Prediction of Median House Values in Philadelphia

1. Introduction

In the previous assignment, we applied ordinary least squares (OLS) analysis to examine how socioeconomic independent variables of census block groups could influence their median house values. Such factors include: 1) number of households living in poverty, 2) the percentage of individuals with Bachelor's Degrees or Higher, 3) the percentage of vacant houses and 4) the percentage of single house units in tracts as our predictors. However, the OLS Regression is not an ideal analysis model for analysis that contains spatial components as spatial datasets themself could present correlation between each other based on proximities that violates the OLS assumption of independence of observations and residuals (Scott, 2009).

In this project, we are going to give predictions of median house values in Philadelphia Census Block Group based on our investigation of spatial regressions models using R, GeoDa and ArcGIS. Regression models include spatial lag, spatial error, and geographically weighted regression. We are particularly interested to see whether those spatial regression models could have better performances than OLS models.

2. Methods

2.1 Spatial Autocorrelation

Spatial objects have one unique characteristic that differs it from the normal quantitative dataset, which is encapsulated as "everything is related to everything else, but near things are more related than distant things". Such a summary is called *Tobler's First Law of Geography* (Waters, 2016). Thus, it is important to examine the relationship between adjacent spatial objects when conducting the spatial analysis. Moran's I is a common measure of spatial autocorrelation in statistics, with its formula presented below. The results of Moran's I range from 1 to -1, with large positive values as strong positive autocorrelation, large negative values as strong negative autocorrelation, while 0 means no spatial autocorrelation. The equation of Moran's I is as follows:

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

Where w_{ij} is the spatial weight between I and j, X_i and X_j is the variable value at a particular location i or j, $\bar{X}$ is the mean of variable X and n is the number of observations.

The calculation of Moran's I requires the weight matrix that presents the relationship between all of the objects in the dataset. In this study, we applied the Queen Weight Matrix - neighboring objects are defined by having intersections with the original object in both points and segments. However, most statisticians tend to use more than one weight matrix in their spatial autocorrelation studies to increase the accuracy.

To test its significance, we then conducted a process of permutation. It randomly reshuffles all spatial features of our dataset around to create a new spatial pattern, and calculate the Moran's I of each object in that new pattern. The newly permuted Moran's I value will apparently be different because it will have new Queen neighbors every time after reshuffling. By repeating the process and re-calculation 999 times, we have a total of 1000 data, composed of 1 actual observation data and 999 permuted data with varying Moran's I values.

We state the null-hypothesis that there is no spatial autocorrelation of the median house values, while the alternative being there is spatial autocorrelation. To reject that, we need to demonstrate that our original data had a significant P-value among the randomly generated datasets. Among the 1000 datasets, we calculate the P-value by sorting out the rank of the original observation's Moran's I by descending order, then dividing by the sample size of 1000. When the p-value is less than 0.05, which means as long as the original Moran's I rank above the 50, we can reject the null hypothesis and favor that the original Moran's I have less than 5% of being non-significant, meaning there is significant spatial-autocorrelation.

With the given Global Moran's I, we also calculated the Local Moran's I (LISA) for each of the spatial objects using the Queen Matrix in our data. Compared to the Global Moran's I, LISA more focuses on the relationship of each object to its spatially surrounding objects. At location i, Local LISA statistics may be positive or negative. We want to test whether there is statistically significant spatial autocorrelation at location i, and we state that the null hypothesis is that there is no local spatial autocorrelation at location i, while the alternative hypothesis is that there is negative or positive spatial autocorrelation at location i.

2.2 OLS Regression and Assumptions

Ordinary Least Square (OLS) regression is a statistical method used for predicting a variable (predictor) based on the observed pattern of the dependent variables. Certain assumptions should be satisfied before conducting OLS regression, such as the linearity, normality of residuals, homoscedasticity, non-multicollinearity of the original datasets, and most importantly, the independence of observations, which means one observation occurrence should provide no information about the other observation occurrence. ("NEDARC - Statistical Terms Dictionary" 2021)

However, with the requirement of independence of observations, the OLS regression model is inappropriate for analyzing spatial data because of its contradiction with the First Law of Geography mentioned above. In this case, we can use Moran's I to test the spatial autocorrelation of the residuals, expressed as ρ (in GeoDa, this is known as λ). We can also regress OLS residuals on datasets' Queen neighbors. To test OLS residuals for spatial autocorrelation is to regress them on their nearby residuals, where we defined neighboring block group by the Queen Matrix. ρ (rho), a term known as *lambda* (λ) in GeoDa, is used to measure the spatial autocorrelation that's the beta coefficient when OLS residuals are regressed on their (queen) neighbors.

GeoDa allows the test of a few other assumptions like homoscedasticity and the normality of errors. Breusch-Pagan Test and Koenker-Bassett test are used by GeoDa for homoscedasticity. We stated the null hypothesis that there is homoscedasticity among our spatial data, by rejecting it if $p<0.05$, we can favor of the alternative hypothesis of the heteroscedasticity of our data. Normality of Residuals, or the Normality of Errors, could be tested using the Jarque-Bera Test in GeoDa: we listed the null hypothesis that the residuals of data are from a normal distribution. By rejecting the null hypothesis when $p<0.05$, we reject H0 of normality for the Ha of non-normality. (MIT Library, 2013)

2.3 Spatial Lag and Spatial Error Regression

The spatial lag regression assumes that the value of one dependent variable is associated with the values of that variable in the nearby locations, and the 'nearby' is defined according to the weights matrix W, such as rook and queen. The equation of spatial lag regression is as follows:

$$y = \rho W_y+\beta_0+ \beta_1 X_1+ \beta_2 X_2+ \beta_3 X_3+ \beta_4 X_4+\varepsilon \tag{2}$$

Where:

- y denotes the dependent variable LNMEDHVAL;
- ρ is the coefficient of the y-lag variable Wy;
- β_0 denotes the intercept, a constant, meaning the expected value of y when all x=0;
- ε denotes residuals, suggesting the difference between the observed and predicted values;
- β_i is the called the standard coefficient of predictor Xi, which is a k-length vector of parameters to be estimated;
- X_1 denotes LNNBELPOV;
- X_2 denotes PCTBACHMOR;
- X_3 denotes PCTSINGLES;
- X_4 denotes PCTVANT.

The spatial error regression assumes that the value of residuals is associated with its nearby residuals. It is a two-step regression: first, OLS regression is conducted, where Y is regressed on its predictors; second, residuals are regressed on the nearest neighbor residuals. The equation of spatial lag regression is as follows:

$$y = \beta_0+ \beta_1 X_1+ \beta_2 X_2+ \beta_3 X_3+ \beta_4 X_4+\lambda W\varepsilon+u \tag{3}$$

Where:

- y denotes the dependent variable LNMEDHVAL;
- β_0 denotes the intercept, a constant, meaning the expected value of y when all x=0;
- β_i is the called the standard coefficient of predictor Xi, whichis a k-length vector of parameters to be estimated;
- The OLS residuals ε is split into two parts: one with the spatially lagged residuals, denoted as λW_ε, and the other is the random noise, denoted as u.

In GeoDa and R, we run the spatial lag and spatial error regressions. The goal of both spatial lag and spatial error regression models is to consider the fact that the spatial dependencies might exist in the data or residuals. Both models might result in less heteroscedastic or no spatially correlated residuals. Then the residuals can be examined for spatial autocorrelation (Moran's I). For both regression models, we assume that each predictor is linearly related with the DV, the residuals should be normal, and no multicollinearity exists among predictors.

Both spatial lag regression and spatial error regression are conducted in our study, and the results of two models with OLS are compared to decide which spatial model perform better, based on the following three criteria:

(1) Akaike Information Criterion/Schwarz Criterion
Both Akaike Information Criterion (AIC) and Schwarz Criterion (SC) are measures of the goodness of fit for an estimated model. They measure the lost information relatively when a given modelis used to describe the reality. The smaller value of AIC/SC is, the better fit of regression model is.

(2) Log likelihood
The log likelihood is associated with the maximum likelihood of the fitness of a statistical model. Maximum likelihood is used for estimating the parameters of given distribution, by picking up the values of model parameters that make the data more possible than other values of parameters that would make them. The higher log likelihood is, the better fit of regression model is. It should be used for comparing nested models only.

(3) Likelihood Ratio Test
The likelihood ratio test is a test of hypothesis where OLS model is compared with the spatial model to decide whether to reject the null hypothesis or not. Both null and alternative hypotheses are proposed as the following:
H_0 (Null Hypothesis): spatial lag (error) model is not a better model than the OLS model.
H_A (Alternative Hypothesis): spatial lag (error) model is not a better model than the OLS model.
If $p<0.05$, we reject the null hypothesis, and assume that the spatial lag (error) model performs better than the OLS module.

In addition to these criteria, we can also look at the Moran’s I of regression residuals to see which module performs better. First, run OLS regression and examine whether there is significant spatial autocorrelation in residuals. If so, then examine the spatial autocorrelation of the residuals of spatial lag and spatial error models. The module with lower spatial autocorrelation in residuals performs better.

In our study, we will compare the models based on the Akaike Information Criterion/Schwarz Criterion, the Log Likelihood, and the Likelihood Ratio Test. We also looked at Moran’s I scatterplot of regression residuals. A better module should have smaller AIC/SC, higher log likelihood, less significant spatial autocorrelation in residuals and null hypothesis of likelihood ratio test should be rejected.

2.4 Geographically Weighted Regression

Geographically Weighted Regression (GWR), one spatial regression technique in geography discipline, is a local linear regression form that can model spatially variation relationships ("How GWR Works—Arcmap | Documentation" 2021) and GWR analyses will be conducted in ArcGIS and R. Separate, local regression models for each location are built in GWR model, which makes up for the lack of accuracy result generated from the single, global regression. Simpson's Paradox refers to the reversal results when groups of data are analyzed separately and then combined, which stresses the risk of analyzing aggregate data (Lee and Schuett 2014). According to Simpson's Paradox, the estimation of relationships on the global scale can be very misleading. GWR, on the other hand, can improve the performance and accuracy by examining the local spatial variation, by using other observations in the dataset to run the regression.

