Hanyong Xu, Xiaoran Wang

Prediction of Median House Values in Philadelphia Block Groups

1. Introduction

This project examines the relationship between the median house values and the other kinds of neighborhood characteristics in the study area of Philadelphia. The data will be organized in the census block group unit. Four neighborhood characteristics are identified that may potentially affect house values: the proportion of residents with at least a bachelor's degree, the proportion of vacant housing units, the percent of detached single-family houses, and the number of households with income below 100% poverty level. These four characteristics are the major predictors for this analysis.

In the previous study, we carried out OLS regression to examine the relationship between the median house values and the predictors. The median house value was the dependent variable and the four predictors are the same as the four predictors in this project. Median house value and the number of households with income below 100% were log-transformed for normalizing the distribution. However, OLS analysis is often inappropriate when dealing with datasets that have spatial components, since it estimates the correlation in a linear regression model regardless of the spatial proximities between variables. In this project, the spatial context of our dataset also reduced the effectiveness of OLS analysis. Thus, the purpose of this report is to use spatial lag, spatial error, and geographically weighted regression to see whether these methods perform better than OLS when dealing with a dataset with spatial context.

2. Methods

2.1 The Concept of Spatial Autocorrelation

Waldo Tobler came up with the first law of geography: "everything is related to everything else, but near things are more related than distant things." This law implies the idea of positive spatial autocorrelation. The concept of spatial autocorrelation describes the relationship of values of a single variable at nearby locations. Based on the first law of geography, if observations that are closer to each other in space have related values, they would have a positive spatial autocorrelation. The negative spatial autocorrelation usually describes a contrasting relationship: the observation that is closer to each other has completely different values. The negative spatial autocorrelation occurs fairly rarely, so we mainly focus on positive spatial autocorrelation.

Moran's I is a method of testing spatial autocorrelation or spatial dependencies. It is defined as the following equation:

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

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

where $\overline{X}$ is the mean of the variable X, $X_i$is the variable value at a particular location i, $X_j$is the variable value at a particular location j, $W_{ij}$is the weight indexing the location of i relative to j, and n is the number of observations. Large, positive values (close to 1) shows a positive autocorrelation, that is to say, clustered spatial order. On the other hand, large, negative values (close to -1) demonstrates there is a negative autocorrelation, meaning the distribution of the variable is dispersed and tend to repel each other. Values close to 0 reflects that there is no spatial autocorrelation. The expected value of the Moran's I is equal to -1/(n-1), where n is the number of observations.

When dealing with spatial analysis, a weight matrix needs to be defined for the spatial proximity. A weight matrix is a table summarizing the pairwise spatial relationships in a dataset. Here, we are using the Queen matrix, which means the neighbors of any census block group A are defined as all block groups that intersect with A either at a point or a segment. Statisticians sometimes like to try different spatial weight matrix to make sure that the result is not dependent on the matrix.

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

H_0:There is no spatial autocorrelation associated with median house values and the four predictors

H_{a1}:There is a positive spatial autocorrelation associated with median house values and the four predictors

H_{a2}:There is a negative spatial autocorrelation associated with median house values and the four predictors

Random permutation calculates a reference distribution for the statistic under the null hypothesis of spatial randomness by randomly permuting the value of observations over the location for a certain number of times. The Moran's I value will be calculated at each time of permutation. Then, it needs to rank all Moran's I values in descending order. The pseudo-p-value is calculated by taking the rank of Moran's I and dividing it by the total number of permutations. If the pseudo-p-value is less than the alpha value, which usually is set at 0.05, then there is a significant spatial autocorrelation, and the null hypothesis would be rejected.

For a better understanding of the autocorrelation, LISA (Local Indices of Spatial Autocorrelation) analysis can be conducted. It can demonstrate to what extent the neighboring values of location i of variable X is related to the value at location i. Basically, the deviation of value from the global mean at location i is compared to the average deviation of its neighboring values. The analysis categorizes the block groups into four categories:

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

2.2 OLS Regression and Assumptions

An OLS (Ordinary Least Square) regression of median house value on the proportion of residents with at least a bachelor's degree, the proportion of vacant houses in each block group, the proportion of detached houses, and the number of residents living in poverty is conducted to find out the linear relationship. The regression coefficient for each variable measures the relationship between the dependent and the independent variables. There are six assumptions for linear regression: there should be a linear relationship between the dependent variable and each of the predictors; each predictor should be independent with each other; the residuals should display a normal distribution; the data should display homoscedasticity; there should be no multicollinearity between predictors; there should be no fewer than 10 observations per predictor. For detailed explanations in OLS regression, one can refer to the previous report.

One of the most important assumptions is that observations are independent of each other, but when the data has a spatial component, the assumption that the errors are random/independent often doesn't hold. In this case, we can use Moran's I to test whether the randomness or independence of errors hold. Furthermore, regressing the OLS residuals on the nearby residuals is an alternative to conduct the test. The "nearby residuals" are the residuals at the neighboring block groups, as defined by the Queen matrix.

Assuming that the value of the dependent variable at one location relates to the values of that variable in the nearby locations, the model includes the spatial lag of the dependent variable as a predictor. The rho (ρ) in this spatial lag model indicates whether the spatial autocorrelation exists. The ρ is the slope of the fitted line of regression of the OLS residuals with their neighboring residuals. It demonstrates the relationship between the residuals and their neighbors.

GeoDa, the software we are using for the spatial analysis, provides tools to check other assumptions. One assumption requires the data to have homoscedasticity, which means variance of the dependent variable does not change as the variance of the predictor changes, can be examined by the Breusch-Pagan Test, the Koenker-Bassett Test, or the White Test. The null hypothesis is that there is no heteroscedasticity, while the alternate hypothesis is having heteroscedasticity. A p-value of less than 0.05 would mean one can reject the null hypothesis of having no heteroscedasticity. Another assumption that GeoDa is capable of testing is the normality of errors. The Jarque-Bera Test can be used to test this assumption. The null hypothesis is there is normal distribution in the error terms (residuals), while the alternative hypothesis is there is non-normality in error terms. A p-value of less than 0.05 means rejecting the null hypothesis of normality.

2.3 Spatial Lag and Spatial Error Regression

The powerful GeoDa software is used to conduct spatial lag and spatial error regressions. Further explanations of both models are presented below.

The spatial lag model assumes the value of the dependent variable at individual locations are related to the values in nearby locations. It includes a geographically weighted lag (the spatial lag) as one of the predictors. The model equation is defined as follows:

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

where W_y is the spatial lag, or the average of the neighboring values the variable y, ρ is the coefficient of the y-lag variable W_y, and the value is between -1 and 1. If ρ is significant, the error from OLS is biased and contains spatial autocorrelations. $\beta_0$is the intercept constant when all predictors (including the spatial lag) are zero;ε is the residual or error term.

The spatial error model assumes that the residual at one location is associated with residuals at nearby locations. The nearby locations are defined by the weight matrix. The OLS regression of the dependent variables on the predictors would be run first. Then, we can regress residuals from this regression on the nearest neighbor residuals, thereby filtering the spatial information out of the OLS residuals and decomposing the residuals into two parts: one with a spatial pattern and another one is simply random noise. The model equation is defined as follows:

Spatial error:

$$y = \beta_0 + \beta_1 PCBACHMORE + \beta_2 PCTVACANT + \beta_3 LNNBELPOV + \beta_4 PCTSINGLES + \varepsilon$$

$$\varepsilon = \lambda W_\varepsilon + \mu$$

$$\Downarrow$$

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

Where β_0 is constant intercept, λW_ε is spatially lagged residuals, λ is the autoregressive coefficient, W_ε is a spatial lag for the errors, and μ is random noise. $\lambda$is constrained within -1 and 1, and positive means positive autocorrelation and vice versa.

