Homework 2

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

Yuhao Jia, Zhonghua Yang, Zile Wu

1. Introduction

As an important part of the social economy, housing prices reflect the development of society. Therefore, forecasts of housing prices can help people better judge the economy.

In the previous report, we predicted median values for all owner-occupied housing units using OLS regression models based on the proportion of residents in the neighborhood group with at least a bachelor's degree; the proportion of vacant housing units; the proportion of detached single-family housing units and the number of households with incomes below the 100% poverty level as predictors.

However, the OLS analysis is often inappropriate when dealing with dataset that have a spatial component. In this report, we will use spatial lag, spatial error and geographically weighted regression to see whether these methods perform better than OLS.

2. Methods

a) A Description of the Concept of Spatial Autocorrelation

Waldo Tobler proposed the first law of Geography: “Everything is related to everything else, but near things are more related than distant things.”

Spatial autocorrelation describes the presence of systematic spatial variation in a variable, and Moran’s I is the most widely used method of testing for spatial autocorrelation. The value of Moran’s I is calculated as below:

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

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

In this formula, $\bar{X}$ is the mean of the variable X. X_i is the variable value at a particular location i. X_j is the variable value at another location j. w_{ij} is a weight indexing location of i relative to j. n is the number of observations (points or areal units).

The larger positive this value is (close to 1), the stronger positive autocorrelation it indicates. The larger negative this value is (close to -1), the stronger negative autocorrelation it indicates. If this value is around 0, it indicates that there is no spatial autocorrelation (random pattern). Moran's I is most of times but not always between -1 and 1.

Queen neighbors of a polygon are those intersects it either at a point (common vertex) or a segment (common edge). In this report, Queen Neighbor Matrix will be used as the weight matrix to calculate the spatial autocorrelation and run the spatial regressions.

As different spatial weight matrices have their own limitations and fit for different situation, statisticians usually try several different weight matrices to make sure their results are not merely an artifact of the matrix they are using, unless they have strong theoretical motivations to not do so.

We will calculate the Moran's I for a variable and test whether it is significant with 999 permutes using GeoDa. The null hypothesis H_0. is no spatial autocorrelation (random pattern). One alternative hypothesis H_{a1} is positive spatial autocorrelation and the other one H_{a2} is negative spatial autocorrelation.

To do so, the values of the variable will be randomly shuffled (permuted) 999 times and calculate Moran's I for each permutation to see the value distribution of no spatial autocorrelation. Then, the place original Moran's I (for the observed variable) stands will be compared to the Moran's I values for the random permutations. If the Pseudo P-value (Chance of getting a value as large as the observed value using samples from this distribution) is less than 0.05, the original Moran's I is statistically significantly different from 0, which also means we can reject the H_0 of no spatial auto correlation.

Before the Moran's I is calculated for the whole region (global spatial autocorrelation), we also need to consider spatial autocorrelation in smaller area. Local spatial autocorrelation is the presence of systematic spatial variation in a variable only focusing on the relationships between each observation and its surroundings. In other words, it's the relationship between the variable value in each location and the values in its neighbors.

b) A Review of OLS regression and Assumptions

OLS regression is a statistical method used to examine the relationship between a variable of interest (y) and one or more explanatory variables (x). These two kinds of variables can be related to each other in a deterministic way or non-deterministic way. OLS regression with one predictor is called simple regression, and regression with two or more predictors is called multiple regression.

For simple regression, there are several model assumptions: Linear relationship between y and x; Residuals are normally distributed; Residuals are random; Residuals are homoscedastic; Independence of observations and residuals; y is continuous and preferably normal.

For multiple regression, there are same assumptions as for simple regression and one more: the predictors (x) should not be very strongly correlated with each other.

Please refer to *Using OLS Regression to Predict Median House Values in Philadelphia* for more information about OLS.

When there is a spatial component in the data, the assumption of random errors is usually violated.

We can test this assumption by examining the spatial autocorrelation of the residuals using Moran's I.

Another way to test spatial autocorrelation of OLS residuals is to regress them on nearby residuals (residuals at neighboring block groups as defined by the Queen matrix). rho (ρ) is the parameter used to judge the autocorrelation. It is calculated as the coefficient of nearby residuals in the regression of OLS residuals and their nearby residuals, also known as lambda (λ) in GeoDa.

In this research, OLS regression will be run in GeoDa, where there is also a way of testing other regression.

First, we need to check the assumption of homoscedasticity. To do so, we will test the heteroscedasticity for OLS residuals in GeoDa using the White Test. The null hypothesis H_0 here is that there is homoscedasticity (no heteroscedasticity). The alternative hypothesis H_a is that there is heteroscedasticity. If the p-value is less than 0.05, then we can reject the null hypothesis for the alternate hypothesis of heteroscedasticity.

We also need to check the assumption of normality of errors. We will use Jarque-Bera test in GeoDa to test it. The null hypothesis H_0 here is that the residuals are from a normal distribution. The alternative hypothesis H_a is that the residuals are not from a normal distribution (non-normality). If p<0.05, reject the Null Hypothesis of normality for the alternative hypothesis of non-normality.

c) Spatial Lag and Spatial Error Regression

In this report, we will be using GeoDa for running spatial lag and spatial error regressions.

Spatial lag regression model assumes that the value of the dependent variable at one location is associated with the values of that variable in nearby locations (defined by Queen weights matrix in this report), which means the model includes the spatial lag of the dependent variable as a predictor.

The spatial lag model is run as following equation in this research:

$$LNMEDHVAL = \rho Wy + \beta_0 + \beta_1 PCTVACANT + \beta_2 PCTSINGLES \\ + \beta_3 PCTBACHMOR + \beta_4 LNBELPOV + \varepsilon$$

Here, ρ is the coefficient of the spatial lag variable Wy; β_0 is the intercept; β_1 is the coefficient of the variable $PCTVACANT$; β_2 is the coefficient of the variable $PCTSINGLES$; β_3 is the coefficient of the variable $PCTBACHMOR$; β_4 is the coefficient of the variable $LNBELPOV$; ε is the residual.

Spatial error regression model assumes that the residual at one location is associated with residuals at nearby locations. There are two steps: run OLS regression and regress residuals on the nearest neighbor residuals.

The spatial error model is run as following equation in this research:

$$LNMEDHVAL = \beta_1 PCTVACANT + \beta_2 PCTSINGLES + \beta_3 PCTBACHMOR + \beta_4 LNBELPOV + \lambda W_\varepsilon + u$$

Here, β_0 is the intercept; β_1 is the coefficient of the variable $PCTVACANT$; β_2 is the coefficient of the variable $PCTSINGLES$; β_3 is the coefficient of the variable $PCTBACHMOR$; β_4 is the coefficient of the variable $LNBELPOV$; λ is the coefficient of the spatially lagged residuals W_ε; u is the random noise.

Most assumptions that are needed for OLS are still needed for both spatial lag and spatial error regression models: each predictor is linearly related with the dependent variable; The residuals are normal; There should not be multicollinearity. Only spatial independence of observations is not needed.