To run the local regression, multiple observations rather than a single one should be included in the GWR model. These observations are weighted according to their proximity to location i: the closer observations are to location i, the greater weights they are given and the stronger influence on the estimation of parameters for location i is. A spatial kernel is used to provide the geographic weighting of the model and its size is controlled by bandwidth, a key coefficient in the kernel (2021). Two ways to weigh nearby locations are fixed bandwidth and adaptive bandwidth.

Fixed bandwidth means that the bandwidth distance remains fixed despite the number of observations will vary around each point i. It is more appropriate to apply the fixed bandwidth kernel when distributions vary across space, such as size and number of neighborhoods. On the other hand, adaptive bandwidth function assumes that the area and bandwidth change even if the number of observations remain constant. This bandwidth is appropriate when distribution varies across the space. In this study, significant clustering patterns as well as spatial heterogeneity of variables is detected. Under this circumstance, adaptive bandwidth is used in our GWR analysis.
OLS assumptions still hold for GWR, including linearity, independence of observations, normality of residuals, homoscedasticity and no multicollinearity, otherwise the results will be unreliable, which usually happens in global regression models. When some variables have similar clustering patterns in a certain region, multicollinearity exists. In the attribute table, the conditional number suggests that results are unstable due to local multicollinearity. Caution should be used to where categorical data cluster spatially, indicating the presence of local collinearity ("Arcgis Desktop Help 9.3 - Geographically Weighted Regression (Spatial Statistics)" 2021).

Unlike the global model OLS, which can be accomplished with t-test and their associated P-values are provided, each regression point in GWR is associated with one set of parameters as well as standard errors, therefore, to decide the local significance of parameters, hundreds or thousands of tests should be required. In this 4 predictor model estimated at about 1720 regression points, there would be 8600 (5*1720) significant tests: 4 for predictors and I for the intercept. According to type 1 error, where we would expect 5 tests in every 100 to be significant results but in reality they do not if we used a =0.05 as significance level, we would expect 430 (8600*0.05) of these tests to be significant by chance. Due to this problem with multiple testing, p-values are not part of GWR output.

3. Results

3.1 Spatial Autocorrelation

Figure 1. shows Moran’s scatterplot of the LNMEDHVAL in relation to the average value of its Queen neighbors, with the global Moran’s I value of 0.794, which means there is possibly positive spatial autocorrelation of the LNMEDHVAL. Figure 2 is the permutation result of the global Moran’s I, showing the p-value of 0.001, in which the original Moran’s I rank the top one value among the Moran’s I of the reshuffled neighbor patterns. Based on that, we can reject the null hypothesis and argue that there is significant spatial-autocorrelation.

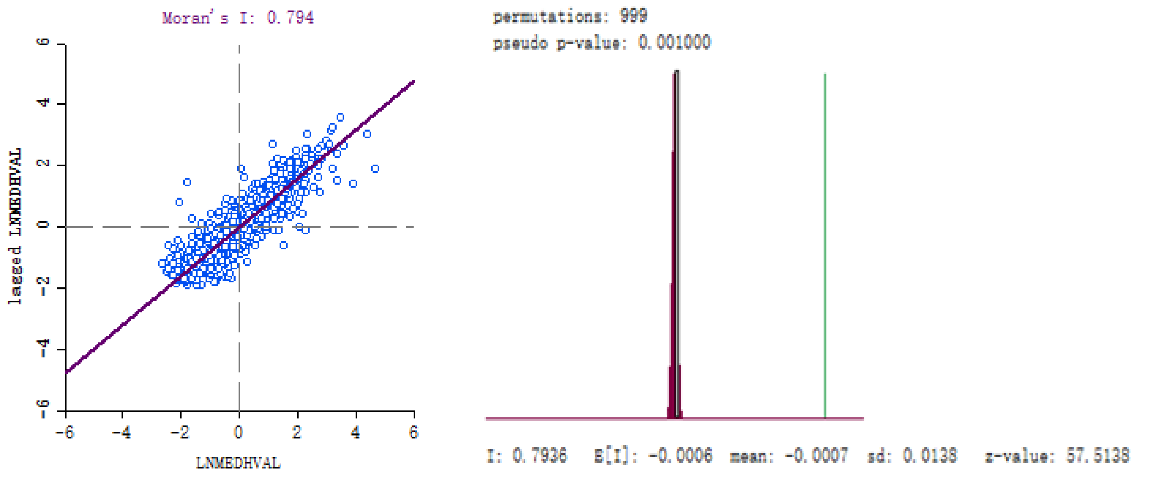


Figure 1 & 2. Moran’s I and permutation scatterplot of LNMEDHVAL

Figure 3 applies the local Moran’s I. We calculated into the map of Philadelphia, PA. By looking at the map, we can observe that there are high-high spatial autocorrelations in LNMEDHVAL (those areas with high LNMEDHVAL are surrounded by neighborhoods that also have high LNMEDHVAL) blocks clustered in the Northeast Suburbs, Northwest Suburbs, Central City District, and the Southern University City, plus a small block in the southern City; Blocks with low-low LNMEDHVAL relations (neighborhoods with low LNMEDHVALsurrounded by neighbors that also has low LNMEDHVAL) are clustered in areas north and southeast to the Central City, north and south to the University City, with an isolated small cluster in the northeast, likely to be the Kensington. There aren’t large clusters of high-low and low-high blocks, they are observed in the Southeast Downtown and the area adjacent to Northeast suburbs.

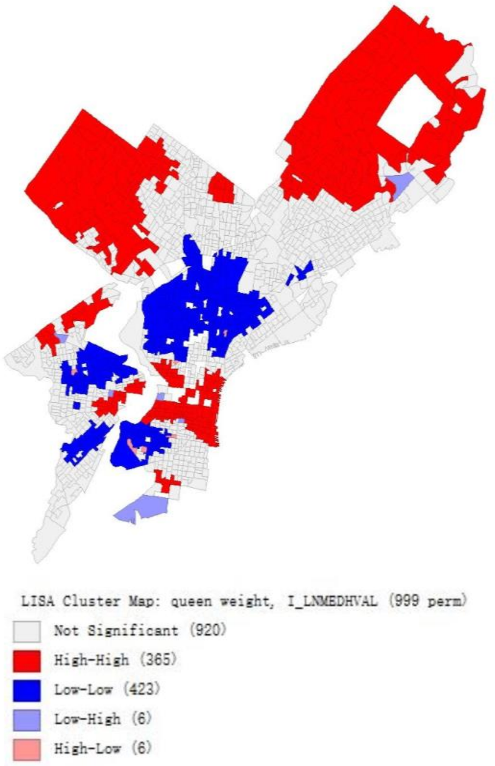


Figure 3. LISA Cluster Map

Figure 4 shows the significance value of the LISA. As the map shows, most of the block groups that are being identified “correlated” on Figure 3 have a p-value less than 0.05, which means we can reject the null hypothesis for the alternative hypothesis that the local Moran’s I has less than 5% of being non-significant, meaning there is significant spatial-autocorrelation.