Similar to the OLS regression, the six assumptions need to be satisfied in the spatial lag and spatial error regressions. One exclusion is the spatial independence of observations. Other than that, the data need to have each of the predictors linearly related to the dependent variable, have residuals in normal distribution, no heteroscedasticity, and there should not be multicollinearity.

Note that the interpretation of the β coefficients for each predictor in the spatial lag and spatial error model is different from that of the OLS regression, so are their standard errors and the corresponding significance. Due to the inclusion of spatial lag in the model, the assumption of independence of residuals is hopefully met, and the beta coefficients will not be as biased as before.

The goal of the spatial regressions is to consider the spatial autocorrelation in the residuals or in the data. We hope to see the resultant residuals no longer spatially correlated and have less heteroscedasticity.

The fit of the spatial models will be compared with the OLS model by evaluating four criteria, the AIC/SC, the Log Likelihood, the Likelihood Ratio Test, and the Moran's I. The AIC or SC (Akaike Information Criterion / Schwarz Criterion) measures the lost information when using the model to predict values in reality. It reflects the tradeoff between precision and complexity of the model. Thus, the smaller the AIC or SC, the better. Also, note that the comparison of AIC and SC between models are only meaningful when the difference is greater than 3. The Log Likelihood is related to the maximum likelihood method of fitting a model to the data. The maximum likelihood optimizes the values of the model parameters and choose the ones that are "most likely". The higher (the less negative) the Log Likelihood, the better the model fits. Note that this criterion can only be used for comparing nested models, meaning one can only compare the Log Likelihood of OLS with spatial lag or between OLS and spatial error. The Likelihood Ratio Test compares the OLS with the spatial model. The null hypothesis is spatial lag or spatial error model is not a better fit than the OLS, while the alternative hypothesis is the reverse. This means with a p-value less than 0.05 one can reject the null hypothesis of the spatial model is not better. The last criterion is the Moran's I, which is already explained above. When looking at the Moran's I of the residuals, the closer the Moran's I to zero, the less spatial autocorrelation, the better the model.

2.4 Geographically Weighted Regression

The ArcGIS will be used to finish the Geographically Weighted Regression analysis (GWR). Further explanations of GWR analysis are presented below.

GWR is a way to deal with spatial non-stationarity. It assumes that the modeled relationships are not constant across space. Instead of having a single global-level regression, we have a separate, local regressions for each location. The location could be in point, block group, or tract. The local difference makes the relationship no longer constant. This scenario can be explained by the Simpson's paradox: the trend appears in multiple different groups reverses or disappears when these groups are combined together. In other words, when we disaggregate dataset and observe different regions as individual groups, we may see a different or reversed relationship between two variables in different areas. Thereby, rather than fitting them into a single regression model, we can use local regression to fit the subsets at different locations into different models, and it is what the local regression stands for.

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

and i = 1 ... n

In the above equation, for each location i:

β_{i1}= when *PCBACHMORE* increases 1 percent, the *LNMEDHVAL* would increase β_{i1}unit, *MEDHVAL* increase by $(e^{\beta_{i1}} - 1) * 100\%$, holding other variables constant
β_{i2}= when *PCTVACANT* increases 1 percent, the *LNMEDHVAL* would increase β_{i2}unit, *MEDHVAL* increase by $(e^{\beta_{i2}} - 1) * 100\%$, 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 $(1.01^{\beta_{i3}} - 1) * 100\%$, holding other variables constant
β_{i4}= when *PCTSINGLES* increases 1 percent, the *LNMEDHVAL* would increase β_{i4}dollars, *MEDHVAL* increase by $(e^{\beta_{i4}} - 1) * 100\%$, holding other variables constant
ε_i= the residuals
There are n locations (block groups) in total.

The local regression of GWR is the regression for each point or area. In our case, it is a regression for each census block group. To run a regression for each location, multiple observations or locations are needed. The GWR uses other observations from the dataset to run the regression. Observations closer to a location usually have more influence on the result of parameters for that location. In other words, the observations are assigned with weights and closer observations have higher weights. The weight of an observation is different in each local regression for different locations.

The weights discussed in the previous paragraph can be defined with a weighing function with different bandwidth. There are two types of bandwidth, fixed and adaptable. A fixed bandwidth (kernel) means for each location, the observations within the fixed bandwidth distance are taken and are regressed upon according to the weighing function to these observations. The number of observations included in each local regression can be different. An adaptive bandwidth (kernel) means for each location the number of observations is fixed, and the bandwidth is changed according to that. The weighing function applies to the observations within the bandwidth accordingly. In each local regression, the weighing function gives the closer observations heavier weights, while it gives any observations outside the bandwidth distance a weight of zero. Generally, fixed bandwidth kernel is better when the distribution of observations is relatively even across space, while adaptive bandwidth kernel is better when the distribution of observations varies across space. In our case, the distribution of the block groups in Philadelphia varies a lot, with small, dense block groups in the center and large, more spread-out ones at the edge of the city. Thus, we will be using the adaptive bandwidth.

We run OLS prior to the GWR to ensure that the model is reasonable and there are relationships which are worth exploring, and many OLS assumptions still hold in GWR, such as the normality of residuals, homoscedasticity, no multicollinearity, and the residuals are close to normal. When there is a value of an independent variable spatially cluster in a substantial way, it means there is a problem with local multicollinearity. If the GWR has two or more variables that have similar patterns of value at all locations in a certain part of the study area, it also implies the problem with multicollinearity. We can use the Condition Number from the "*Cond.Number*" field to find out when the results are unstable due to the local multicollinearity. The condition number would indicate the instability of the results due to local multicollinearity. When the Condition number of a variable is larger than 30, equal to Null, or equal to - 1.7976931348623158e + 308, the variable would potentially cause the instability of the results.

P-values are not included in the GWR due to its complexity. In GWR, a regression is run for each observation, thus resulting in a large number of regressions with multiple parameters need to be estimated. Recall there exists the possibility of type I and type II errors with the OLS regression, which are rejecting the null hypothesis when it is true and falsely rejecting the null hypothesis when it is not true. The same error exists in GWR, and with the large number of regressions run during the GWR, the total number of significant estimations due to chance is greatly increased. The p-values in these cases are not reliable in terms of rejecting the null hypothesis. Thus, they are not included.

This report will not focus on local regression results, except the local R-squared results. Instead, the report will mainly focus on more global diagnostics, such as AIC and the Moran's I of the GWR residuals that are obtained from GeoDa. In addition, the ratio of the beta coefficients and the standard error for the local regressions will be examined.

3. Results

3.1 Spatial Autocorrelation

We calculated global Moran's I values for the dependent variable and plotted a scatter plot to show the spatial lag among the levels of the dependent variable in figure 1. The global Moran's I value is 0.794, which is larger than 0.5, and it means that the dependent variable has a high positive spatial autocorrelation. The similar median house values tend to cluster together. Then we conducted the random permutation test on the results, and the k is 999 times. The Moran's I for the 999 permutations is 0.7636, which is lower than the Moran's I for the median house value. The p-value is 0.001, so there is not a single statistic computed from the randomly generated samples is larger than the observed values. The dependent variable is statistically significantly spatial autocorrelated, so we reject the null hypothesis of no spatial autocorrelation.