The goal of spatial lag and spatial error regression is to reduce the influence of spatial components on the model and thus improve the predictive power of the model. We hope that after using these methods, the regression residuals will not show spatial autocorrelation anymore.

We will compare the results of spatial lag regression with OLS and the results of spatial error regression with OLS, then we will decide whether the spatial models perform better than OLS based several criteria including Akaike Information Criterion (AIC)/Schwarz Criterion (SC), Log Likelihood, and Likelihood Ratio Test.

The AIC and SC are measures of the goodness of fit of the model. They are relative measures of the information lost, which describes the tradeoff between precision and complexity of the model. The lower the AIC and SC, the better the fit of this model.

The Log Likelihood associated with the maximum likelihood method of fitting a statistical model to the data and estimating model parameters. The higher the log likelihood, the better the fit of this model. It should only be used for comparing nested models, which means we can use it to compare OLS model with spatial lag model or with spatial error model, but we can not use it to compare spatial lag model with spatial error model.

The Likelihood Ratio Test compares the OLS model with the spatial model. The null hypothesis H_0 here is spatial lag (or spatial error) model is not a better specification than the OLS model. The alternative hypothesis H_a is that spatial lag (or spatial error) model is a better specification than the OLS model. If p value is less than 0.05, then we can reject the null hypothesis and state that the spatial model is better than the OLS model.

Another way of comparing OLS results with spatial lag and spatial error results is by looking at Moran’s I of regression residuals. With significances of residuals’ Moran's I for all three models, the closer to 0 the Moran’s I is, the better the model is.

d) Geographically Weighted Regression

In this report, we will do geographically weighted regression analyses in ArcGIS. We need to conduct spatial regression for different geographical observations so as to explore the impact of prediction variables on the results of the dependent variable for different spatial locations.

Compared with spatial lag regression and spatial error regression, the major difference of geographical weighted regression is the non-stationarity of geographical spatial relations. Spatial stationarity is the assumption that in the model, all correlations are invariant in the global space, and the relationship between the predictors and dependent variable at any specified location is the same. But in practice it is not always correct.

Local regression is very useful in practice, because in many cases, it is not a single, global regression. The correlation between variables is closely related to geographical location and environmental factors. We need to conduct separate local regression for different locations, such as spots, area, tract, etc. which are with clear geographic information.

Simpson's Paradox explains this difference and visualizes it. Plots below are two scatter plots showing the relationship between variables and contains the fitting results after regression. However, the left uses global regression, and the right is local regression. As you can see, the scatter on the left show different clusters depending on the median home price, but the regression result shows that burglary and median home price are negatively correlated. In the figure on the right, the distribution of scatter points is the same as that on the left, but these data are decomposed and different regressions are carried out according to the regions, we will find that there are completely different correlation relationships between variables in different regions.

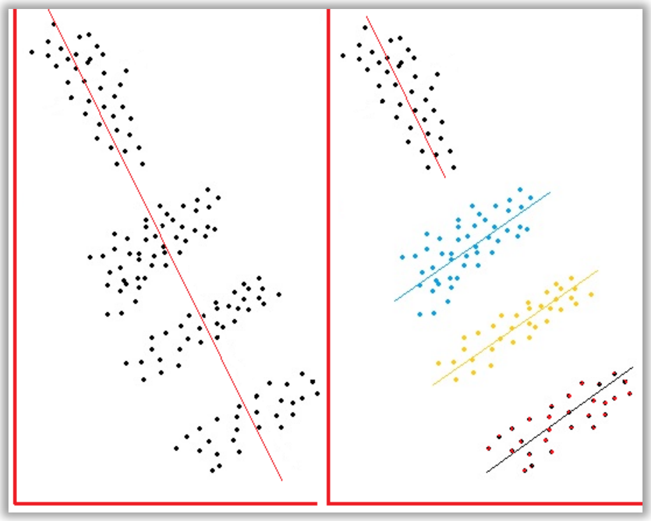


Figure1

Formula below shows the equation for GWR. In the formula, i represents an observed value, $i = 1 \dots n$, this formula shows the dependent variable y and multiple predictive variables x_k, $(k = 1 \dots m)$, and the formula is based on the geographical location of the observed value i, and the relationship is specific to that location.

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

The operation of local regression requires multiple observations and locations. If there is only one observation value in the model, local regression cannot be run. The GWR data set contains many observations, each of which will participate in the regression. However, each observation value plays a different role in the regression according to different geographical locations, which requires that different weights should be assigned to each observation value, and the closer the observation value is to the location i, the higher the weight will be.

The weight is calculated based on the distance between the observed value and the location. There are two ways to calculate the weight: fixed bandwidth and adaptive bandwidth. Fixed bandwidth means that the distance or area around an observation i remains constant regardless of the number of observations. The equation can be expressed as formula below. $distance_{ij}$ refers to the distance between regression point i and data point j. h refers to the fixed bandwidth.

$$w_{ij} = \begin{cases} e^{-0.5\left(\frac{distance_{ij}}{h}\right)^2}, & if\ distance_{ij} \le h \\ 0, & otherwise \end{cases}$$

Unlike fixed bandwidth, although the number of observed values in the model does not change, the area of the weight will be different. In the formula below, $distance_{ij}$ refers to the distance between regression point i and data point j. The h for adaptive bandwidth can be varied, and the value of this h has different values for different observed values. For example, we stipulate that each observation value should contain 10 nearest values. Observation value 1 may need h= 5000 to obtain 10 nearest values, but observation value 2 only needs h=2000 to meet the requirement.

$$w_{ij} = \begin{cases} \left[1 - \left(\frac{distance_{ij}}{h}\right)^2\right]^2, & if\ j\ is\ one\ of\ i's\ N\ nearest\ neighbors \\ 0, & otherwise \end{cases}$$

The most important part of GWR is the calculation of weights, because different weights have completely different effects on the results during the regression process. In the calculation process, the weight calculation is also the choice of bandwidth. In GWR, either fixed bandwidth or adaptive bandwidth can be used, so next we need to discuss whether to use fixed or adaptive bandwidth to calculate the weight.

The fixed bandwidth is more suitable for the stable data distribution, and the quantity distribution and aggregation degree of the observed values are very uniform. The adaptive bandwidth is more suitable for considering the geographical spatial variation, that is, the observed values are clustered or polygons are heterogeneously shaped or sized. Therefore, for GWR, we need to conduct local regression for different spatial locations, and adopt different bandwidths for different local regressions, which is also the characteristics of adaptive bandwidth.

OLS needs to be run before running geographically weighted regression to ensure the rationality of the model, because it is difficult to check the linearity of the data relationships. This is the same as when we are using spatial lag regression and spatial error regression, the OLS model should be run first. If the OLS result is reasonable, GWR can continue. We know that many of our assumptions in OLS are still true in GWR, such as residual normality, homoscedasticity, and no multicollinearity. However, GWR requires an index called AIC, which is used to measure model performance and help compare different regression models. The smaller AIC is, the better the model can fit the observed data. However, when the amount of data is small, AIC is more likely to select models with too many parameters, and small sample size needs to be corrected, which is called AICc. Neither AIC nor AICc are absolute measures of goodness of fit, but as long as they apply to the same dependent variable, they are useful for comparing models with different explanatory variables.