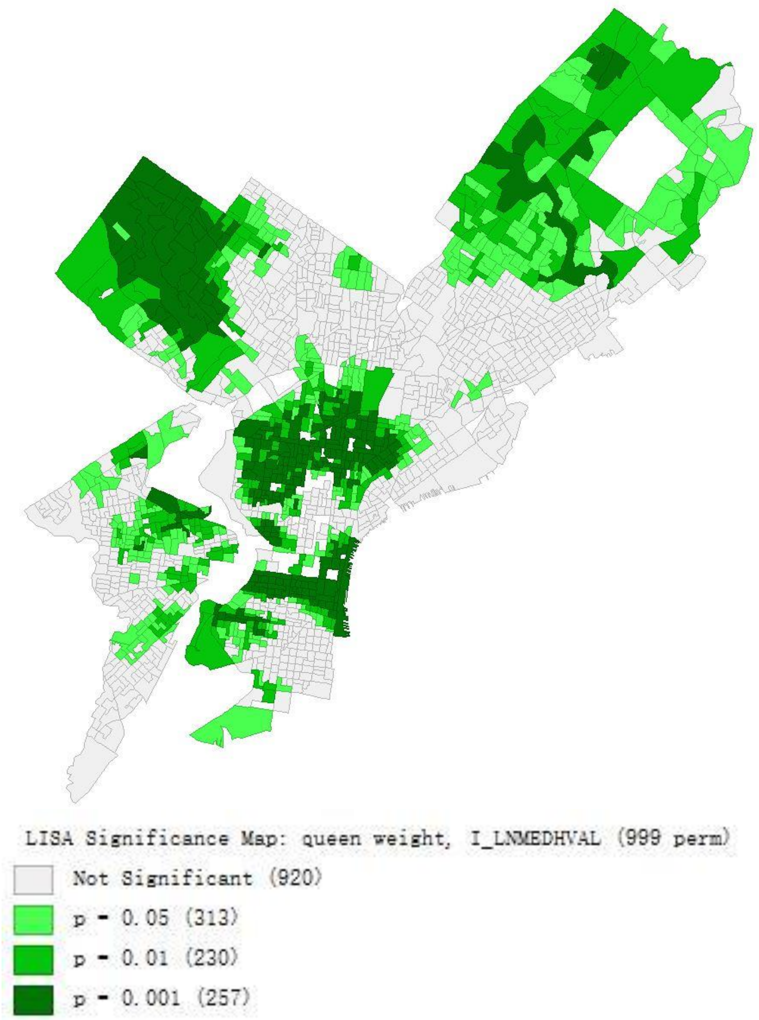


Figure 4. LISA Significance Map

3.2 OLS Regression and Assumptions

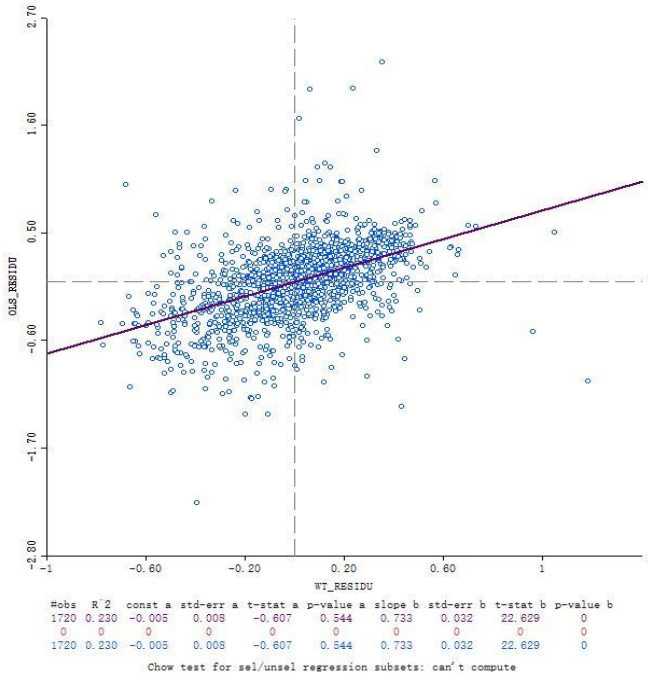
In Table 1, we measured coefficients, standard errors, t-statistic, and also the p-value for each predictor in relation to LNMEDHVAL. We noticed that after conducting a T-test based on 1720 df for each predictor, all predictors are significant (p<0.05). The residual standard error is 0.3665, and R Square equals 66.23%, which is the coefficient of multiple determination, as well as the proportion of variance in the model explained by all 4 predictors. In addition, the adjusted R Square equals 66.15%, which can be explained by predictors adjusted for the number of 4 predictors. We observed that all p-values associated with an F-ratio of 840.9 are less than 0.0001.

Table 1 also shows that the p-value for Breusch-Pagan test and Koenker-Bassett test is almost 0, less than 0.05, meaning we can reject the null-hypothesis of heteroscedasticity and favor the alternative of homoscedasticity; the Jarque-Bera test has the p-value of 0, meaning we can favor the alternative hypothesis of non-normality.

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set          :  Regression Data
Dependent Variable  :   LNMEDHVAL  Number of Observations: 1720
Mean dependent var  :      10.882  Number of Variables   :    5
S.D. dependent var  :     0.62972  Degrees of Freedom    : 1715

R-squared           :    0.662300  F-statistic           :    840.869
Adjusted R-squared  :    0.661513  Prob(F-statistic)     :          0
Sum squared residual:     230.332  Log likelihood        :   -711.493
Sigma-square        :    0.134304  Akaike info criterion :    1432.99
S.E. of regression  :    0.366475  Schwarz criterion     :    1460.24
Sigma-square ML     :    0.133914
S.E of regression ML:    0.365942

-----------------------------------------------------------------------------
Variable     Coefficient     Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
CONSTANT       11.1138      0.0465318       238.843     0.00000
LNNBELPOV    -0.0789035     0.0084567       -9.3303     0.00000
PCTBACHMOR    0.0209095    0.000543184      38.4944     0.00000
PCTSINGLES   0.00297695    0.000703155      4.23371     0.00002
PCTVACANT    -0.0191563    0.000977851     -19.5902     0.00000
-----------------------------------------------------------------------------
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   12.990609
TEST ON NORMALITY OF ERRORS
TEST                DF          VALUE           PROB
Jarque-Bera          2         778.9646        0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                DF          VALUE           PROB
Breusch-Pagan test    4        162.9108        0.00000
Koenker-Bassett test   4        61.6992        0.00000
SPECIFICATION ROBUST TEST
TEST                DF          VALUE           PROB
White                14        111.3224        0.00000
```

Table 1. OLS Regression Output

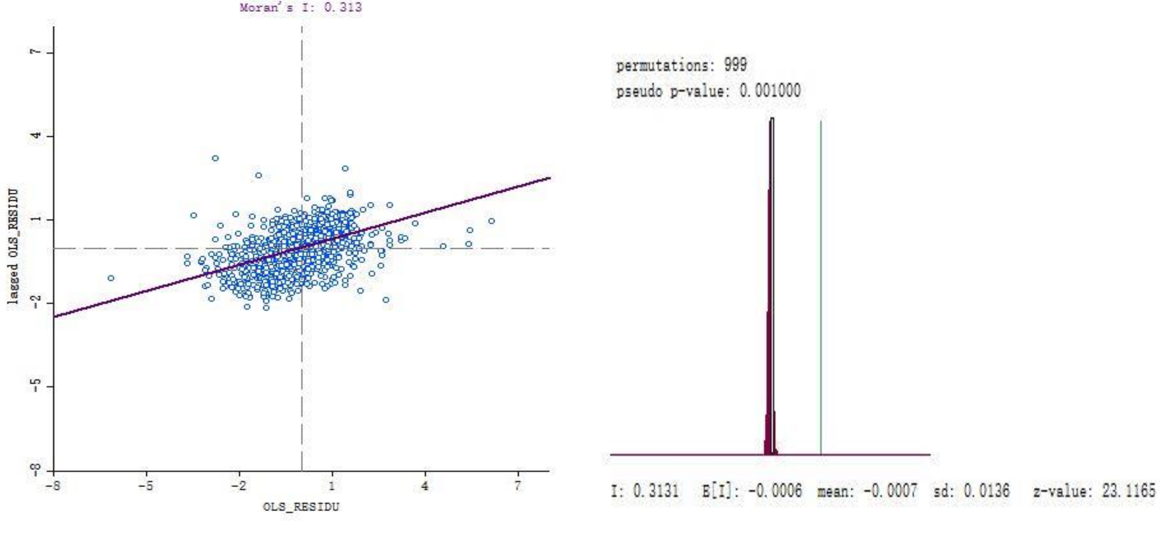
Figure 5 below is the scatterplot of the weighted residuals (WT_RESIDU) and residual values (OLS_RESIDU) calculated in the OLS model by GeoDa. The results show that there is a positive linear correlation between those two. Meanwhile, the slope b value - which population correlation coefficient of ρ has the value of 0.733, which is not equal to 0, thus we can reject the null hypothesis and favor the alternative that the correlation coefficient is spatially autocorrelated.



```
#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  0.230  -0.005    0.008     -0.607     0.544     0.733     0.032     22.629       0
  0     0       0        0          0          0         0         0          0          0
1720  0.230  -0.005    0.008     -0.607     0.544     0.733     0.032     22.629       0
              Chow test for sel/unsel regression subsets: can't compute
```

Figure 5. Scatterplot of WT_RESIDU and OLS_RESIDU

Figure 6 below shows the Moran's I scatterplot between the OLS residuals (OLS_RESIDU) and its lagged Residual values of the all Queen neighbor of the original block (OLS-RESIDU). We can observe that data are split in all of the four quadrants, with most of them distributed on the first (positive-positive) and third (negative-negative) quadrants: which means those are the blocks groups with positive values are surrounded by other positive values, and negative values are surrounded by other negative values. By looking at the pseudo p-value, the results show that there is significant spatial autocorrelation of the residuals.



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

```
I: 0.3131   E[I]: -0.0006   mean: -0.0007   sd: 0.0136   z-value: 23.1165
```

Figure 6 & 7. OLS Moran's I and permutation scatterplot

3.3 Spatial Lag and Spatial Error Regression

W_LNMEDHVAL is the spatial lag of the dependent variable LNMEDHVAL. In other words, it's the value of LNMEDHVAL in nearby areas. It is a significant predictor since the probability of W_LNMEDHVAL is less than 0.05. That is, median house value in an area is associated with median house value in surrounding areas. The coefficient ρ (0.651) is significant as well, and it indicates that W_LNMEDHVAL is positively associated with the dependent variable LNMEDHVA.

Looking at Table 2, we also found that the other predictors LNNBELPOV, PCTBACHMOR, PCTSINGLES, and PCTVACANT are significant since all of their p-values are less than 0.05. The predictors are significant in both OLS and SL models. The positive sign of the predictor's coefficients indicates changes in PCTBACHMOR and PCTSINGLES are associated with a positive change in the dependent variable LNMEDHVAL. On the contrary, the negative sign of the predictors' coefficients means that changes in LNNBELPOV and PCTVACANT are associated with a negative change in LNMEDHVAL.

Heteroscedasticity could be tested by Breusch-Pagan test. Since the p-value of Breusch-Pagan test is less than 0.05 in Table 2, we can reject the null hypothesis of homoscedasticity by having in favor of the alternative hypothesis of heteroscedasticity of the residuals. The AIC of OLS regression is about 1432.99, which is greater than the spatial lag regression's AIC (523.123). The SC of OLS regression is about 1460.24, which is greater than the spatial lag regression's SC (555.824). The smaller the value of AIC/SC is, the better fit of the regression model is. In this case, the spatial lag model does a better job than the OLS model. The log likelihood of the spatial lag regression is -255.562, which is greater than the OLS regression's log likelihood (-711.493). Since a higher log likelihood indicates a better fit of the regression, the spatial lag model performs better than the OLS model. According to the Likelihood Ratio Test result, the p-value is close to zero (p <0.05), which indicates we can reject the null hypothesis and assume that the spatial lag model is a better fit than the OLS regression model.

