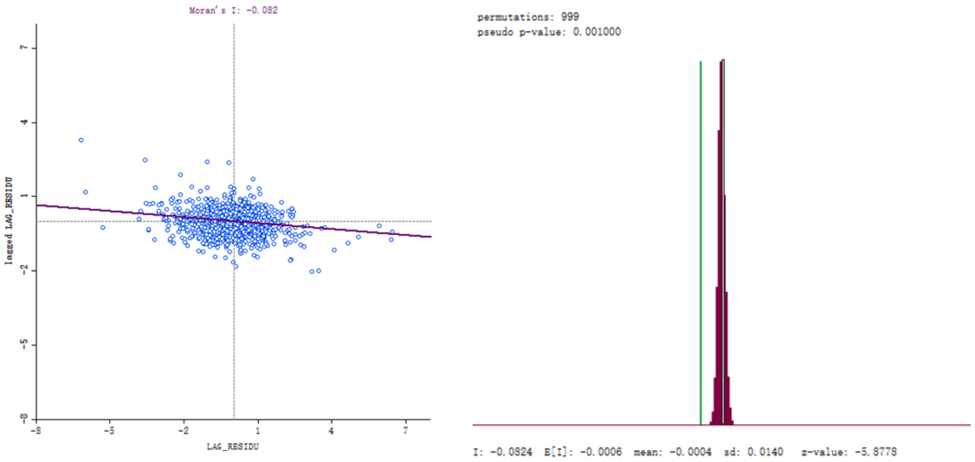
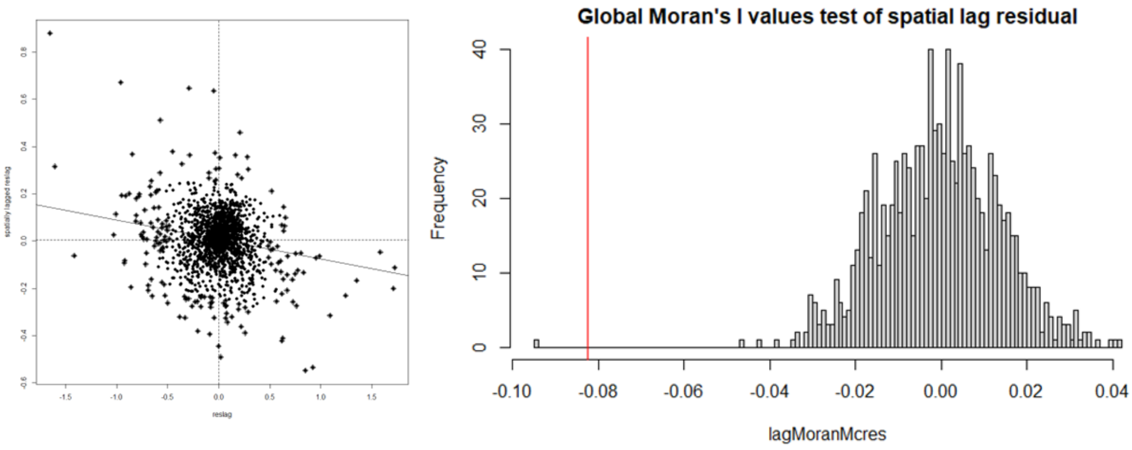


Figure 9: R Moran’s I scatter plot & permutation plot



The Moran’s I for the spatial lag model is -0.082, closer to zero than that of the OLS model, suggesting that there seems to be less spatial autocorrelation in these residuals than in OLS residuals. The pseudo-p-value of the 999 random permutation test for both the spatial lag model and the OLS model is 0.001, indicating that there are spatial autocorrelations in both models. However, according to the Moran’s I scatter plot and the residual histogram, both are much closer to the ideal Moran’s I for random distribution, meaning that there’s less autocorrelation in the spatial lag residuals than in OLS residuals.