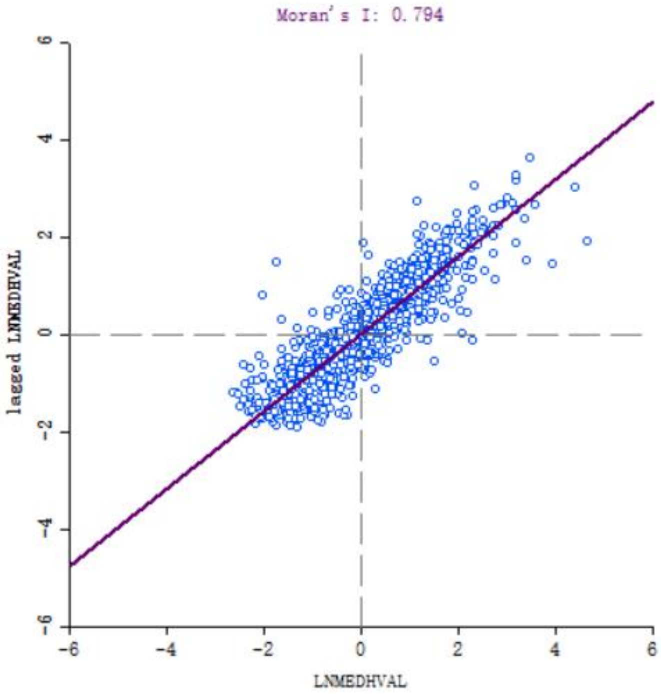


Fig. 1. Global Moran's I value for median house value.

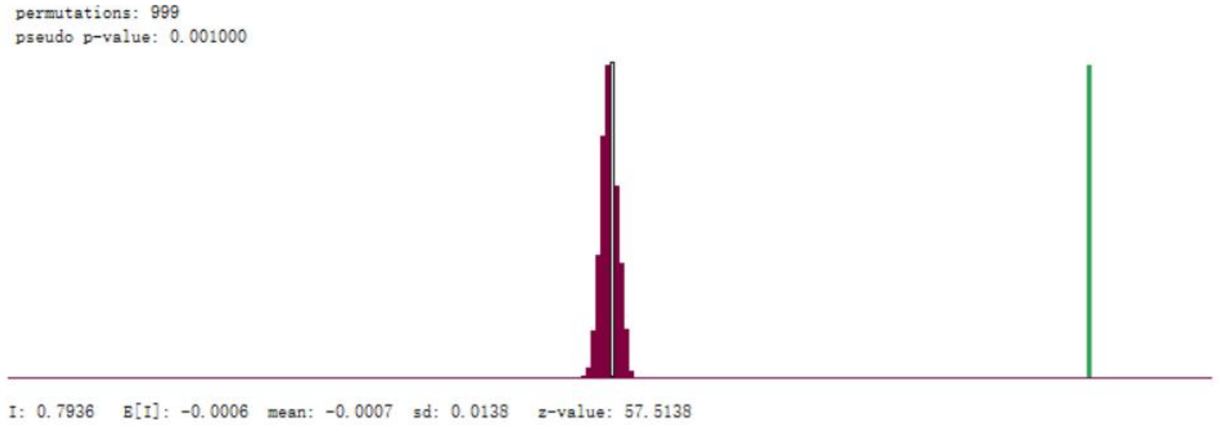


Fig. 2. Histogram of Moran's I value for random permutation test.

For the local Moran's I, we plotted the significance map and cluster map to display the results (Fig. 3, Fig. 4). The majority of Philadelphia are in the not significant, high-high and low-low areas. Half of the Northwest and West Philadelphia, ⅓ of Northeast and South Philadelphia, and the area at Kensington, Bridesburg, Richmond are the not-significant areas. The high-high areas mainly cluster at the Northeast, Northwest and South Philadelphia. A small proportion of the high-high area is at the Center City and West Philadelphia. The low-low areas cluster at the North Philadelphia, and parts of the West and Southwest Philadelphia. The low-high area separately distributes at Northeast, West and South Philadelphia. The high-low areas appear at the West and South Philadelphia.

In terms of all significant areas, the high-high areas at Northwest and South Philadelphia are more significant due to the low p-values, and it means these areas have strong positively spatial autocorrelations. Similarly, the significant low-low area at North Philadelphia has a strong positively spatial autocorrelation (Fig. 4).