```
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set          : Regression Data
Spatial Weight     : queen weight
Dependent Variable  :   LNMEDHVAL  Number of Observations: 1720
Mean dependent var  :     10.882  Number of Variables   :   6
S.D. dependent var  :    0.62972  Degrees of Freedom   : 1714
Lag coeff.   (Rho)  :   0.651107

R-squared         :   0.818603  Log likelihood       :   -255.562
Sq. Correlation    : -          Akaike info criterion :    523.123
Sigma-square       :  0.0719325  Schwarz criterion    :     555.824
S.E of regression   :   0.268202

-----------------------------------------------------------------------------
Variable      Coefficient     Std.Error      z-value    Probability
-----------------------------------------------------------------------------
W_LNMEDHVAL     0.651107      0.0180482        36.076     0.00000
CONSTANT       3.89835       0.20109       19.3861     0.00000
LNNBELPOV    -0.0340632     0.00629222       -5.41355     0.00000
PCTBACHMOR    0.00851569     0.00052192        16.3161     0.00000
PCTSINGLES    0.00202905     0.00051571        3.93448     0.00008
PCTVACANT   -0.00852676     0.00074357       -11.4673     0.00000
-----------------------------------------------------------------------------
REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                            DF      VALUE       PROB
Breusch-Pagan test                  4       220.5298     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : queen weight
TEST                            DF      VALUE       PROB
Likelihood Ratio Test               1       911.8633     0.00000
```

Table 2. Spatial Lag Regression Output

Looking at Figure 8 and 9 below, there is slight negative spatial autocorrelation in the spatial lag residuals since Moran's I is less than zero, which indicates observations that are closer to each other markedly different values. According to the results of the pseudo p-value, mean of the permuted residuals, there is still spatial autocorrelation of the residuals. Moran's I of the LAG residuals is -0.082, smaller than the one of OLS residuals (0.313), which indicates there is less spatial autocorrelation in LAG residuals than in OLS residuals. Based on all of these criteria, the spatial lag model performs better than the OLS model.

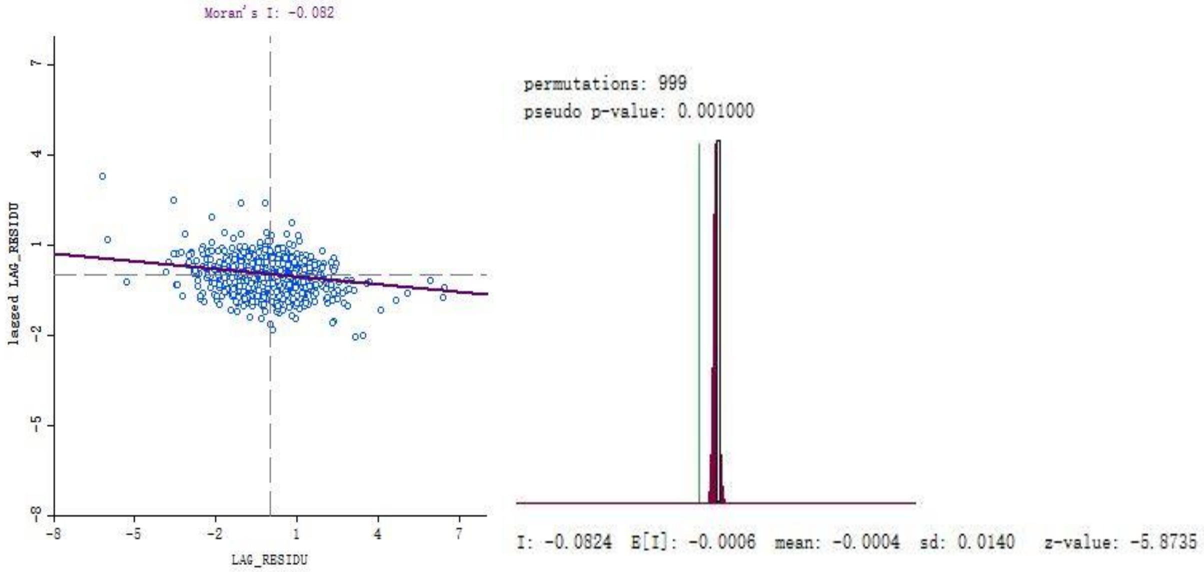


Figure 8 & 9. Spatial Lag Regression Moran's I and permutation scatterplot

The coefficient of lagged residuals is denoted as LAMBDA (λ) in GeoDa. As rho (ρ) in the spatial lag model, lambda (λ) in the spatial error model is another measure of spatial autocorrelation. Lambda (0.814872) is significant since its p-value is close to zero, indicating that the positive spatial autocorrelation in LNMEDHVAL (or OLS residuals) is accounted for by spatially lagged residuals. In this model, lagged residuals can be thought of as OLS residuals at nearby locations while residuals are from the OLS model. According to the Likelihood Ratio Test result, the p-value is close to zero ($p < 0.05$), which indicates we can reject the null hypothesis and assume that the spatial lag model is a better fit than the OLS regression model.

Looking at Table 3, we also found that the other predictors LNNBELPOV, PCTBACHMOR, PCTSINGLES, and PCTVACANT are significant since all of their p-values are less than 0.05. The predictors are significant in both OLS and SE models. The positive sign of the predictor's coefficients indicates changes in PCTBACHMOR and PCTSINGLES are associated with a positive change in the dependent variable LNMEDHVAL. On the contrary, the negative sign of the predictors' coefficients means that changes in LNNBELPOV and PCTVACANT are associated with a negative change in LNMEDHVAL.

Heteroscedastic could be tested by Breusch-Pagan test. Since the p-value of Breusch-Pagan test is less than 0.05 in Table 3, we can reject the null hypothesis of homoscedasticity by having in favor of the alternative hypothesis of heteroscedasticity of the residuals. The AIC of OLS regression is about 1432.99, which is greater than the spatial error regression's AIC (754.985). The SC of OLS regression is about 1460.24, which is greater than the spatial error regression's SC (782.235). The smaller the value of AIC/SC is, the better fit of the regression model is. In this case, the spatial error model does a better job than the OLS model. The log likelihood of the spatial error regression is -372.493, which is greater than the OLS regression's log likelihood (-711.493). Since a higher log likelihood indicates a better fit of the regression, the spatial error model performs better than the OLS model. According to the Likelihood Ratio Test result, the p-value is close to zero (p

<0.05), which indicates we can reject the null hypothesis and assume that the spatial error model is a better fit than the OLS regression model.

```
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM
LIKELIHOOD ESTIMATION
Data set           : Regression Data
Spatial Weight      : queen weight
Dependent Variable  :  LNMEDHVAL  Number of Observations: 1720
Mean dependent var  :   10.882000  Number of Variables   :    5
S.D. dependent var  :    0.629720  Degrees of Freedom    : 1715
Lag coeff. (Lambda) :    0.814872