We know that when we run a global regression model (such as OLS), if there is multicollinearity between two or more variables, the results of the regression model are not reliable. Multicollinearity means that a variable is redundant in the model, which has almost the same effect as a collinear variable, which can have a negative effect on the production of results. GWR will construct local regression equations for each factor in the data set. When the values of the prediction variables are clustered in a substantial way, you will very likely have multicollinearity problems. This actually shows that the prediction variables play the same role in the position of each factor, and there is not enough variability in the model. Similarly, if more than two variables in the model have similar clustering patterns in the local region, the model will also encounter the multicollinearity problem. For example, multicollinearity problems can occur if the values of two variables in the data set are both very high or very low at a certain location. When we run the GWR model in the algorithm, we get a property sheet with a number of conditions by which we can determine whether the model is unstable due to multicollinearity. In general, you cannot trust conditions that are large (greater than 30), equal to null, or small.

In GWR, multicollinearity should be avoided as much as possible. Sometimes, some problems need to be paid attention to in data processing. The inclusion of classified data in the GWR model should be used with caution, as it may lead to unnecessary spatial clustering of prediction variables, such clustering problems may lead to multicollinearity. For example, assigning a value of 1 to an area outside the central city and a value of 0 to a central city area will result in unnecessary data aggregation. In fact, GWR allows coefficients of explanatory variables to vary, so it is not necessary to reassign features to spatially categorized explanatory variables. If any clustering is present in the dummy variable, then there may not be any variability in the predictor for position i, which is not meaningful for the regression model.

It is worth noting that p-values are not included in GWR's output. In the global regression model, p-value is usually used to test whether the estimated value of the parameter is significantly different from the null hypothesis, and T-test is the process of calculating p-values. However, for GWR, the calculation methods of local regression and global regression test are quite different. Because for GWR, each regression point will have a set of parameters, as well as a set of standard deviations, and there may be hundreds of regression points during the regression process, it may be necessary to have thousands of tests to determine whether the parameter is significant in the local regression. According to type I error, if α=0.05 is used as the significance level, it means that we would predict that 5 out of 100 results would be significant, but in fact this is not correct. If four variables prediction model has an estimated 2000 regression results, so will need 10000 significance tests, each regression results will contain an intercept and four predictor variable inspection, for type I error, we would expect there are 500 test the returned result is remarkable, but it only returns the results by accident, that they are not actually significant in reality. Therefore, the significance test results in local regression cannot be represented by p-values.

3. Results

a) Spatial Autocorrelation

Below are the results of the global Moran's I. The green vertical line represents the Moran's I for the data is 0.79 with Queen Matrix. The red part represents the Moran's I for all 999 permutations, the Pseudo P-value = 0.001 < 0.05 which means there is a significant spatial autocorrelation in the dependent variable, LNMEDHVAL.

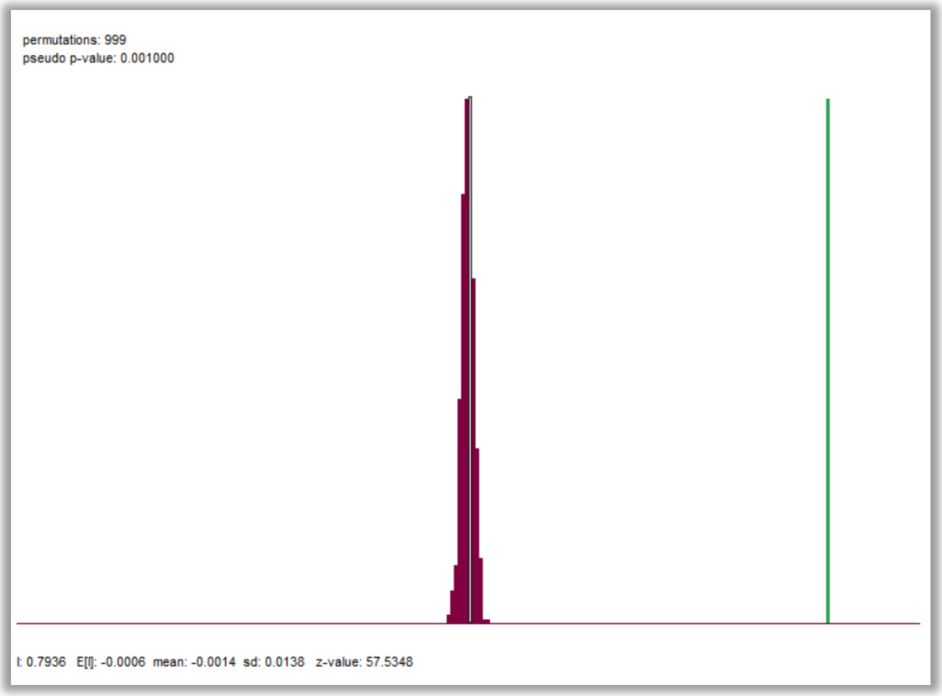


Figure 2

Based on the global Moran' I values, we understand that there is spatial autocorrelation in LNMEDHVAL. For further study, we use LISA (Local Indices of Spatial Autocorrelation) to measure which locations have significant spatial autocorrelation. The following map filled by different shades of green is the LISA Significance Map, the darker the color, the more significant spatial autocorrelation exists in that location, and the white part indicates that there is no significant spatial autocorrelation at that place. The map filled by red and blue is LISA Cluster Map, and the dark red part indicates that the LNMEDHVAL around the census tract is high and the LNMEDHVAL at the location is also high. The dark blue part indicates that the LNMEDHVAL around the census tract is low while the LNMEDHVAL at that location is also low. The light red part indicates that the LNMEDHVAL around the census tract is high while the LNMEDHVAL at this location is low. The light blue part indicates that the LNMEDHVAL around this census tract is low while the LNMEDHVAL at this location is high.

From the graph, we can see that the areas in Northwest and Northeast Philadelphia appear to have a concentration of high housing prices to live in, and the areas in North Philadelphia and West Philadelphia mantua areas appear to have a concentration of low housing prices to congregate, which is consistent with the actual situation.