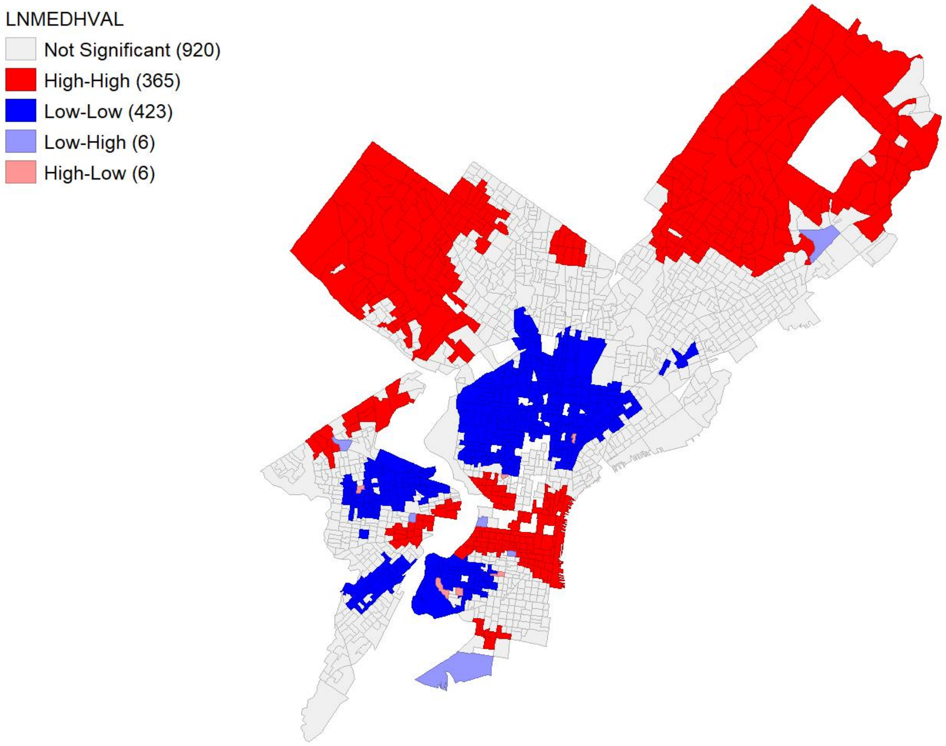


Fig. 3. Map of the clusters of local Moran’s I

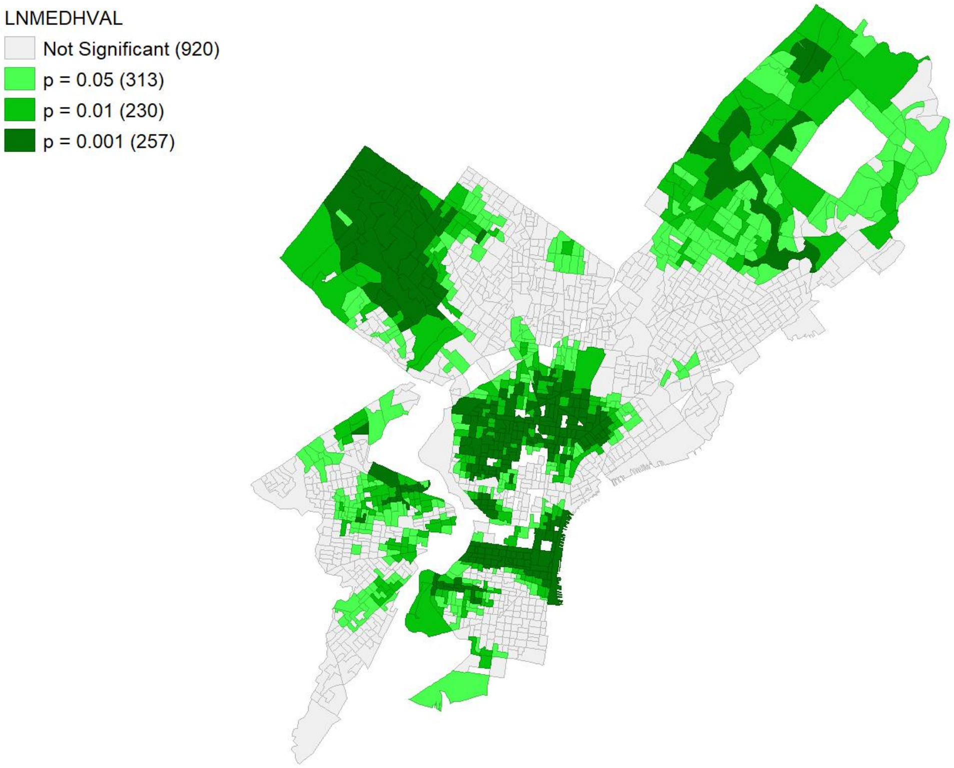


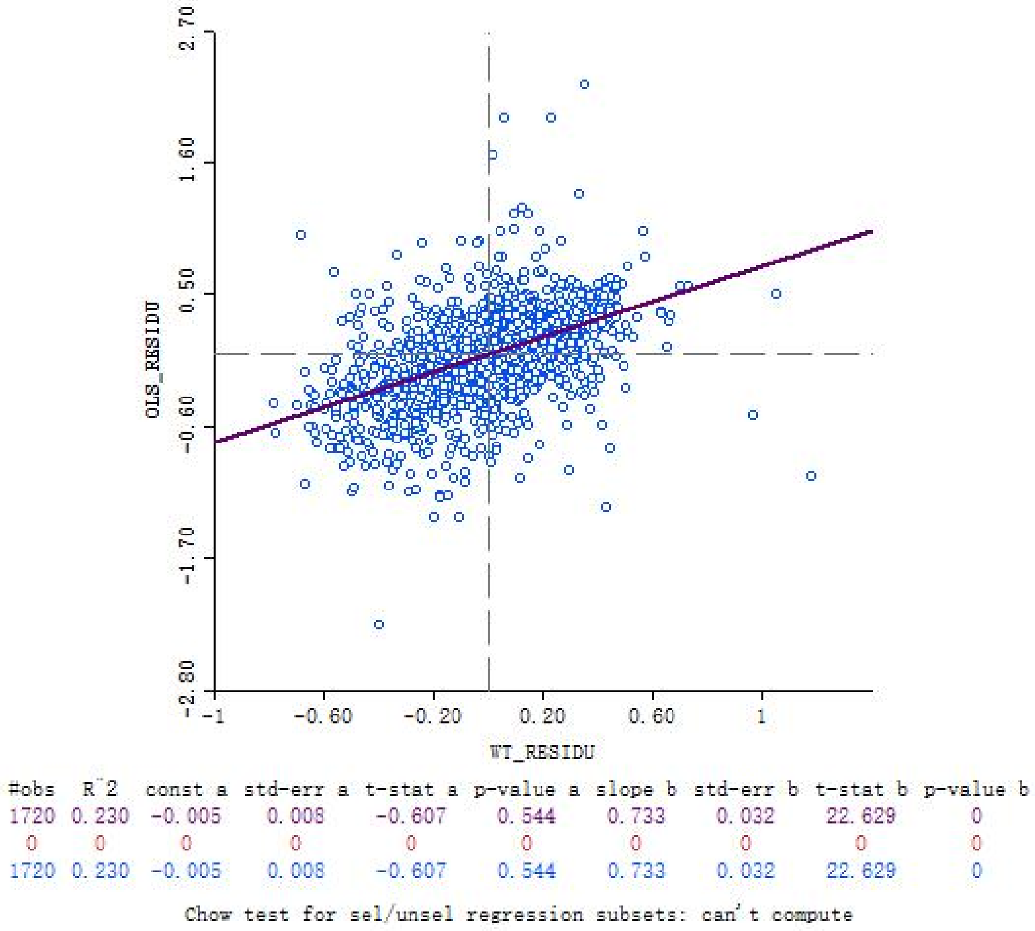
Fig. 4. Significant map for Local Moran’s I

3.2 OLS Regression Results

Table 1. OLS Regression Results

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

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

-----------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT       11.1138      0.0465318       238.843     0.00000
         LNNBELPOV    -0.0789035      0.0084567       -9.3303     0.00000
        PCTBACHMOR     0.0209095    0.000543184       38.4944     0.00000
        PCTSINGLES    0.00297695    0.000703155       4.23371     0.00002
         PCTVACANT    -0.0191563    0.000977851      -19.5902     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   12.990609
TEST ON NORMALITY OF ERRORS
TEST                   DF          VALUE             PROB
Jarque-Bera             2          778.9646          0.00000

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

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : RegressionData
   (row-standardized weights)
TEST                          MI/DF        VALUE          PROB
Moran's I (error)             0.3131       22.3763        0.00000
Lagrange Multiplier (lag)       1         930.5854        0.00000
Robust LM (lag)                 1         441.1036        0.00000
Lagrange Multiplier (error)     1         491.0070        0.00000
Robust LM (error)               1           1.5252        0.21684
Lagrange Multiplier (SARMA)     2         932.1106        0.00000
```

We regressed median household value (*LNMEDHVAL*) on the number of households living in poverty, the percent of individuals with bachelor's degrees or higher, the percent of vacant housing units, and the percent of single house units. About 66% of the variance of the dependent variable *LNMEDHVAL* is explained by the predictors as indicated by the adjusted R-squared. All four predictors are highly significant as they have a p-value less than 0.05. All three results from the Breusch-Pagan test, the Koenker-Bassett test, and the White test have a p-value of close to zero, which means one can reject the null hypothesis that there is no heteroscedasticity. The three tests results are consistent with each other. This indication of high probability of heteroscedasticity is problematic. The normality of the errors are checked by the Jarque-Bera test, which gives a p-value almost equal to zero, meaning one rejects the null hypothesis that the error is distributed normally. Thus, there is a problem with lack of normality.