In general, according to all the above analyses, the spatial lag model does better than the OLS model in explaining the relationship between the dependent variable and the predictors.

d) Spatial Error Regression Results

Table 3.1 : GEODA Spatial Error Regression Results

```
---------------------------------------------------------------------------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set             : Regression Data
Spatial Weight       : Qweight
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
        PCTSINGLES    0.00267792    0.000620832        4.31343     0.00002
         PCTVACANT   -0.00578308    0.000886701       -6.52201     0.00000
         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 : Qweight
TEST                                DF      VALUE         PROB
Likelihood Ratio Test               1       677.6059      0.00000
---------------------------------------------------------------------------
```

Table 3.2 ***:*** R Spatial Error Regression Results

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

Residuals:
      Min        1Q    Median        3Q       Max
-1.926477 -0.115408  0.014889  0.133852  1.948663

Type: error
Coefficients: (asymptotic standard errors)
---------------------------------------------------------------------------
               Estimate  Std. Error  z value              Pr(>|z|)
(Intercept) 10.90643419  0.05346781 203.9813 < 0.00000000000000022
LNNBELPOV   -0.03453407  0.00708933  -4.8713      0.00000110882576
PCTVACANT   -0.00578308  0.00088670  -6.5220      0.00000000006937
PCTBACHMOR   0.00981293  0.00072896  13.4615 < 0.00000000000000022
PCTSINGLES   0.00267792  0.00062083   4.3134      0.00001607387769
---------------------------------------------------------------------------
Lambda: 0.81492, LR test value: 677.61, p-value: < 0.000000000000000222

AIC: 759.38, (AIC for lm: 1435)
Likelihood ratio = 677.61, df = 1, p-value < 0.00000000000000022
Breusch-Pagan test  BP = 23.213, df = 4, p-value = 0.0001148
studentized Breusch-Pagan test  df = 4, p-value = 0.271
Jarque Bera Test X-squared = 3507, df = 2, p-value < 0.00000000000000022
---------------------------------------------------------------------------
```