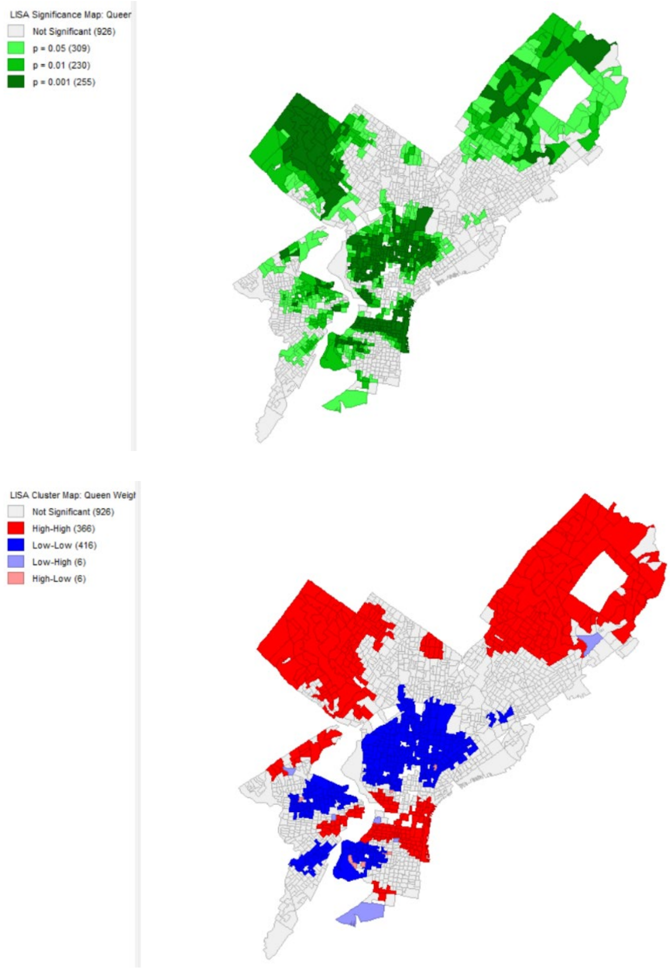
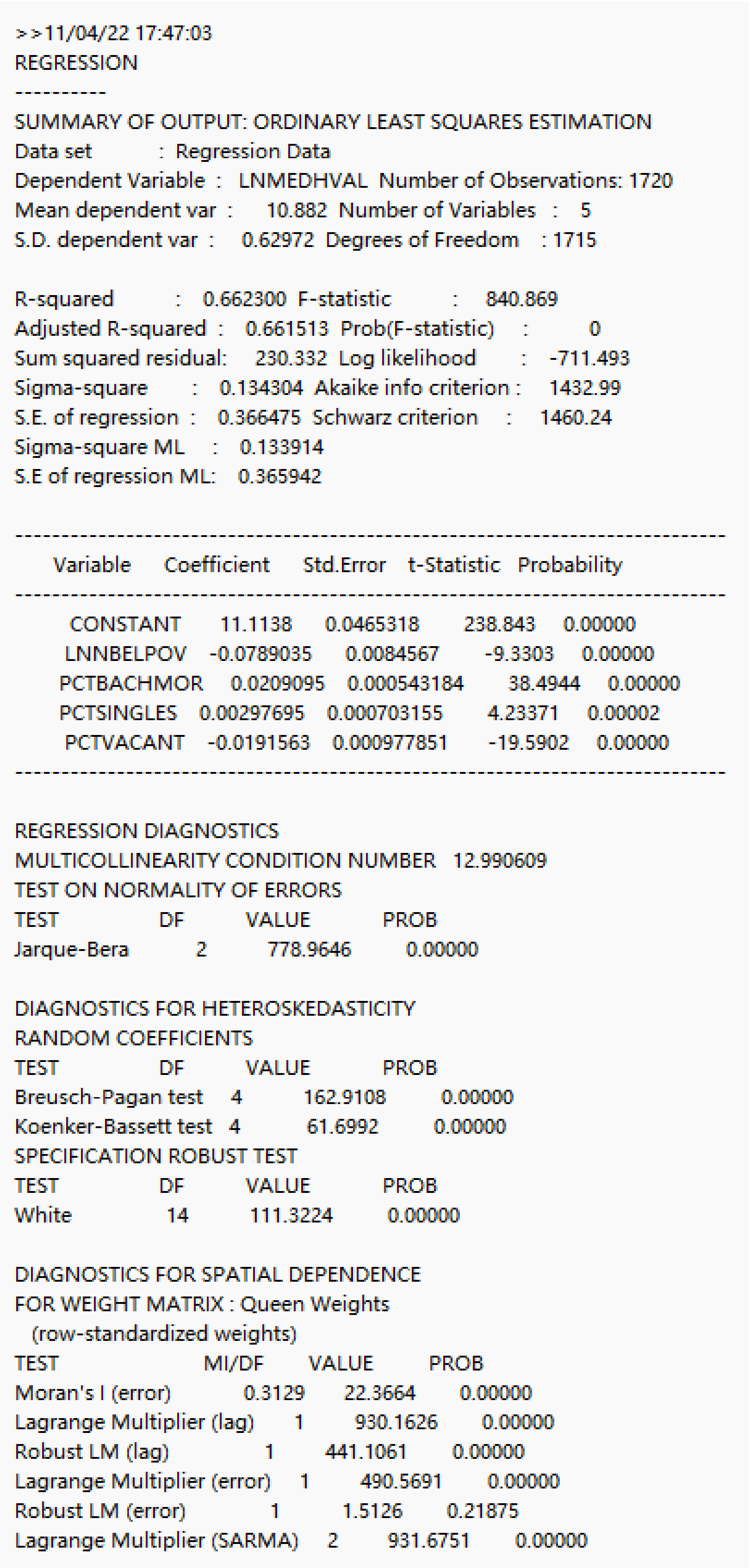


Figure 3

b) A review of OLS Regression and Assumptions: Results



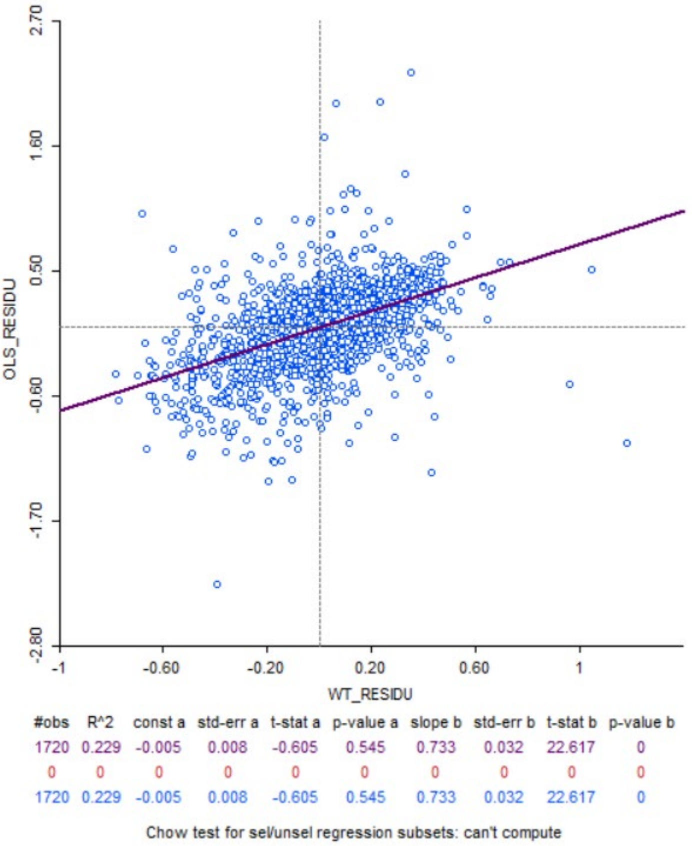
```
>>11/04/22 17:47:03
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  Regression Data
Dependent Variable  :  LNMEDHVAL  Number of Observations: 1720
Mean dependent var  :     10.882  Number of Variables   :    5
S.D. dependent var  :    0.62972  Degrees of Freedom    : 1715

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

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

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

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

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : Queen Weights
   (row-standardized weights)
TEST                          MI/DF        VALUE          PROB
Moran's I (error)            0.3129       22.3664        0.00000
Lagrange Multiplier (lag)       1        930.1626        0.00000
Robust LM (lag)                 1        441.1061        0.00000
Lagrange Multiplier (error)     1        490.5691        0.00000
Robust LM (error)               1          1.5126        0.21875
Lagrange Multiplier (SARMA)     2        931.6751        0.00000
```