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

-----------------------------------------------------------------------------
Variable      Coefficient     Std.Error      z-value    Probability
-----------------------------------------------------------------------------
CONSTANT        10.9062      0.0534556       204.023     0.00000
LNNBELPOV    -0.0345369     0.00708851      -4.87224     0.00000
PCTBACHMOR   0.00982427    0.000728944       13.4774     0.00000
PCTSINGLES   0.00266586    0.000620803       4.29421     0.00002
PCTVACANT   -0.00577991    0.000886626        -6.519     0.00000
LAMBDA         0.814872      0.0163744        49.765     0.00000
-----------------------------------------------------------------------------
REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                          DF      VALUE        PROB
Breusch-Pagan test             4    211.1640    0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : queen weight
TEST                          DF      VALUE        PROB
Likelihood Ratio Test          1    678.0016    0.00000
```

Table 3. Spatial Error regression Output

Looking at Figure 10 and11, there is slight negative spatial autocorrelation in the spatial error residuals since Moran's I is less than zero, which indicates observations that are closer to each other have different values. Since Moran's I is less than zero, there is a negative linear correlation, which indicates observations that are closer to each other have markedly different values. According to the results of the pseudo p-value, mean of the permuted residuals, there is still spatial autocorrelation of the residuals. Moran's I of the SE residuals is -0.095, smaller than the one of OLS residuals (0.313), which indicates it substantially lower the spatial autocorrelation in SE residuals than in OLS residuals. Based on all of these criteria, the spatial error model performs better than the OLS model.

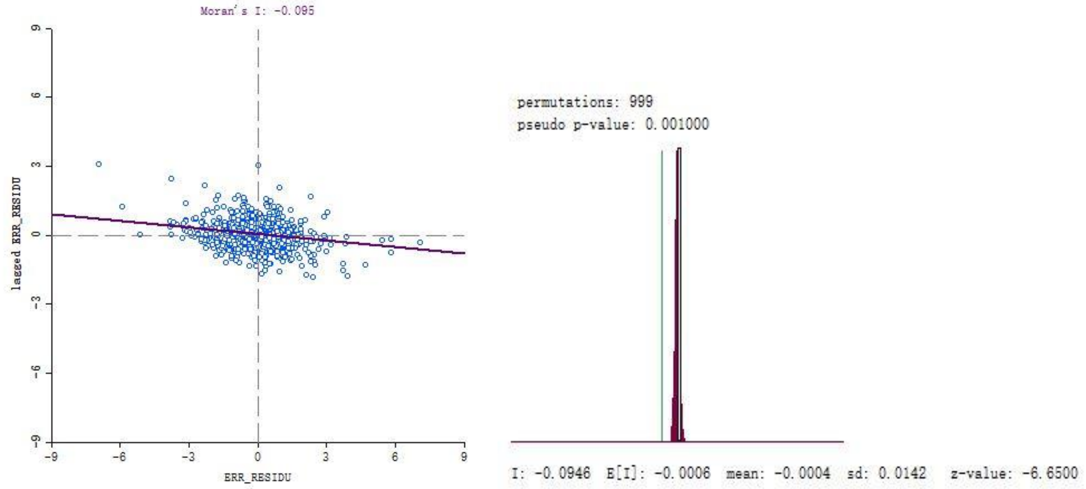
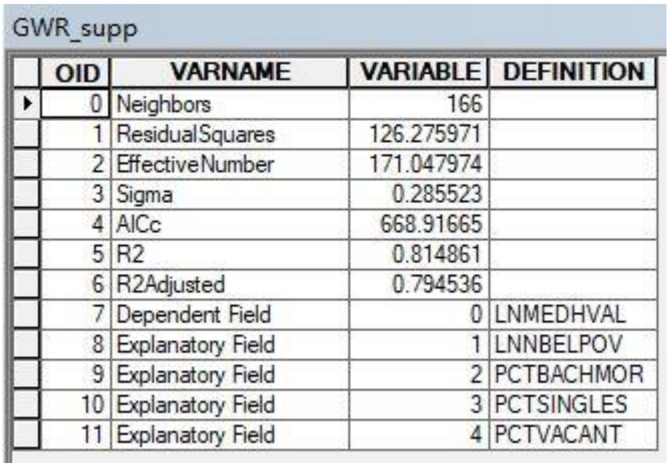


Figure 10 & 11. Spatial Error Moran's I and permutation scatterplot

Moran's I of the SE residuals is -0.095, which is similar to the one of the SL residuals is -0.082. Moran's I from both models are close to zero, which indicates spatial regression models have a substantially lower spatial autocorrelation than in OLS. In the further comparison, the AIC/SC of Spatial Lag model is lower than the ones of Spatial Error, illustrating Spatial Lag model does a better job than Spatial Error model.

3.4 Geographically Weighted Regression Results

The overall R-squared of GWR regression is about 0.81, which is significantly greater than the R-square of OLS regression (0.66), which means that more proportion of the variance of dependent variable can be explained by its predictors in GWR model than the OLS model. Therefore, GWR model performs a better job of explaining the variance in the dependent variable than the OLS model. In addition, the AIC of GWR is about 668.92, which is greater than the spatial lag regression's AIC (523.123) and smaller than the spatial error regression's (754.985). Since the lower AIC means the better fit, spatial lag model does a better job than GWR as well as spatial error model.


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	LNNBELPOV
9	Explanatory Field	2	PCTBACHMOR
10	Explanatory Field	3	PCTSINGLES
11	Explanatory Field	4	PCTVACANT

Table 4. GWR results

Figure 12 and 13 shows the Moran’s I scatter plot and results of 999 permutations for GWR residuals. The Moran’s I for GWR residuals is 0.083, which is smaller than the Moran’s I for OLS residuals (0.313), suggesting that GWR residuals have less spatial autocorrelation and confirm that GWR model performs better than OLS model. Both the Moran’s I values for spatial lag (-0.082) and spatial error residuals (-0.095) are negative, which is opposite to the GWR residuals and suggested a little negative spatial autocorrelation. The Moran’s I for the residuals of all GWR model, spatial log model, and spatial error model are close to approximately 0, indicating they substantially lower the spatial autocorrelation than the OLS model.

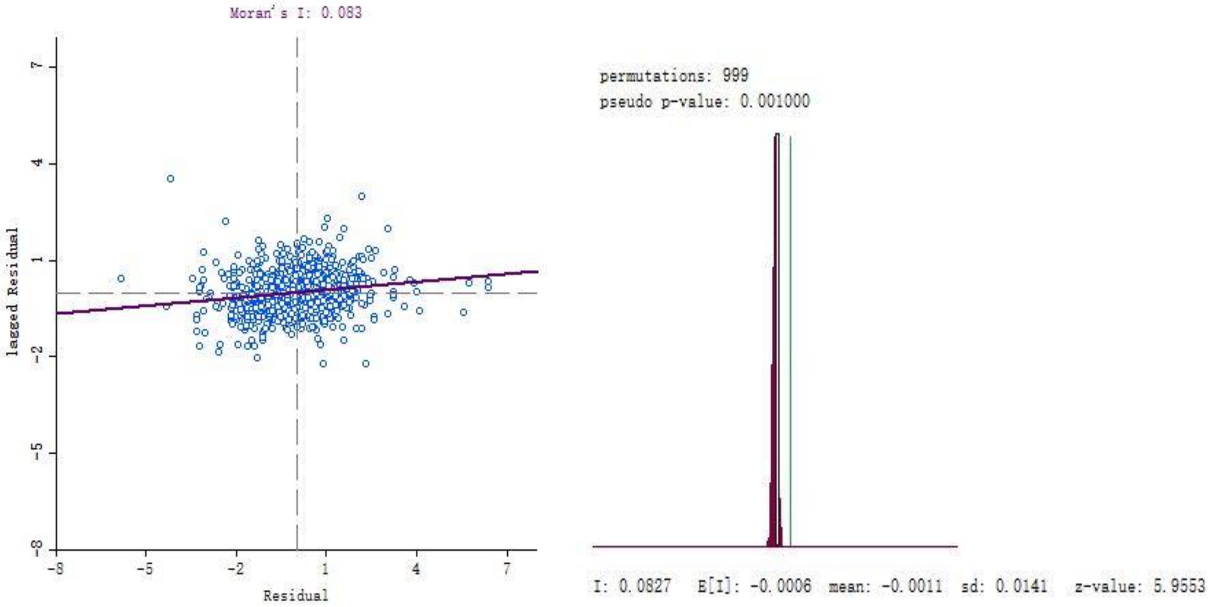


Figure 12&13. Spatial Error Moran’s I and permutation scatterplot

According to the local regression results in Figure 14, there is no positive relationship with the dependent variable that’s possibly significant (dark red area) in LNNBELPOV and PCTVACANT. Some regions in the northwestern and southeastern corner, east side and the central city show a negative relationship with LNMEDHVAL that’s possibly significant in LNNBELPOV, while for PCTVACANT, some regions in southern, western, northwestern, and eastern also show a negative relationship with the LNMEDHVAL that’s possibly significant. Most areas in Philadelphia, especially the western half, have a positive, possibly significant relationship between PCTBACHMOR and LNNBELPOV. North Philadelphia indicates a positive, possibly significant relationship between PCTSINGLES and LNNBELPOV.

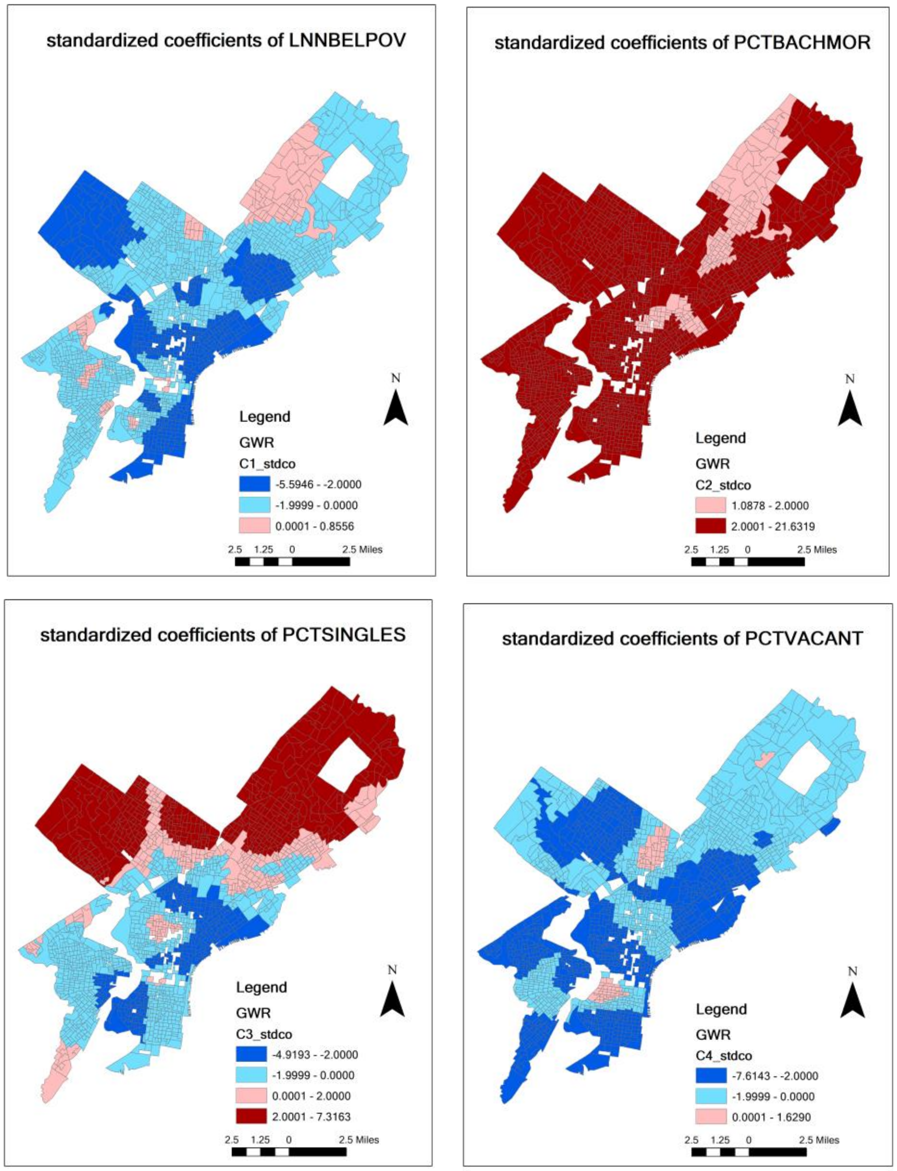


Figure 14. local regression results