```
#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

Fig. 5. OLS Residuals vs Respective Weighted Residuals

Figure 5 demonstrates the relationship between OLS residuals and their respective weighted residuals. The weighted residuals are calculated as the average residuals of the queen neighbors of each point. As weighted residual of a point increases, its residual increases. This indicates that there is a positive correlation between the point and their neighbors. The best fit line has a slope value of 0.73 and a p-value almost equal to zero, meaning the correlation between residuals and their neighbors is highly significant. This is an indication of high spatial correlation.

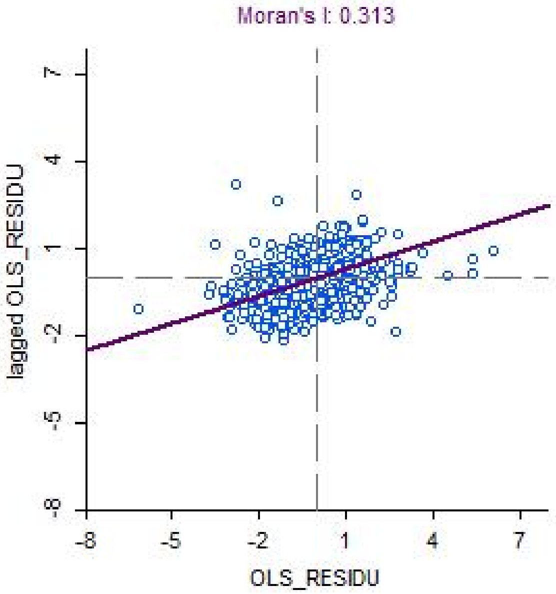


Fig. 6. Moran’s I for OLS residuals of median house values.

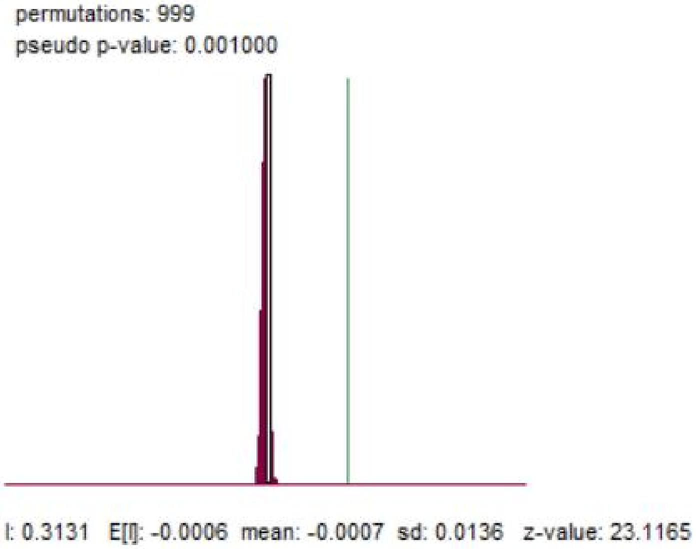


Fig. 7. Histogram of Moran’s I for random permutation test for OLS residuals.

The Moran’s I for the OLS residuals is 0.313, which is quite big, which indicates positive spatial autocorrelation. The pseudo-p-value of the 999 random permutation test for the OLS residuals is 0.001, which is highly significant since it is far less than 0.05. This means that one can reject the null hypothesis that there is no spatial autocorrelation. This is problematic because there are unexplained spatial variances within the residuals, meaning the model does not capture all the variances in the dependent variable.

3.3 Spatial Lag and Spatial Error Regression Results

Table 2. Spatial Lag Regression Results

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : RegressionData
Spatial Weight      : RegressionData
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 : RegressionData
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1       911.8633     0.00000
```

We ran the spatial lag model for *LNMEDHVAL* with the lagged variable (W_*LNMEDHVAL*) and the four original predictors (the number of households living in poverty, the percent of individuals with bachelor's degrees or higher, the percent of vacant housing units, and the percent of single house units). The lagged variable is the weighted dependent variable, which is produced from the Queen weight matrix. The R-squared in the spatial lag model no longer has the same interpretation as in OLS, so here we will not give further interpretation. The p-value of the lagged variable is less than 0.05, so the variable is significant in the model. One would reject the null hypothesis that the median house value in an area is not associated with the median house value in surrounding areas. For the other four original predictors, the p-values of them are all smaller than 0.05, so they are still significant in this model. For both spatial lag model and OLS model, the p-values of the four original predictors are extremely small and close to zero. In terms of the standard errors in the two models, the standard error of the four predictors in the spatial lag model is slightly higher than that in the OLS model, so the sample mean deviates more from the population in the spatial lag model. The Breusch-Pegan test results in a p-value smaller than 0.05, which indicates that one would reject the null hypothesis that there is no problem of heteroscedasticity.

We can use the Akaike Information Criterion (AIC) and Schwartz Criterion (SC) to measure the fitness of the model. Based on Table 3, the AIC and SC values of the OLS model is almost three times that of the spatial lag model. The spatial lag model is a better fit than the OLS model. The Log Likelihood and the Likelihood Ratio Test are the other ways of testing the model fitness. The higher Log Likelihood of spatial lag indicates that it performs better than OLS does. In the Likelihood Ratio Test, the p-value of the spatial lag model is less than 0.05, so one would reject the null hypothesis that the OLS model is better.

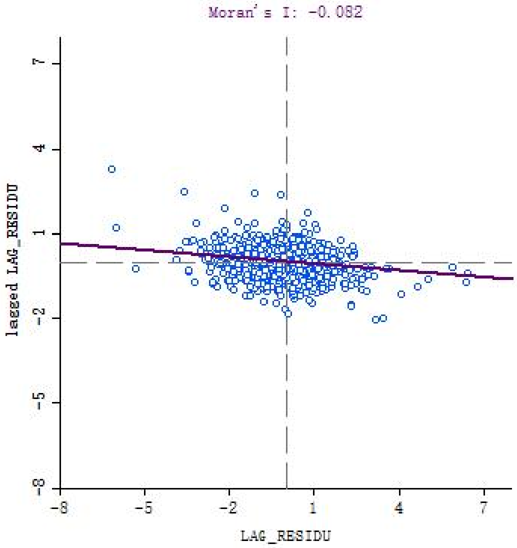


Fig. 8. Moran’s I for Spatial Lag residuals of median house values.

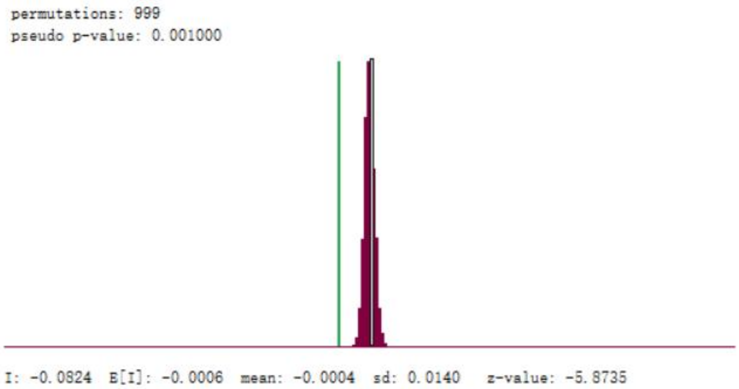


Fig. 9. Histogram of Moran’s I for random permutation test for Spatial Error residuals.

The Moran’s I (Fig 8) for spatial lag model residual is -0.082, which is closer to zero and reflect a negative spatial autocorrelation. The value suggests that the spatial lag model performs better than the OLS model in interpreting spatial variance. In the histogram of 999 random permutation test (Fig 9), both p-values of spatial lag and OLS models are 0.001, which indicates spatial autocorrelation, but both the residual histogram and the Moran’s I of spatial lag model are closer to the expected Moran’s I.

Based on the above comparisons, spatial lag model does better in explaining the *LNMEDHVAL* with other predictors and has better fitness in our data.

Table 3. Spatial Error Regression Results

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : RegressionData
Spatial Weight      : RegressionData
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 : RegressionData
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                   1       678.0016     0.00000
```