Table 1

The OLS regression output shows that PCTSINGLES and PCTBACHMOR are highly significant and positively associated with LNMEDHVAL, while PCTVACANT and LNNBELPOV100 are highly significant and negatively associated with LNMEDHVAL ($p<0.05$ for all four variables). More than 60% of variance in LNMDEHVAL has been explain by the model (R^2 and Adjusted R^2 are 0.6623 and 0.6615). The p value with F-statistic 841 is less than 0.05, so we can reject the H_0 that all β coefficients are 0.

From the OLS results calculated by GeoDa, we can see three different heteroskedasticity diagnostics, The Breusch-Pagan Test, The Koenker-Bassett Test, and The White Test, From the GeoDa results we can see that the p-values of all three diagnostics are less than 0.05, so there is a problem with heteroscedasticity.

Also, the Jarque-Bera test can be used to check the assumption of normality of errors in GeoDa. From table 1, we can see that p value of the Jarque-Bera test is less than 0.05, which indicating that the residuals are not from a normal distribution(non-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  0.229  -0.005    0.008     -0.605     0.545     0.733     0.032     22.617       0
   0      0       0        0          0         0         0         0          0       0
1720  0.229  -0.005    0.008     -0.605     0.545     0.733     0.032     22.617       0

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

Figure 4

To examine the spatial autocorrelation of the OLS Regression, the scatterplot of OLS_RESIDU (OLS residuals) by WT_RESIDU (residuals at neighboring block groups as defined by the Queen matrix) are presented above. We can see from the picture that the value of ρ (referred to as Slope b in the results) is 0.733. That indicates a significant spatial autocorrelation in the OLS Regression model.

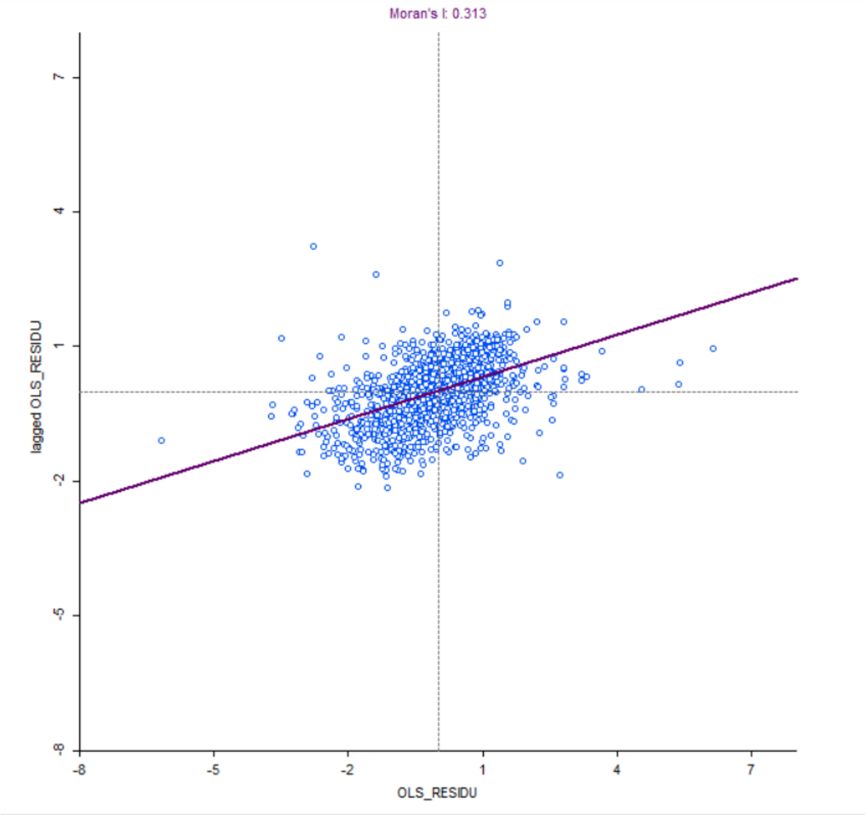
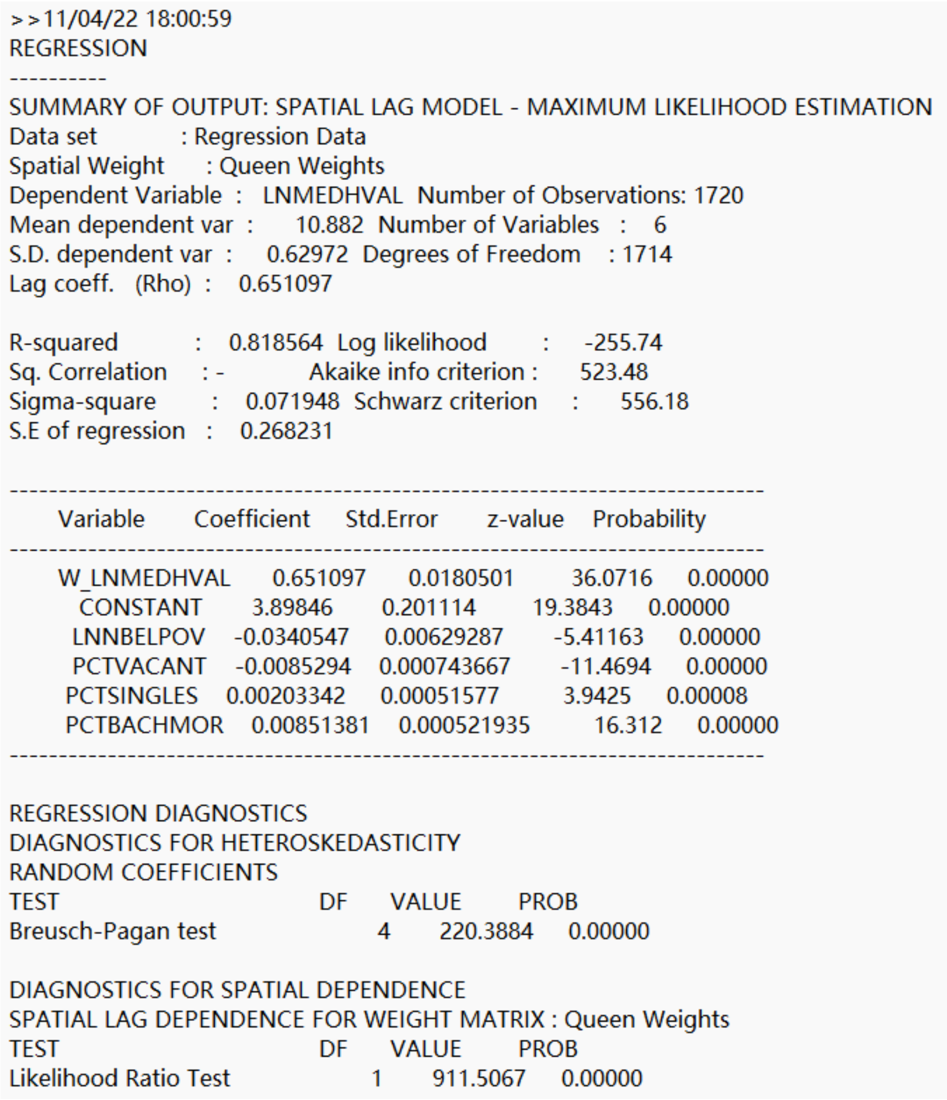