As figure 15 shows, local R-squared varies greatly throughout the city. It is poorly fit with most parts of Philadelphia, especially in the center where local R-squared is below 0.22. A few parts in northwestern and western Philadelphia show good fit. The vast variation in coefficient estimates suggest that GWR model omits some important predictors, such as median house value and median household income.

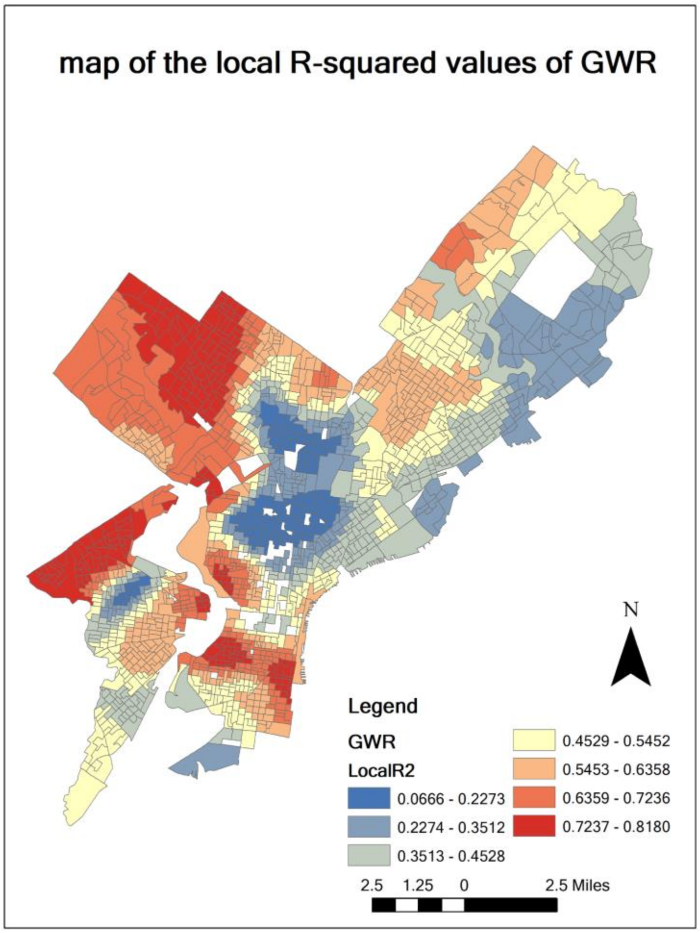


Figure 15. map of local R-squared results

4. Discussion

In this project we examine the relationship between median house values and several neighborhood characteristics based on our investigation of spatial regressions models including spatial lag, spatial error, and geographically weighted regression using GeoDa, ArcGIS and R Studio.

Since Moran's I of OLS indicates there is significant spatial autocorrelation in residuals, we run the spatial lag and spatial error models and examine residuals in each model for spatial autocorrelation. Using AIC, Log Likelihood and the Likelihood Ratio Test, we compared Spatial Lag, Spatial Error model with OLS and found they are better fit than OLS with lower AIC/SC and higher Log Likelihood. Compared to Spatial Lag, Spatial Error performs better with a lower AIC/SC.

Finally, we run the GWR model and compare it with previous regression models. GWR is a better model than OLS with a lower AIC， SC and greater R-squared. We did regressions both from GeoDa and R but the output of GWR is different from two methods. In R output, which is attached in the bottom of research, GWR does a better fit than Spatial Error and Spatial Lag with a lower AIC, SC and lower spatial autocorrelation in residuals. However, AIC of GWR from GeoDa is greater than the one of Spatial Lag model while Moran's I of SL and GWR are similar from GeoDa. In that case, Spatial Lag performs the best. Ultimately, the reduction of Moran's I in three spatial regressions models substantially lowered the spatial autocorrelation compared to the OLS model.

By conducting the spatial autocorrelation tests, we found out that block groups with similar median housing values are clustered. However, there are limitations in models we chose. First, residuals may no longer be spatially correlated after being transformed using the Spatial Lag and Spatial Error methods, which may affect the way we assess the validity of the model. Also, the assumption of the First Law of Geography we hold throughout the report may not always be the case. It is possible that close neighbors may not have similar characteristics due to constraints like geographical and artificial barriers. In the context of the United States, racial segregation policies such as redlining, the construction of interstate highway that divided neighborhoods also make the assumption less plausible: we often observe the pattern that an interstate highways easily draw a border between extreme rich, White communities and poor, Black neighborhoods. Luckily, neighborhoods in Philadelphia are not clearly disintegrated by highways (Kramer, 2018), but we still need to consider this factor well when conducting analysis for other cities. In addition, the accuracy of models should be further improved since the AICs are still high. More potential demographic factors that may affect the median income should be considered in the further analysis.

References

[1] "NEDARC - Statistical Terms Dictionary". 2021. Nedarc.Org. https://www.nedarc.org/statisticalhelp/statisticalTermsDictionary.html.

[2] “Interpreting Regression Output in GeoDa and ArcMap”. 2013. MIT Library. https://libraries.mit.edu/files/gis/regression_output_iap2013.pdf

[3] Waters, N. (2016). T obler's First Law of Geography. International encyclopedia of geography: People, the earth, environment and technology, 1-15.

[4] Scott, Lauren; Pratt, Monica. “Answering Why Questions: An introduction to using regression analysis with spatial data”. ArcUser. 2009. Accessed November 3, 2021.https://www.esri.com/news/arcuser/0309/files/why.pdf

[5] Kramer, R. (2018). Testing the role of barriers in shaping segregation profiles: The importance of visualizing the local neighborhood. Environment and Planning B: Urban Analytics and City Science, 45(6), 1106-1121.

[6] "How GWR Works—Arcmap | Documentation". 2021. Desktop.Arcgis.Com. https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/how-gwr-regression-works.htm.

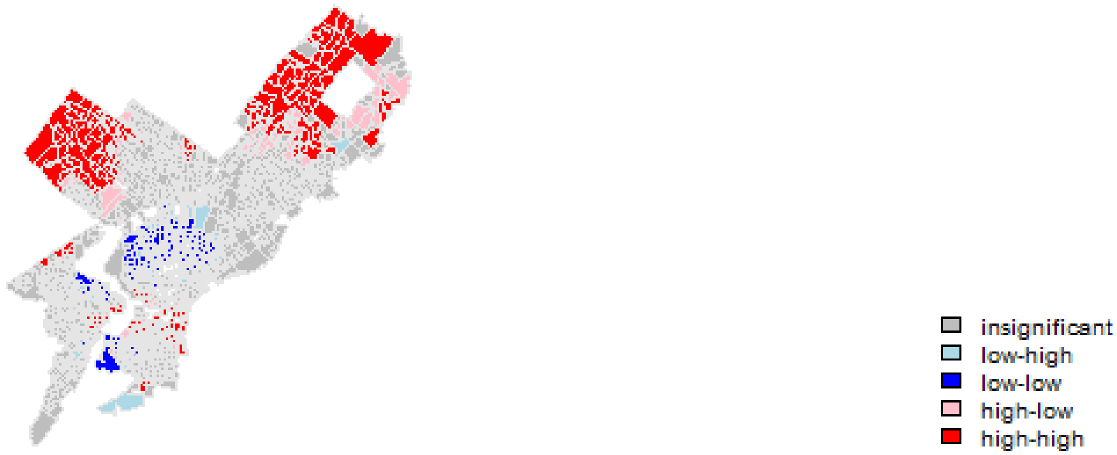
[7] Lee, Kyung Hee, and Michael A. Schuett. 2014. "Exploring Spatial Variations In The Relationships Between Residents' Recreation Demand And Associated Factors: A Case Study In Texas". Applied Geography 53: 213-222. doi:10.1016/j.apgeog.2014.06.018.

[8] 2021.Geos.Ed.Ac.Uk.https://www.geos.ed.ac.uk/~gisteac/fcl/gwr/gwr_arcgis/GWR_Tutorial.pdf

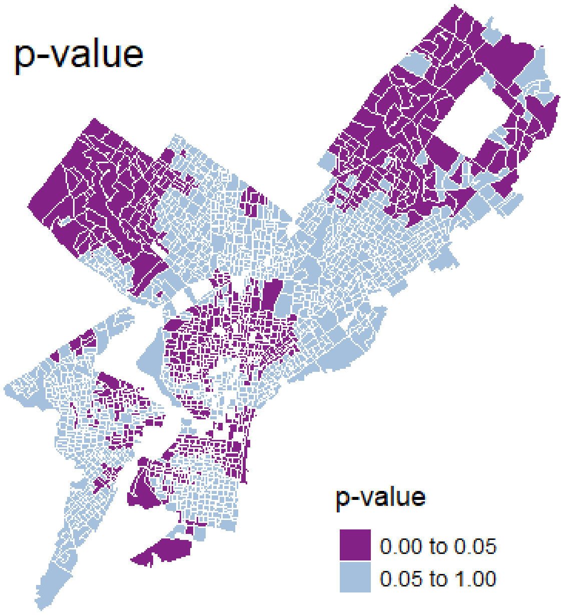
[9] "Arcgis Desktop Help 9.3 - Geographically Weighted Regression (Spatial Statistics)". 2021. Webhelp.Esri.Com. http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Geographically%20Weighted%20Regression%20(Spatial%20Statistics).

Appendix: Plots and Tables from R Studio

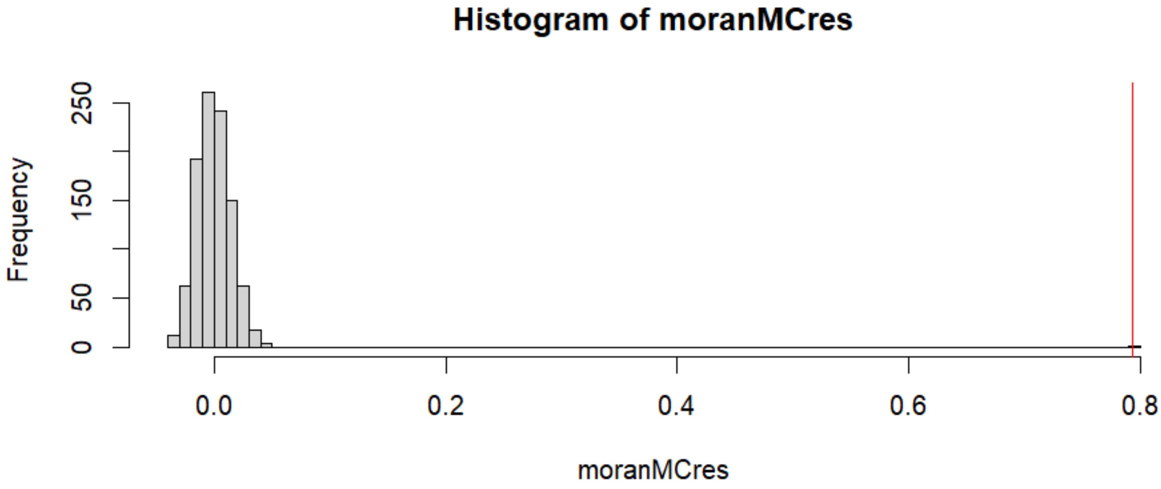
1. LISA Cluster Map



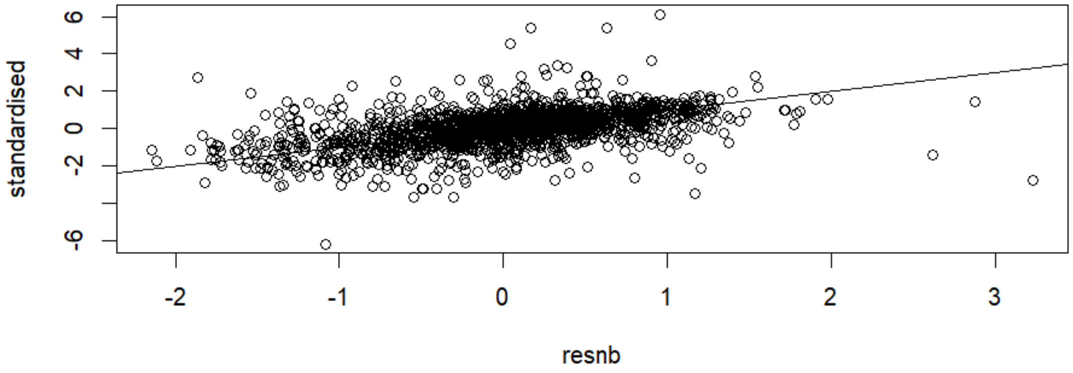
2. P-value Significant Map



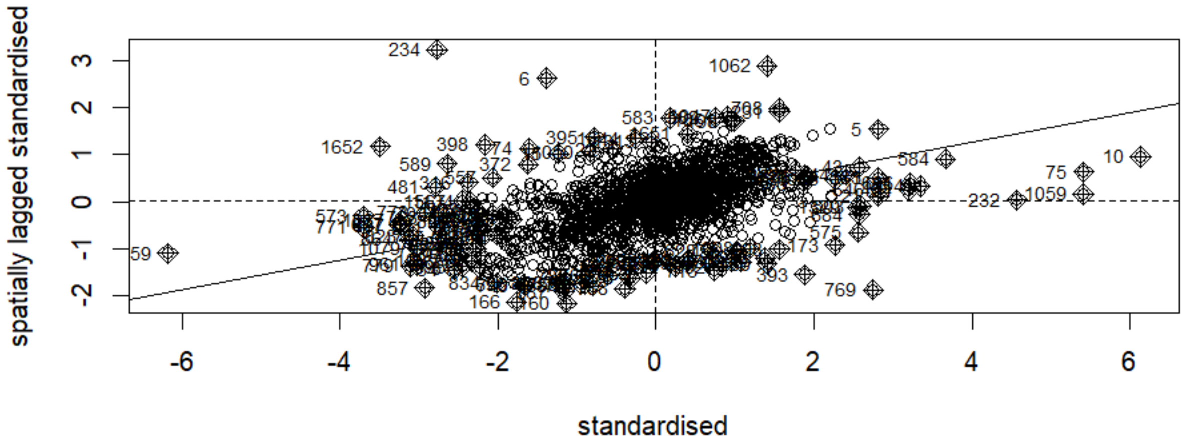
3. Global Moran's I



4. OLS Residuals Scatterplot



5. Moran's I of OLS Residuals Scatterplot

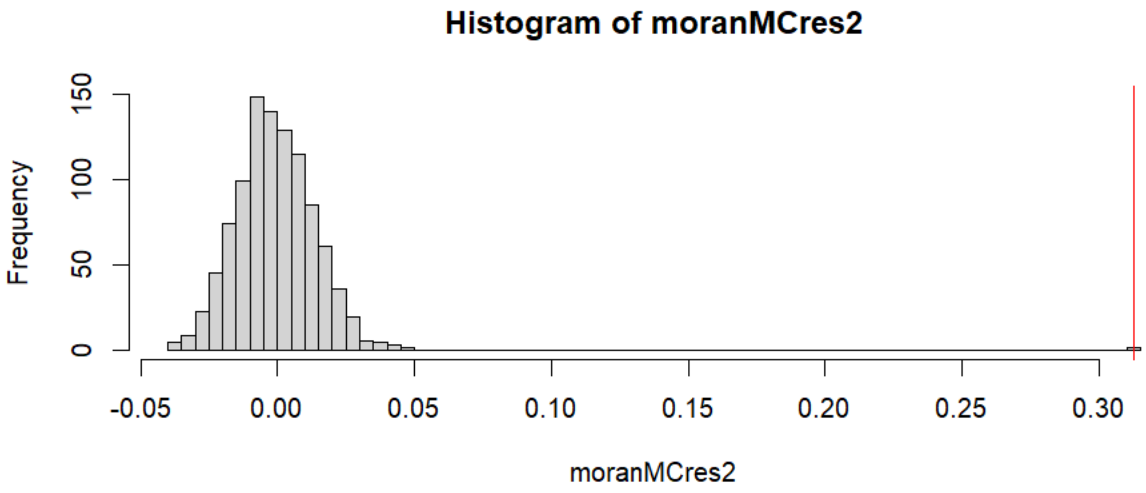


6. OLS Regression Output