We did the spatial error regression for the LNMEDHVAL and all other four predictors. The LAMDA term is around 0.81, which is close to 1, and the p-value for it is close to 0, far less than 0.05, meaning that the correlation between the OLS residuals and their neighbor residuals is significant. Comparing the spatial error results with the OLS results, we can find the remaining terms in the model are significant since the p-value for the spatial error model are all less than 0.05. Based on the Breusch-Pagan test which has a p-value less than 0.05, we can reject the Null Hypothesis that the residuals are homoscedastic for the alternative hypothesis of heteroscedastic, which is problematic.

We use the AIC/SC, the Log Likelihood, and the Likelihood Ratio Test to compare the two models. The AIC and the SC results of the OLS model are 1432.99/1460.24, much higher than that of the spatial error model 755.381/782.631, indicating that the spatial error model fits better than the OLS. The Log Likelihood value of the spatial error model(-372.69) is higher than that of the OLS model(-711.49), again indicating the spatial error model is a better one. In Likelihood Ratio test, the p-value is less than 0.05, so we can reject the null hypothesis of the spatial error model is not a better specification than the OLS model.

Figure 10: Geoda Moran's I scatter plot & permutation plot

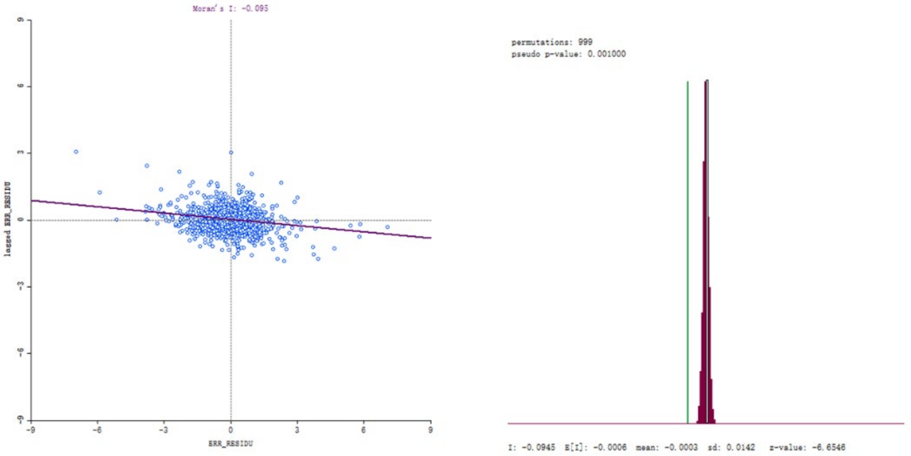
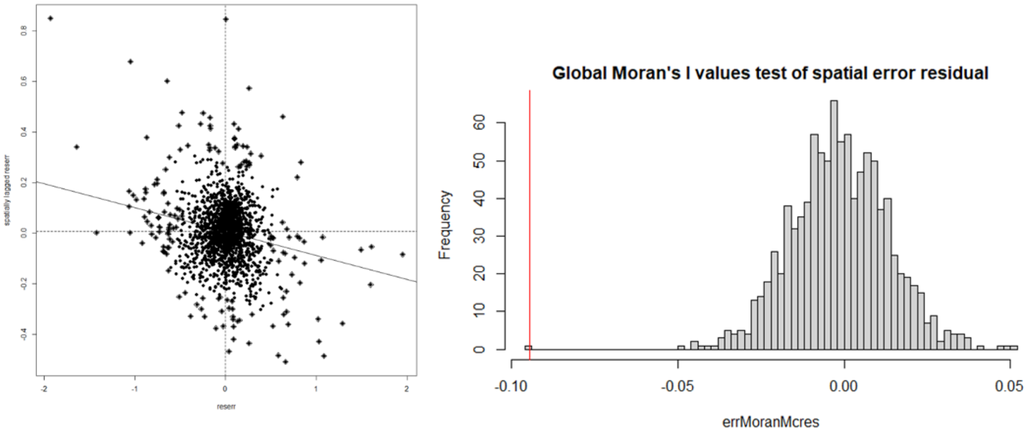


Figure 11: R Moran's I scatter plot & permutation plot



The Moran's I for the spatial error model is -0.095, closer to zero than that of the OLS model, suggesting that there seems to be less spatial autocorrelation in these residuals than in OLS residuals. The pseudo p-value of the 999 random permutation test for the spatial error model is 0.001, indicating that the residuals are still spatially autocorrelated. However, both Moran's I and the histogram are closer to the random distribution of Moran's I, so we can draw the conclusion that there's less spatial autocorrelation in the spatial error residuals compared with the OLS residuals.

In conclusion, according to all the above analysis, the spatial error model does better than the OLS model, since it has a better fit of the data and can explain the LNMEDHVAL better.

When comparing the spatial lag model and the spatial error model, we cannot use the log likelihood, the likelihood ratio test, and the R-squared for comparison. However, it's ok to compare them based on the AIC/SC value. The AIC/SC for the spatial lag model is 523.48/556.18, lower than the AIC/SC for the spatial error model, which is 755.381/782.631. So, the spatial lag model does better than the spatial error model.

e) Geographically Weighted Regression Results

Table 3.1 *:* Geographically Weighted Regression Results