The spatial error regression is run for *LNMEDHVAL* with the four predictors. The LAMBDA term has a value of around 0.81 which is positive, indicating a positive correlation between the OLS residual and the spatially lagged residuals. The p-value of the LAMBDA term is almost zero, meaning the correlation between the OLS residuals and their neighboring residuals is statistically significant. By comparing the coefficient of other terms between the spatial error and the OLS models, one can find that all coefficients still remain statistically significant, while the absolute values of the coefficients decrease. This makes sense since spatially lagged residual is used as an additional predictor, which might have taken some effects from other predictors to the dependent variable. Based on the Breusch-Pagan test, which has a resultant p-value less than 0.05, meaning one can still reject the null hypothesis that the residuals are homoscedastic. Thus, there is a problem of heteroscedasticity in this model.

The model fitness can be checked by comparing the AIC/SCs, the Log Likelihood, and the Likelihood Ratio Test of the spatial error and the OLS regressions. The AIC/SC of the spatial error is 754.985 and 782.235, which is about half the value of AIC/SC in the OLS regressions. It means spatial error fits better than OLS. The Log Likelihood is -372.492533 for the spatial error model, which is much higher (less negative) than that of OLS regression, which is around -711. By looking at the Likelihood Ratio Test, the p-value of close to zero means one can reject the null hypothesis of spatial error is not a better model compared to OLS regression. All three criteria demonstrate that the spatial error model fits the data better than the OLS model.

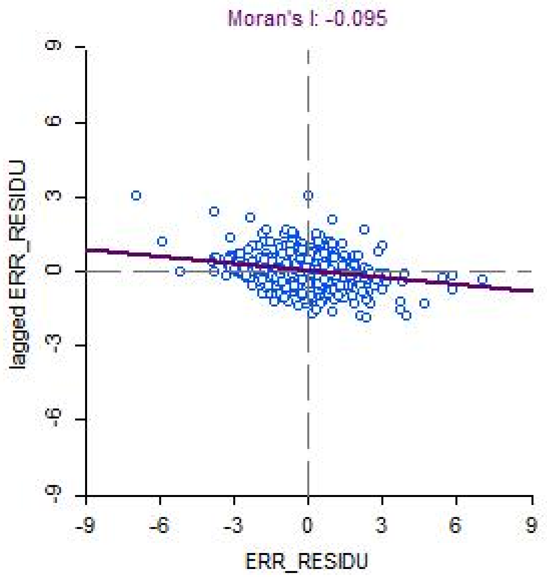


Fig. 10. Moran's I for Spatial Error residuals of median house values.

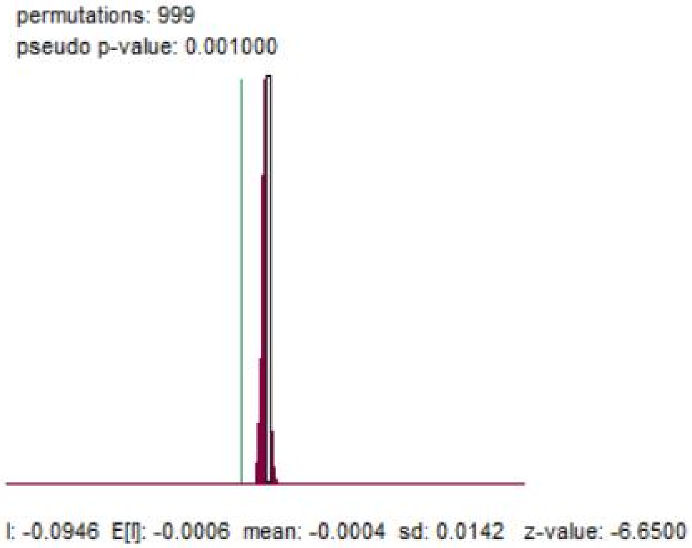


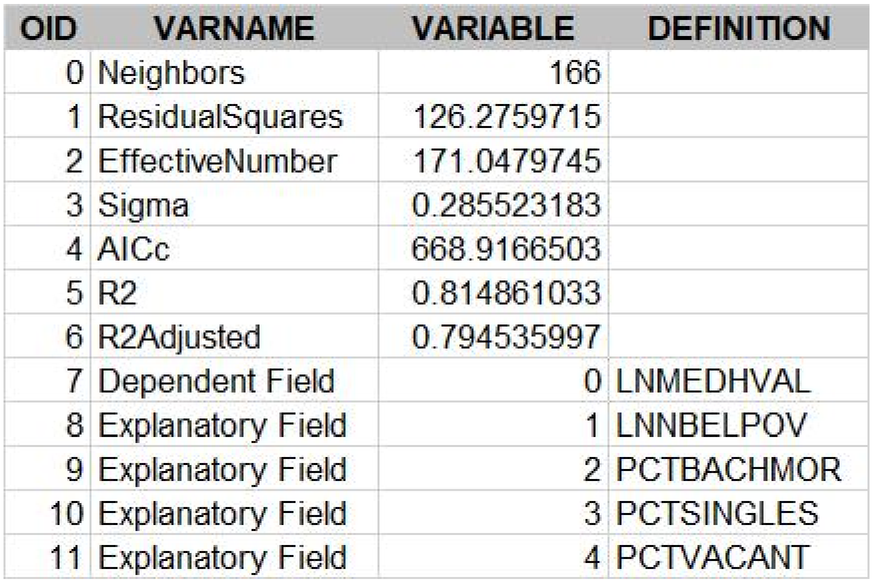
Fig. 11. Histogram of Moran's I for random permutation test for Spatial Error residuals.

The Moran's I (Fig 10) for spatial error model is -0.095, which is much closer to zero compared to the Moran's I for OLS regression. This means the spatial error model does a better job explaining the spatial variances within the dependent variable, *LNMEDHVAL*. By looking at the 999 random permutation histogram (Fig 11), one can see that both the residual histogram and the Moran's I are much closer to the expected Moran's I for random distribution. This also means less spatial autocorrelation is present in the spatial error residuals compared to OLS residuals. However, the p-value is still 0.001, meaning the spatial error residuals are still not random.

In conclusion, the spatial error model is doing a better job than the OLS model, since it fits the data better based on all three criteria and it explains more spatial autocorrelation.

3.4 Geographically Weighted Regression Results

Table 4. GWR Regression Global Results



OID	VARNAME	VARIABLE	DEFINITION
0	Neighbors	166	
1	ResidualSquares	126.2759715	
2	EffectiveNumber	171.0479745	
3	Sigma	0.285523183	
4	AICc	668.9166503	
5	R2	0.814861033	
6	R2Adjusted	0.794535997	
7	Dependent Field	0	LNMEDHVAL
8	Explanatory Field	1	LNNBELPOV
9	Explanatory Field	2	PCTBACHMOR
10	Explanatory Field	3	PCTSINGLES
11	Explanatory Field	4	PCTVACANT

The GWR has an R-squared value of 0.81 and an adjusted R-squared of 0.79, which are all higher than the R-squared value of OLS regression (around 0.66). This means that GWR is doing a better job by explaining 13% more variance within the dependent variable compared to OLS regression. The AIC of GWR is around 669, while the AICs for spatial lag, spatial error, and OLS are approximately 523, 755, and 1433. In this case, GWR is a better fit model that spatial error and OLS regression, but it is not as good as the spatial lag model.