Figure 5

To further examine spatial autocorrelation, the Moran's I scatterplot and results from the 999 permutations for OLS regression residuals are plotted as follows. From the graphs, we can see a significant spatial autocorrelation in this OLS residuals because the p value is less than 0.05. It is problematic and we will attempt to account for that in the following practices in spatial model regressions.

c) Spatial Lag and Spatial Error Regression Results



```
>>11/04/22 18:00:59
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : Regression Data
Spatial Weight      : Queen Weights
Dependent Variable  :  LNMEDHVAL  Number of Observations: 1720
Mean dependent var  :     10.882  Number of Variables   :    6
S.D. dependent var  :    0.62972  Degrees of Freedom    : 1714
Lag coeff.   (Rho)  :   0.651097

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

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
     W_LNMEDHVAL       0.651097      0.0180501        36.0716     0.00000
        CONSTANT        3.89846       0.201114        19.3843     0.00000
       LNNBELPOV     -0.0340547     0.00629287       -5.41163     0.00000
       PCTVACANT     -0.0085294    0.000743667       -11.4694     0.00000
      PCTSINGLES     0.00203342     0.00051577         3.9425     0.00008
      PCTBACHMOR     0.00851381    0.000521935         16.312     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                       4      220.3884     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : Queen Weights
TEST                                    DF      VALUE        PROB
Likelihood Ratio Test                    1      911.5067     0.00000
```

Table 2

The spatial lag regression output from GeoDa is presented above. From Table 2, the coefficient of the spatial lag ρ =0.65, and it is significant ($p < 0.05$). This indicates that a one unit increase of the W_LNMEDHVAL is associated with a 0.65 increase in LNMEDHVAL at one location.

Further, we can see that all remaining predictors (PCTBACHMOR, PCTVACANT, PCTSINGLES, AND LNNBELPOV100) are statistically significant ($p<0.05$ for all variables).

The Breusch-Pagan test, as observed in the GeoDa results summary, indicates that there is an issue with heteroscedastic residuals ($p < 0.05$). There is still a problem with heteroscedasticity.

Based on the Table2 and Table1, we can compare the Spatial Lag regression and OLS regression models based on the Akaike Information Criterion (AIC) /Schwarz Criterion (SC), the Log Likelihood, and the Likelihood Ratio Test.

For the Spatial Lag regression, AIC = 523.48, SC = 556.18. For the OLS regression, AIC = 1432.99, SC = 1460.24. Because the lower the AIC and SC, the better the fits of this model. The results indicates that the Spatial Lag model is a better fit than OLS.

For the Spatial Lag regression, Log Likelihood = -255.74. For the OLS regression, Log Likelihood = -711.493. Because the higher the log likelihood, the better the fit of this model. The results indicates that the Spatial Lag model is a better fit than OLS.

We can also use the Likelihood Ratio Test to compare the OLS model and the Spatial Lag model. The p value of the test is less than 0.05, we reject the null hypothesis and state that the Spatial Lag model is doing a better job than the OLS model.

As is showed in the scatterplot of Spatial Lag regression residuals below, the Moran's I of the Spatial Lag regression residuals is -0.082, which is closer to 0 than the Moran's I = 0.313 of the OLS regression residuals. There seem to be less spatial autocorrelation in these residuals than in OLS residuals.