```
--------------------------------------------------------------------------
Call:
gwr(formula = LNMEDHVAL ~ LNNBELPOV + PCTVACANT + PCTBACHMOR +
    PCTSINGLES, data = shp, gweight = gwr.Gauss, adapt = bw,
    hatmatrix = TRUE, se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.008130619 (about 13 of 1720 data points)

Summary of GWR coefficient estimates at data points:
--------------------------------------------------------------------------
                  Min.    1st Qu.     Median    3rd Qu.       Max.  Global
X.Intercept.  9.6727618 10.7143173 10.9542384 11.1742009 12.0831381 11.1138
LNNBELPOV    -0.2365244 -0.0733572 -0.0401186 -0.0126657  0.0948768 -0.0789
PCTVACANT    -0.0317407 -0.0142383 -0.0089599 -0.0035770  0.0167916 -0.0192
PCTBACHMOR    0.0010974  0.0101380  0.0149279  0.0202187  0.0347258  0.0209
PCTSINGLES   -0.0249706 -0.0075550 -0.0016626  0.0042280  0.0143340  0.0030
--------------------------------------------------------------------------
Number of data points: 1720
Effective number of parameters (residual: 2traceS - traceS'S): 360.5225
Effective degrees of freedom (residual: 2traceS - traceS'S): 1359.477
Sigma (residual: 2traceS - traceS'S): 0.2762201
Effective number of parameters (model: traceS): 257.9061
Effective degrees of freedom (model: traceS): 1462.094
Sigma (model: traceS): 0.2663506
Sigma (ML): 0.245571
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 660.7924
AIC (GWR p. 96, eq. 4.22): 308.7123
Residual sum of squares: 103.7248
Quasi-global R2: 0.8479244
--------------------------------------------------------------------------
```

The GWR R-squared is 0.85, higher than the OLS R-squared, which is 0.66. This means that the GWR is doing a better job since it explains 20% more variance in the dependent variable. The AIC value of GWR is 582.1, lower than that of the OLS regression model and the spatial error model, which are 1432.9 and 755.3, and higher than the spatial lag model, which is 523.5. In this case, the GWR model fits better than the OLS model and the spatial error model, but no as well as the spatial lag model.