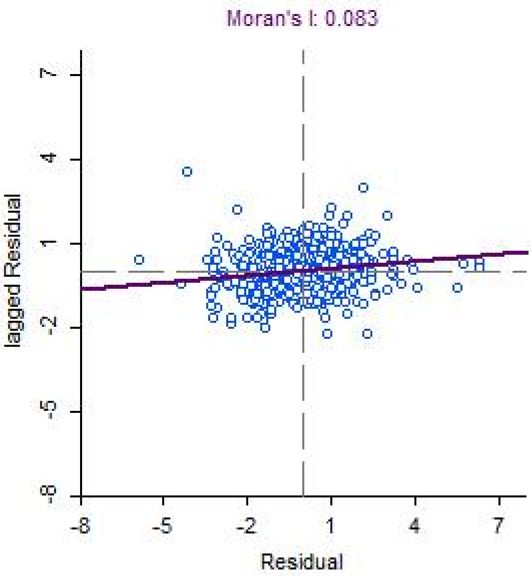


Fig. 12. Moran’s I for GWR residuals of median house values.

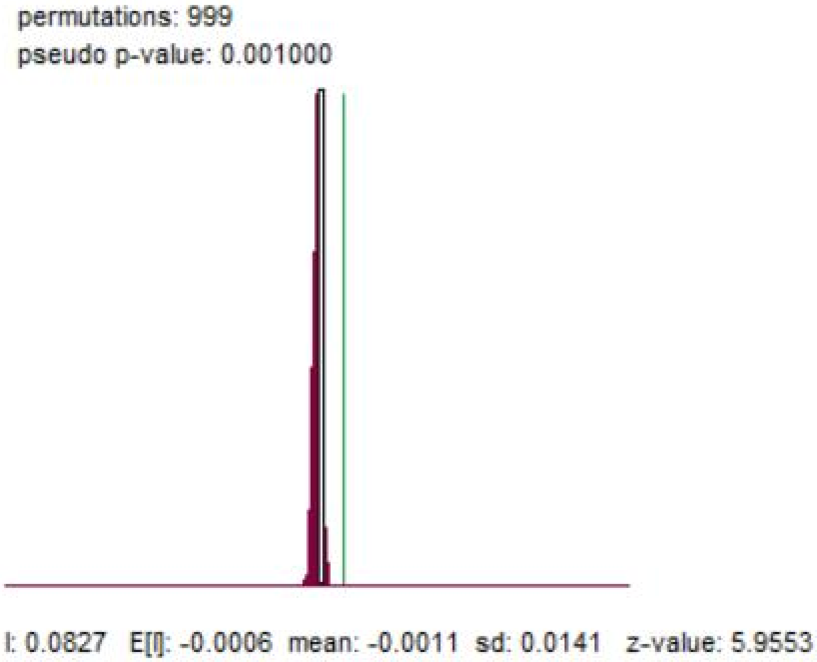


Fig. 13. Histogram of Moran’s I for random permutation test for GWR residuals.

The Moran’s I (Fig 12) for the GWR model is 0.083, which is much closer to zero compared to the Moran’s I for OLS regression of 0.313. This means the GWR model does a better job explaining the spatial variances within the dependent variable, *LNMEDHVAL*. The Moran’s Is for spatial lag and error residuals are -0.082 and -0.095. The Moran’s I for the spatial lag model is the closest to the expected value of -0.0006, with the Moran’s I for GWR residuals ranking the second, the spatial error ranking the third, and the OLS regression the last. Therefore, the spatial lag model has the least spatial autocorrelation in residuals, and GWR does the second-best. By looking at the 999 random permutation histogram (Fig 13), one can see that both the GWR residual histograms and the Moran's I value are much closer to the expected Moran's I for random distribution. This also means less spatial autocorrelation is present in the GWR residuals compared to OLS residuals. However, the p-value is still 0.001, meaning the GWR residuals are still not random.

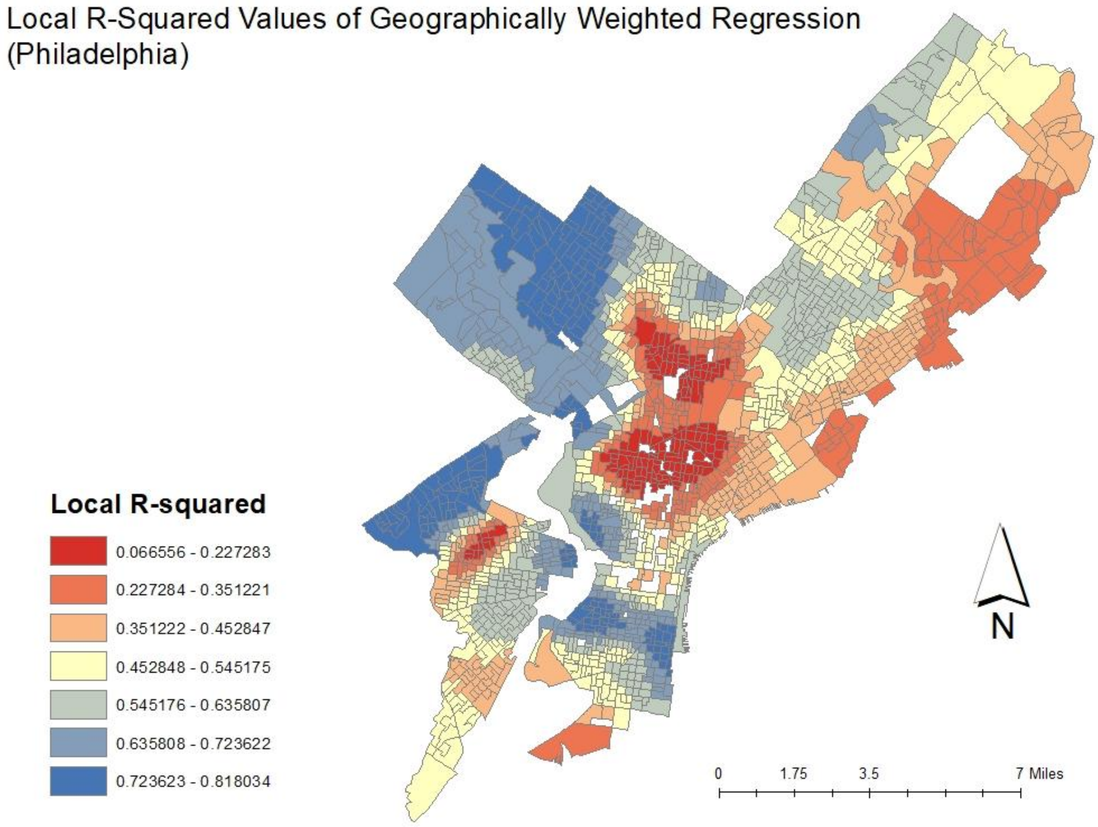
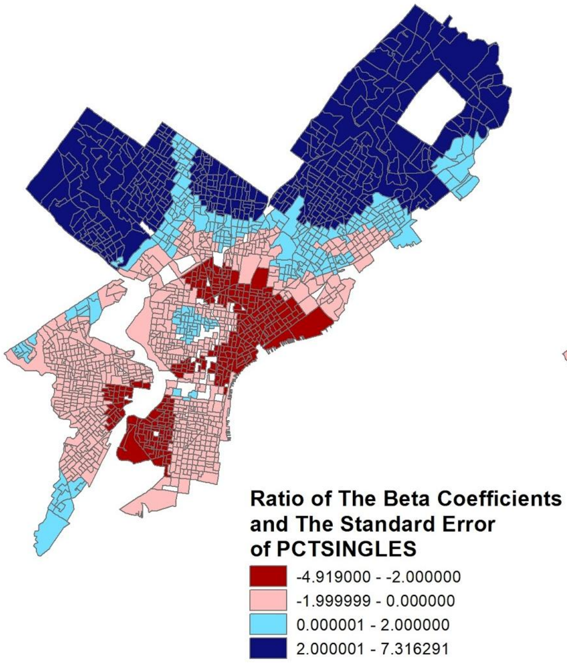
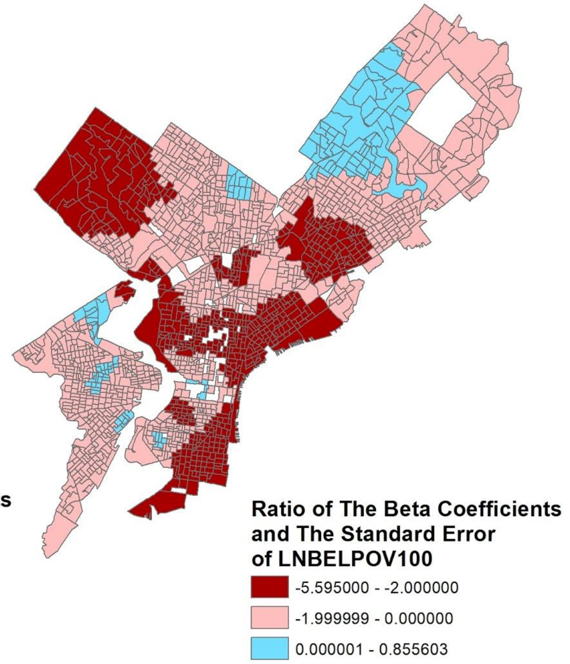