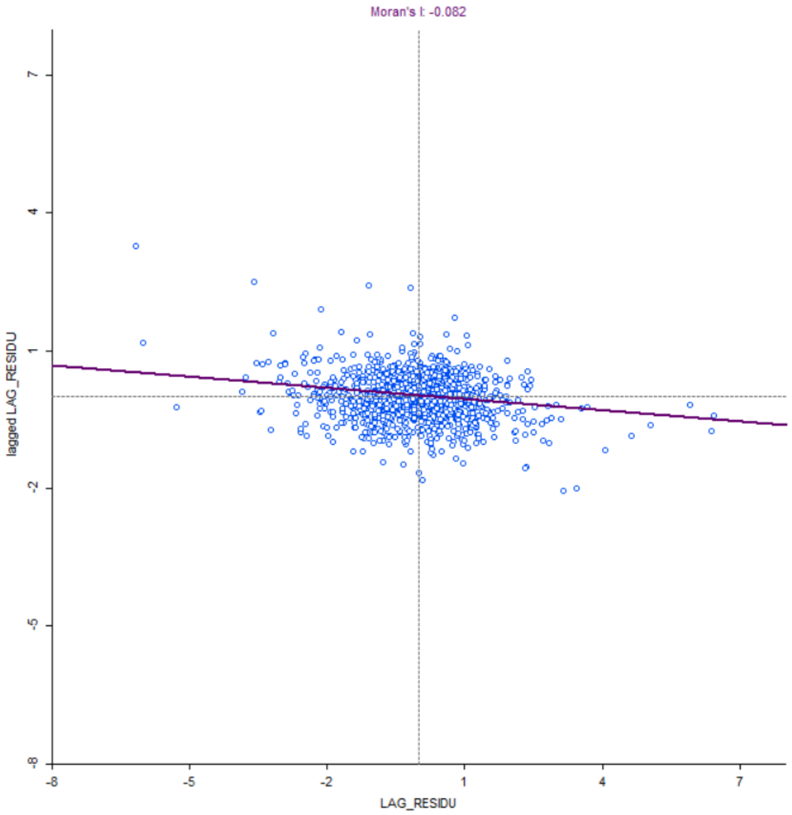
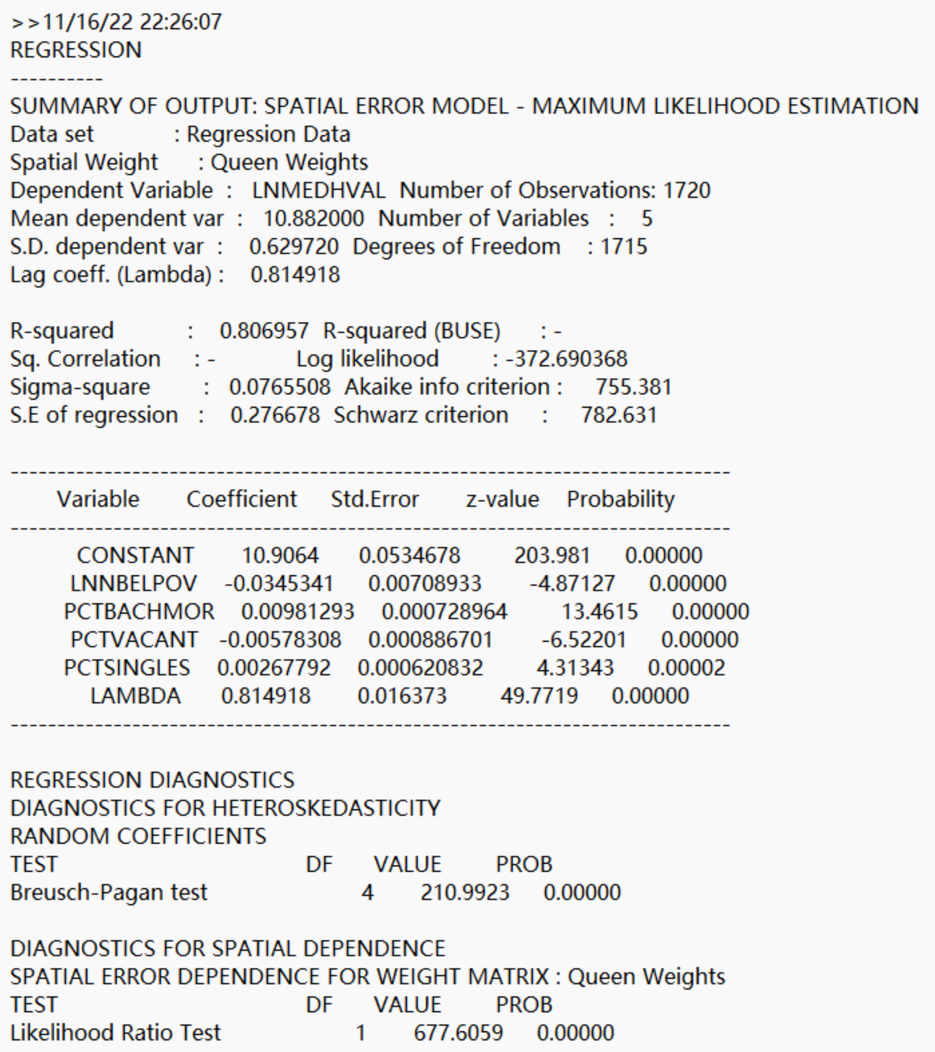


Figure 6

Overall, the Spatial Lag Model performs much better at accounting for the spatial processes that exist within the data based on all of these criteria.



```
>>11/16/22 22:26:07
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : Regression Data
Spatial Weight      : Queen Weights
Dependent Variable  :   LNMEDHVAL  Number of Observations: 1720
Mean dependent var  :   10.882000  Number of Variables   :    5
S.D. dependent var  :   0.629720  Degrees of Freedom    : 1715
Lag coeff. (Lambda) :   0.814918

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

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

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                              DF      VALUE        PROB
Breusch-Pagan test                 4      210.9923     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : Queen Weights
TEST                              DF      VALUE        PROB
Likelihood Ratio Test              1      677.6059     0.00000
```

Table 3

The Spatial Error Regression output from GeoDa is present as follows. From Table 3, the coefficient of the spatial parameter λ is 0.81, and it is significantly important ($p < 0.05$). This indicates that the LNMEDHVAL is highly correlated with some unexplained variation with a spatial component.

After introducing the spatial parameter λ, we can see that the remaining terms (LNNBELPOV, PCTBACHMOR, PCTSINGLES, and PCTVACANT) all are still significant ($p<0.05$ for all variables).

The Breusch-Pagan test, as observed in the GeoDa results summary, indicates that there is an issue with heteroscedastic residuals ($p < 0.05$). There is still a problem with heteroscedasticity.

Based on the Table3 and Table1, we can compare the Spatial Error regression and OLS regression models based on the Akaike Information Criterion (AIC) /Schwarz Criterion (SC), the Log Likelihood, and the Likelihood Ratio Test.

For the Spatial Error regression, AIC = 755.381, SC = 782.631. For the OLS regression, AIC = 1432.99, SC = 1460.24. Because the lower the AIC and SC, the better the fits of this model. The results indicates that the Spatial Error model is a better fit than OLS model.

For the Spatial Error regression, Log Likelihood = -373.690. For the OLS regression, Log Likelihood = -711.493. Because the higher the log likelihood, the better the fit of this model. The results indicates that the Spatial Error model is a better fit than OLS model.

We can also use the Likelihood Ratio Test to compare the OLS model and the Spatial Error model. The p value of the test is less than 0.05, we reject the null hypothesis and state that the Spatial Lag model is doing a better job than the OLS model.

As is showed in the scatterplot of Spatial Error regression residuals below, the Moran's I of the Spatial Lag regression residuals is 0.095, which is closer to 0 than the Moran's I = 0.313 of the OLS regression residuals. There seem to be less spatial autocorrelation in these residuals than in OLS residuals.