Figure 12: Geoda Moran's I scatter plot & permutation plot

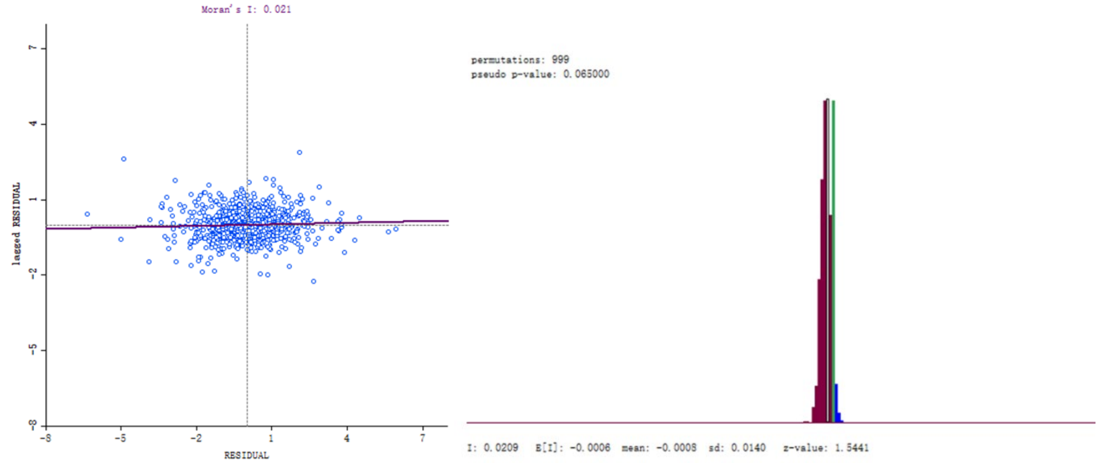
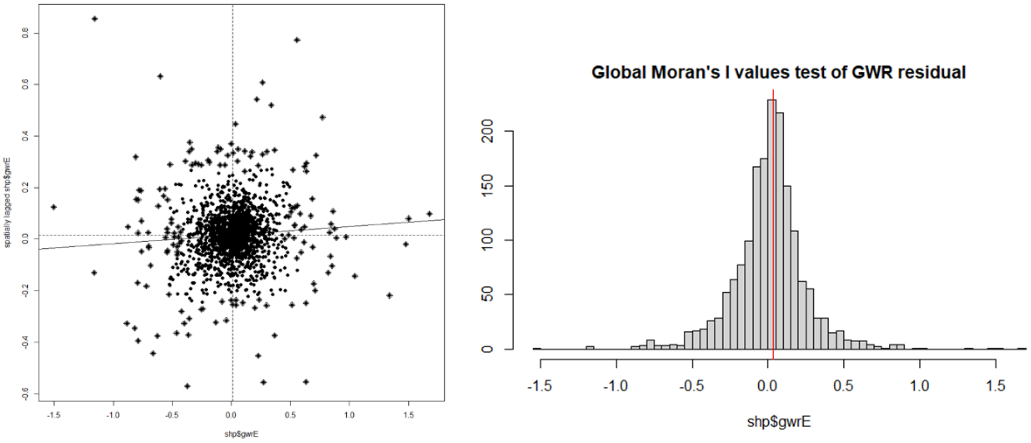
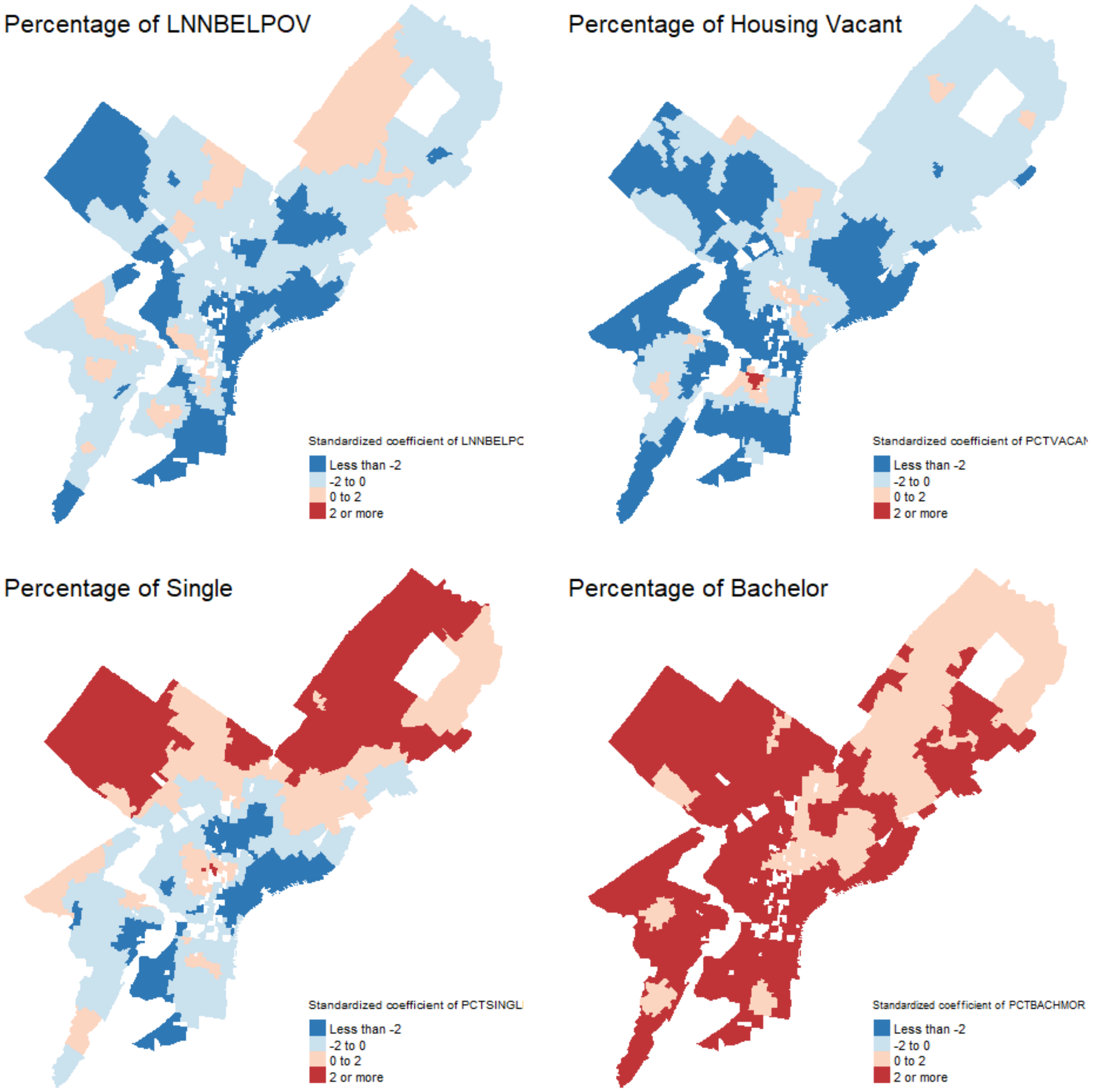


Figure 13: R Moran's I scatter plot & permutation plot



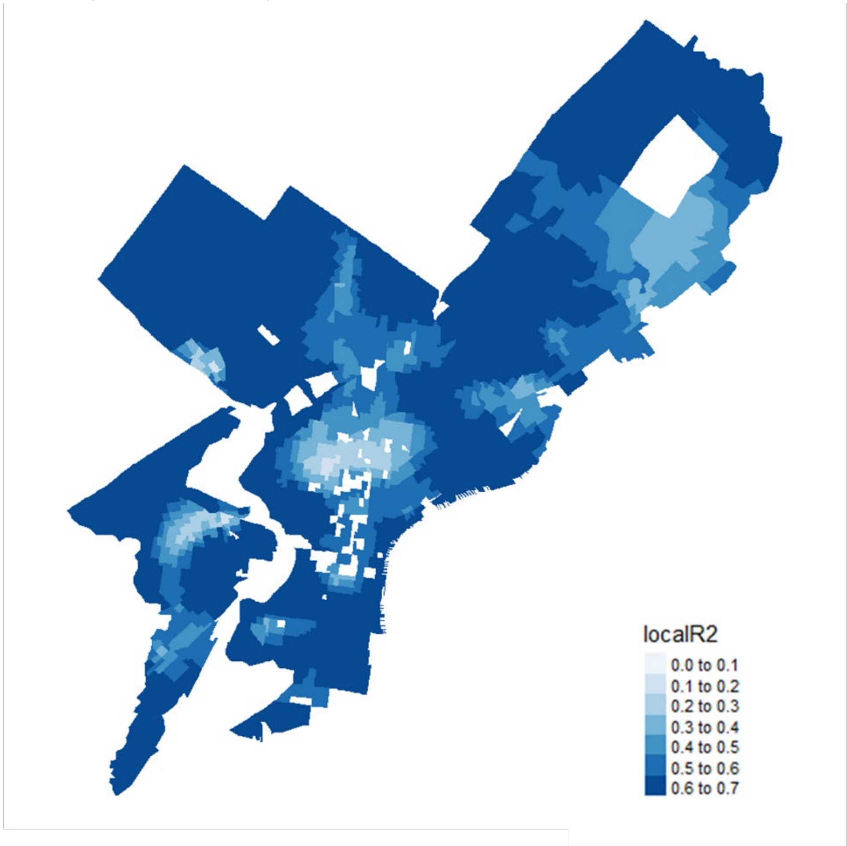
Moran's I for the GWR model is 0.021, closer to zero than that of the OLS model, suggesting that there seems to be less spatial autocorrelation in these residuals than in OLS residuals. The Moran's I for the spatial lag model is -0.082, and for the spatial error, model is -0.095. Comparing all these three models, Moran's I for the GWR model is the closest one to 0, therefore, the residuals in the GWR model have the least spatial autocorrelation. By looking at the residual histogram, the pseudo p-value of the 999 permutations is 0.065, which is larger than 0.05, meaning that there's no longer spatial autocorrelation in the residuals in this model.

Figure 14: Choropleth map of standardized coefficient results



The four maps in Figure 14 shows the local regression results for the four predictors. In general, the percentage of bachelor or more shows more positive relations with the median house value, and the percent of vacant and the number of households living in poverty have generally negative relationships with the median house value. For the result of the regression between the percent of singles and the median house value, the variable has a generally positive effect on the dependent variable in the center, the northwest, and northeast Philadelphia, while in east and middle north Philly, it has a negative effect on the median house value.

Figure 15: Choropleth map of local R-squared results