```
Call:
lm(formula = standardised ~ resnb)
Residuals:
Min     1Q  Median     3Q    Max
-5.3685 -0.4450  0.0585  0.4618  5.4435

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01281    0.02121  -0.604    0.546
resnb       0.73235    0.03244  22.576   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8793 on 1718 degrees of freedom
Multiple R-squared:  0.2288,      Adjusted R-squared:  0.2283
F-statistic: 509.7 on 1 and 1718 DF,  p-value: < 2.2e-16
```

```
lm(formula = standardised ~ resnb)
Residuals:
Min     1Q  Median     3Q    Max
          -5.3685 -0.4450  0.0585  0.4618  5.4435

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01281    0.02121  -0.604    0.546
resnb       0.73235    0.03244  22.576   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.8793 on 1718 degrees of freedom
Multiple R-squared:  0.2288,      Adjusted R-squared:  0.2283
F-statistic: 509.7 on 1 and 1718 DF,  p-value: < 2.2e-16
```

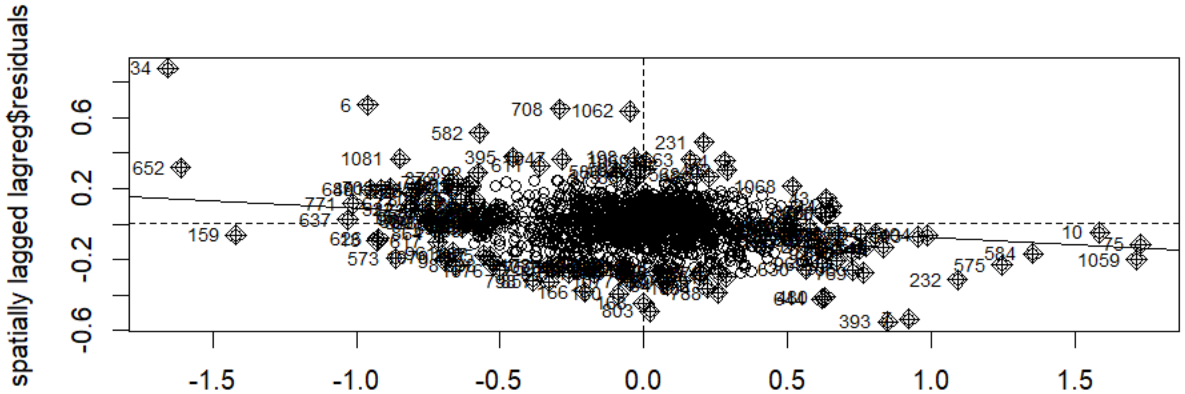
7. Moran’s I of OLS Residuals Histogram



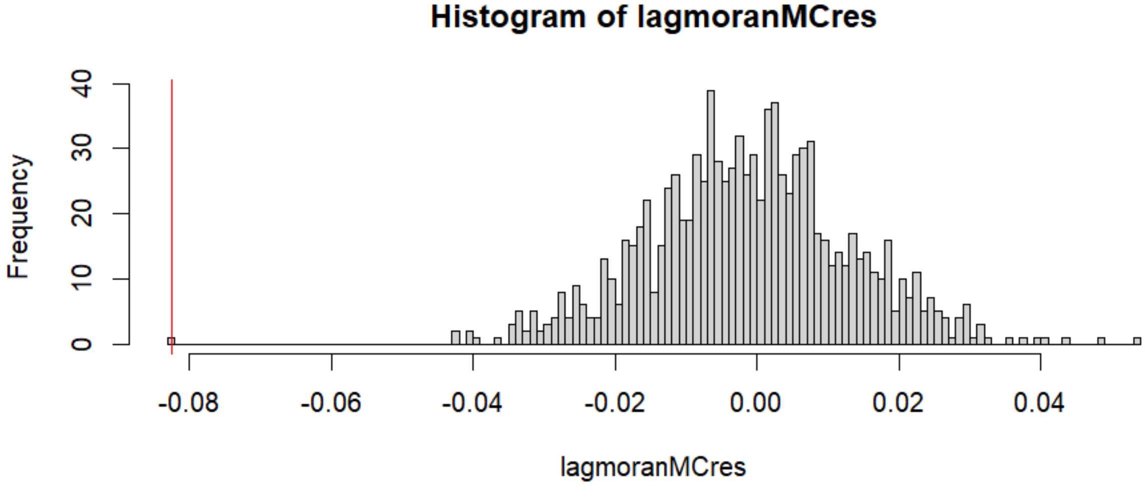
8. Spatial Lag Regression Output

```
Call:lagsarlm(formula = LNMEDHVAL ~ PCTBACHMOR + PCTVACANT +
PCTSINGLES + LNNBelPov100, 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)   3.89845489  0.20111357  19.3843 < 2.2e-16
PCTBACHMOR    0.00851381  0.00052193  16.3120 < 2.2e-16
PCTVACANT    -0.00852940  0.00074367 -11.4694 < 2.2e-16
PCTSINGLES    0.00203342  0.00051577   3.9425 8.064e-05
LNNBelPov100 -0.03405466  0.00629287  -5.4116 6.246e-08

Rho: 0.6511, LR test value: 911.51, p-value: < 2.22e-16
Asymptotic standard error: 0.01805
z-value: 36.072, p-value: < 2.22e-16
Wald statistic: 1301.2, p-value: < 2.22e-16