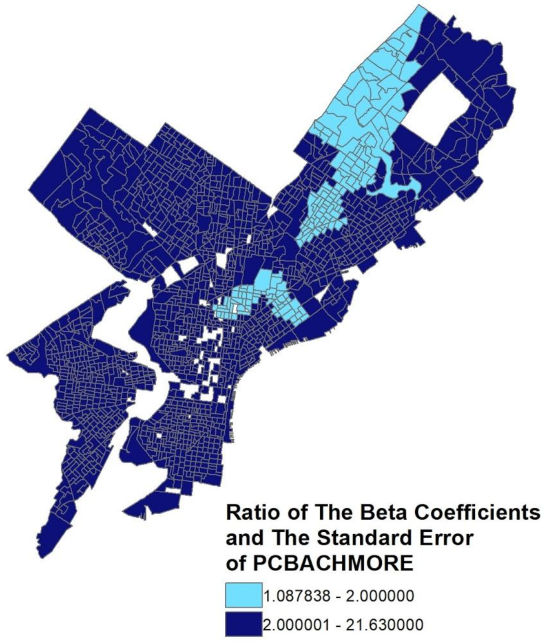
Fig. 14. Local R-Squared Values of GWR Model

Figure 14 shows the local R-squared values presented spatially. North Philadelphia and a small portion of the West Philadelphia have the lowest local R-squared, ranging from 0.067 to 0.23, which means the four chosen predictors explain particularly poorly for these regions, and only 6% to 23% of the variance in the median house value is explained. The south part of the Northeast Philadelphia also has a relatively low R-squared. In contrast, the Northwest Philadelphia and South Philadelphia have very high local R-squared value, ranging from 0.72 to

0.82, which means 72% to 82% of the variance within the median home value is explained by the predictors.




Ratio of The Beta Coefficients and The Standard Error of LNBELPOV100
-5.595000 - -2.000000
-1.999999 - 0.000000
0.000001 - 0.855603


Ratio of The Beta Coefficients and The Standard Error of PCBACHMORE
1.087838 - 2.000000
2.000001 - 21.630000

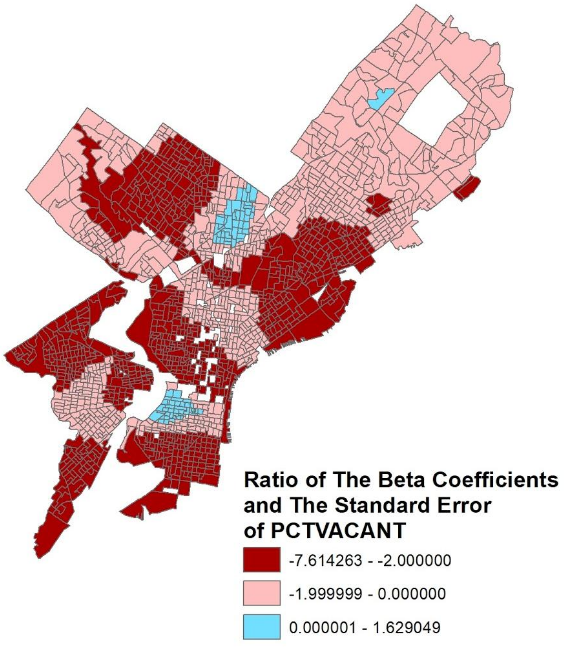

Ratio of The Beta Coefficients and The Standard Error of PCTVACANT
-7.614263 - -2.000000
-1.999999 - 0.000000
0.000001 - 1.629049

Fig. 15. GWR Local Regression Results

Figure 15 shows the local regression results from GWR. The beta coefficients of the percentage of single housings have a balanced range, since there are both very high coefficients and very negative coefficients. In the Northeast and Northwest Philadelphia, this predictor has a stronger positive influence on the housing values, while in Kensington, Bridesburg, Richmond, and the west side of South Philadelphia, the predictor has stronger negative influence on the housing values. The predictors of number of households below poverty and the percentage of vacant housing influence the housing values in mostly a negative way. In contrast, the predictor of percentage of bachelor degrees or higher influence the housing values in mostly a positive way.

4. Discussion

This project we used spatial lag, spatial error and geographically weighted regression (GWR) to predict the median house values in Philadelphia block groups. We assumed these models would perform better than the OLS model do due to the account for the spatial autocorrelations. Based on the above analysis, we conclude that the OLS model is not an appropriate choice to analyze the data with spatial context. For the project on predicting median house values, the spatial lag model is the regression method with the best fitness.

The three methods provide more sophisticated explanations to our prediction, but there are still multiple limitations associated with the models. The results of Breusch-Pagan tests and other similar tests for the spatial lag and spatial error models reject the null hypothesis. Said differently, our assumption on homoscedasticity was violated. In addition, there still exists spatial autocorrelation associated with the residuals for all three models since the p-values are still less than 0.05. The residuals are not random. Thus, the three models still cannot fully account for the spatial autocorrelation.