The Figure 15 shows the local R-squared results' spatial distribution. The lowest R-squared distribution in the center-north and part of southwest Philadelphia, ranging from 0-0.2, meaning that in these places, about only 0%-20% of the variance in the median house values are explained by the predictors, is relatively poor. The highest R-squared appears in the northwest, northeast, and part of south Philadelphia, ranging from 0.6-0.7, which means about 60%-70% of the dependent variable can be explained by the four predictors, indicating the predictors and the dependent variable have more correlation in these areas.

4. Discussion

In this project, we conducted the OLS regression to predict the median house values in Philadelphia. However, the results of the OLS model show the residuals have significant spatial autocorrelation with each other and that means there might be problems of under-or-over-prediction in certain areas. So, we also use the spatial lag model, the spatial error model, and the GWR model to help improve the results, which can explicit spatial interaction in residuals and can address the possible influence of spatial autocorrelation on simple linear regression models that assume spatial independence. After comparing all these four models, the GWR model has the lowest AIC value, which means it has the best fitness in predicting the median house value. In addition, all three models do better than the original OLS model.

Though our three models can provide more explanations for the prediction, there are still some limitations. For both spatial lag and spatial error models, the pseudo-p-values are all 0.001, lower than 0.05, meaning that there’s still spatial autocorrelation in the residuals in each model. The results of the Breusch-Pagan test for both the spatial lag and the spatial error model show the p-value are all less than 0.05, so we can reject the null hypothesis that the residuals are homoscedastic for the alternative hypothesis of heteroscedastic, which is problematic. In conclusion, although the spatial lag and spatial error model does better than the OLS model, they still cannot fully explain the spatial autocorrelation in residuals.

Reference

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