Log likelihood: -255.74 for lag model
ML residual variance (sigma squared): 0.071948, (sigma: 0.26823)
Number of observations: 1720
Number of parameters estimated: 7
AIC: 525.48, (AIC for lm: 1435)
LM test for residual autocorrelation
test value: 67.737, p-value: 2.2204e-16
```

9. Moran's I of Spatial Lag Residuals Scatterplot



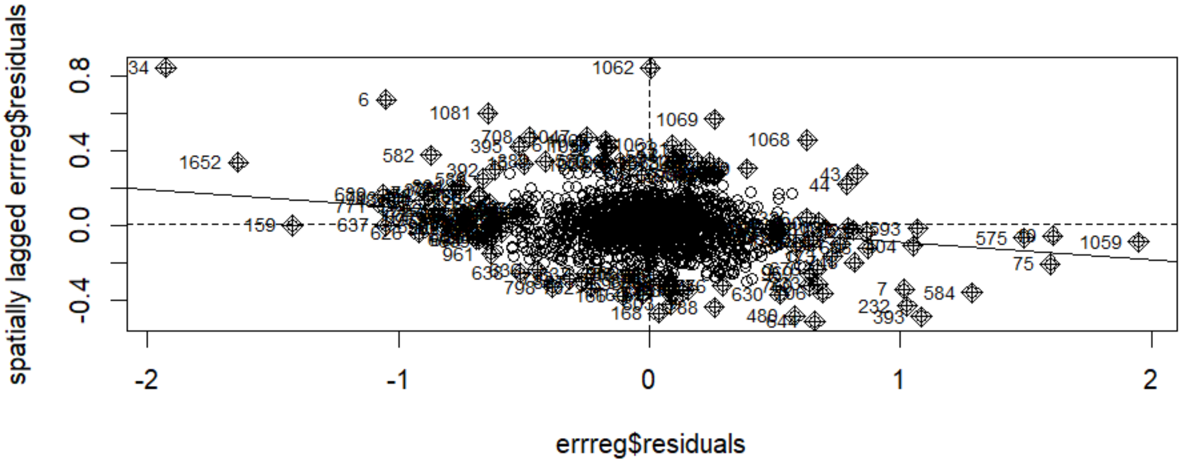
10. Moran's I of Spatial Lag Residuals Histogram



11. Spatial Error Regression Output

```
Call:errorsarlm(formula = LNMEDHVAL ~ PCTBACHMOR +
PCTVACANT + PCTSINGLES + LNNBelPov100, 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)  10.90643420  0.05346781 203.9813 < 2.2e-16
PCTBACHMOR    0.00981293  0.00072896  13.4615 < 2.2e-16
PCTVACANT    -0.00578308  0.00088670  -6.5220 6.937e-11
PCTSINGLES    0.00267792  0.00062083   4.3134 1.607e-05
LNNBelPov100 -0.03453407  0.00708933  -4.8713 1.109e-06

Lambda: 0.81492, LR test value: 677.61, p-value: < 2.22e-16
Asymptotic standard error: 0.016373
z-value: 49.772, p-value: < 2.22e-16
Wald statistic: 2477.2, p-value: < 2.22e-16

Log likelihood: -372.6904 for error model
ML residual variance (sigma squared): 0.076551, (sigma:
0.27668)
```

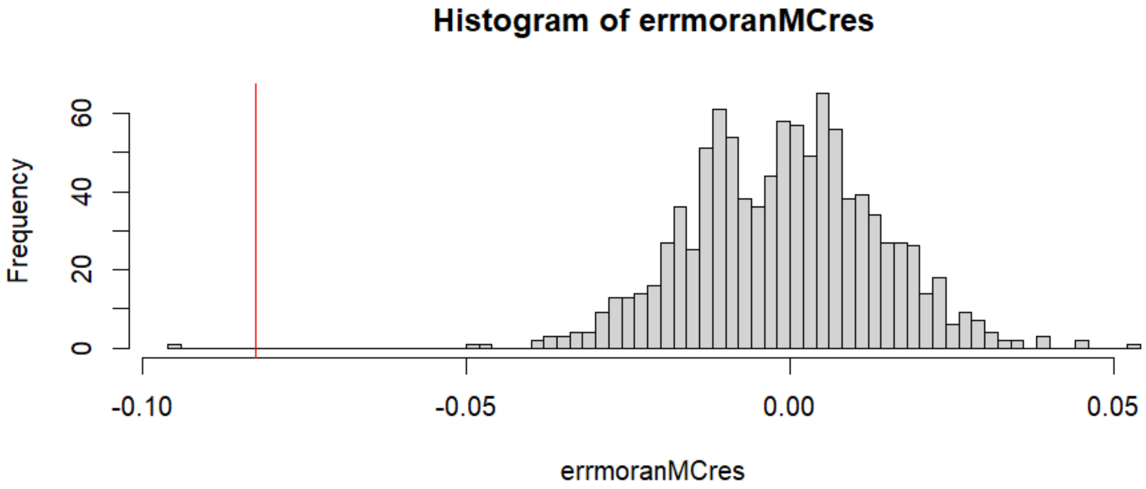
12. Moran's I of spatial error residuals Scatter Plot



13. Moran's I of spatial error residuals Significance Test

```
Monte-Carlo simulation of Moran I

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

statistic = -0.094532, observed rank = 1, p-value = 0.999
alternative hypothesis: greater
```

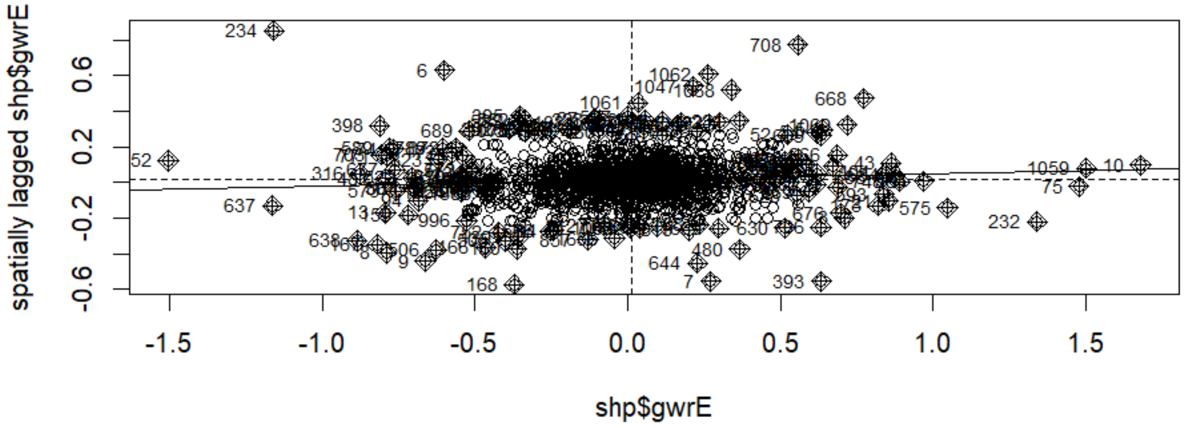
14. Moran's I of spatial error residuals Histogram



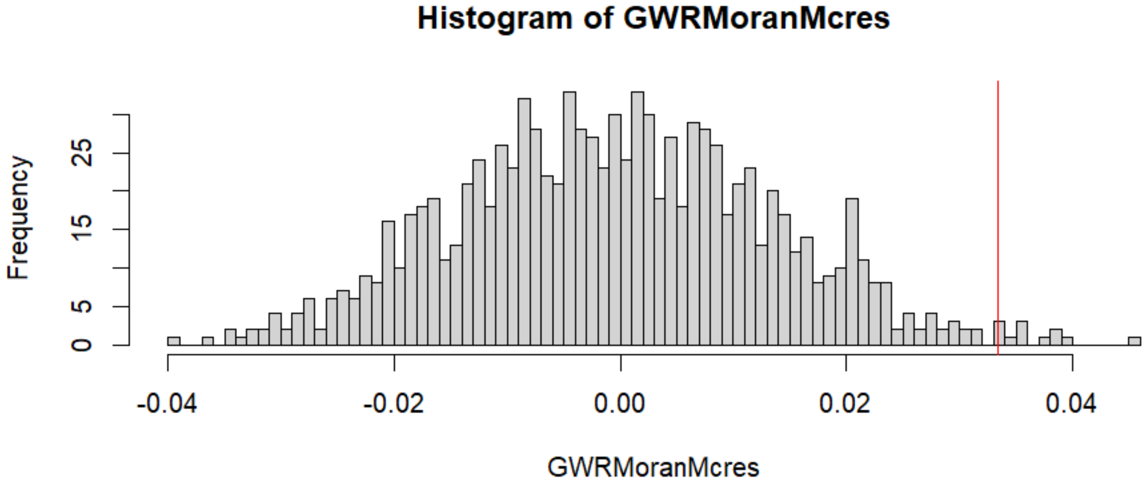
15. Geographically Weighted Regression Output

```
Call:
gwr(formula = LNMEDHVAL ~ PCTBACHMOR + PCTVACANT + PCTSINGLES + LNNBelPov100, 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
PCTBACHMOR    0.0010974  0.0101380  0.0149279  0.0202187  0.0347258  0.0209
PCTVACANT    -0.0317407 -0.0142383 -0.0089599 -0.0035770  0.0167916 -0.0192
PCTSINGLES   -0.0249706 -0.0075550 -0.0016626  0.0042280  0.0143340  0.0030
LNNBelPov100 -0.2365244 -0.0733572 -0.0401186 -0.0126657  0.0948768 -0.0789
Number of data points: 1720
Effective number of parameters (residual: 2traceS - traceS'S): 360.5225
Effective degrees of freedom (residual: 2traceS - traceS'S): 1359.477
Sigma (residual: 2traceS - traceS'S): 0.2762201
Effective number of parameters (model: traceS): 257.9061
Effective degrees of freedom (model: traceS): 1462.094
Sigma (model: traceS): 0.2663506
Sigma (ML): 0.245571
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 660.7924
AIC (GWR p. 96, eq. 4.22): 308.7123
Residual sum of squares: 103.7248
Quasi-global R2: 0.8479244
```

16. Moran's I of the GWR residuals Scatter Plot



17. Moran's I of the GWR residuals Histogram


Histogram of GWRMoranMcres
Frequency
0
5
15
25
-0.04
-0.02
0.00
0.02
0.04
GWRMoranMcres