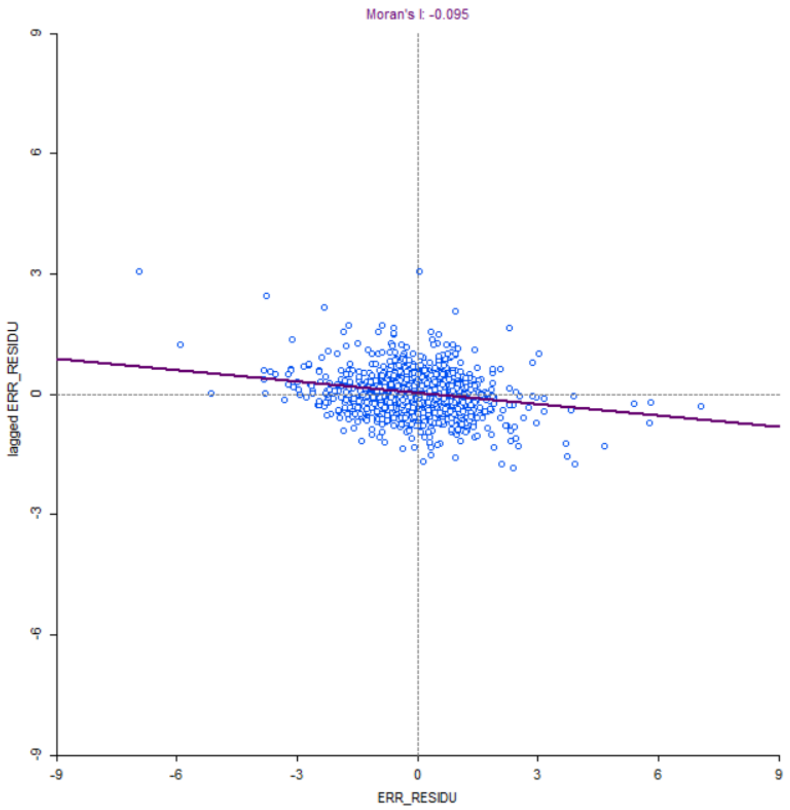


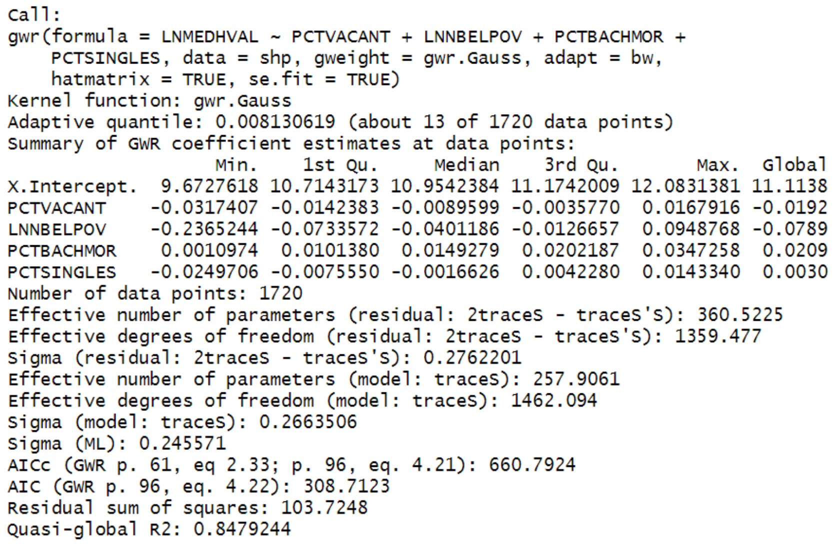
Figure 7

Overall, based on the Akaike Information Criterion/Schwarz Criterion, the Log Likelihood, the Likelihood Ratio Test, and the Moran's I scatterplots, we can confidently say the spatial error regression model is doing better than the OLS regression model.

From the regression results that are listed above, we can see that the value of the Akaike Information Criterion (AIC) for Spatial Lag Regression model is 523.48, and the value for Spatial Error Regression model is 755.381. That says, the Spatial Lag Regression model is doing better than the Special Error Regression model. When it comes to the Schwarz criterion (SC), the value for Spatial Lag Regression model is 556.18, and the value for Spatial Error Regression model is 782.631, also indicating the Spatial Lag Regression model is better.

d) Geographically Weighted Regression Results

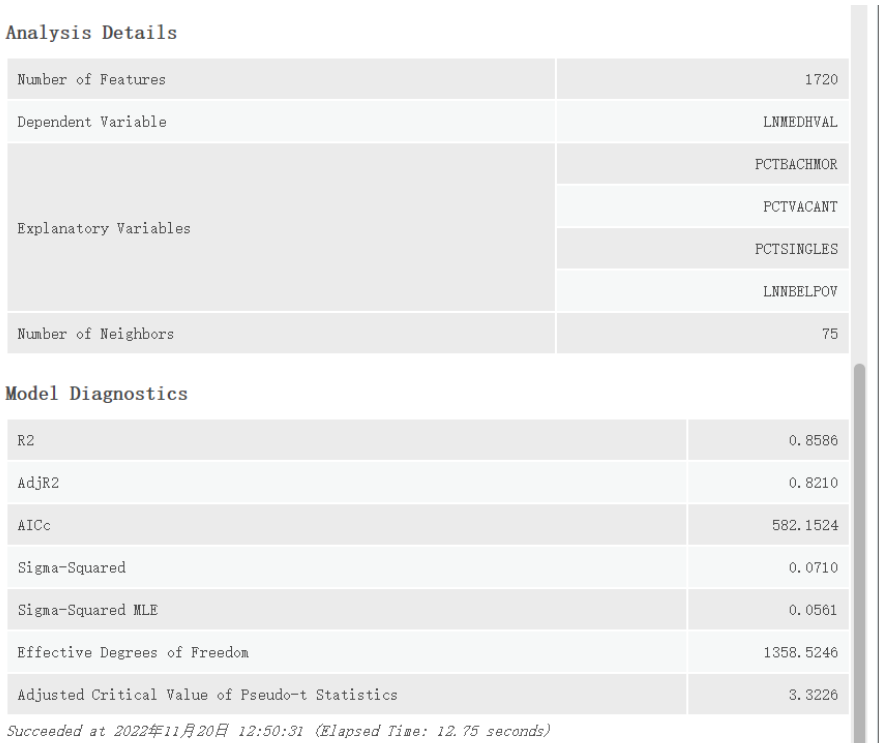
We performed GWR in R, and the following is my presentation and explanation of the results.



```
Call:
gwr(formula = LNMEDHVAL ~ PCTVACANT + LNNBELPOV + PCTBACHMOR +
    PCTSINGLES, data = shp, gweight = gwr.Gauss, adapt = bw,
    hatmatrix = TRUE, se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.008130619 (about 13 of 1720 data points)
Summary of GWR coefficient estimates at data points:
                   Min.    1st Qu.     Median    3rd Qu.       Max.  Global
X.Intercept.  9.6727618 10.7143173 10.9542384 11.1742009 12.0831381 11.1138
PCTVACANT    -0.0317407 -0.0142383 -0.0089599 -0.0035770  0.0167916 -0.0192
LNNBELPOV    -0.2365244 -0.0733572 -0.0401186 -0.0126657  0.0948768 -0.0789
PCTBACHMOR    0.0010974  0.0101380  0.0149279  0.0202187  0.0347258  0.0209
PCTSINGLES   -0.0249706 -0.0075550 -0.0016626  0.0042280  0.0143340  0.0030
Number of data points: 1720
Effective number of parameters (residual: 2traceS - traceS'S): 360.5225
Effective degrees of freedom (residual: 2traceS - traceS'S): 1359.477
Sigma (residual: 2traceS - traceS'S): 0.2762201
Effective number of parameters (model: traceS): 257.9061
Effective degrees of freedom (model: traceS): 1462.094
Sigma (model: traceS): 0.2663506
Sigma (ML): 0.245571
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 660.7924
AIC (GWR p. 96, eq. 4.22): 308.7123
Residual sum of squares: 103.7248
Quasi-global R2: 0.8479244
```

Table 4

We also used arcGISpro to conduct GWR operation. The following figure is the result of the supplementary table:



Analysis Details

Number of Features	1720
Dependent Variable	LNMEDHVAL
Explanatory Variables	PCTBACHMOR PCTVACANT PCTSINGLES LNNBELPOV
Number of Neighbors	75

Model Diagnostics

R2	0.8586
AdjR2	0.8210
AICc	582.1524
Sigma-Squared	0.0710
Sigma-Squared MLE	0.0561
Effective Degrees of Freedom	1358.5246
Adjusted Critical Value of Pseudo-t Statistics	3.3226

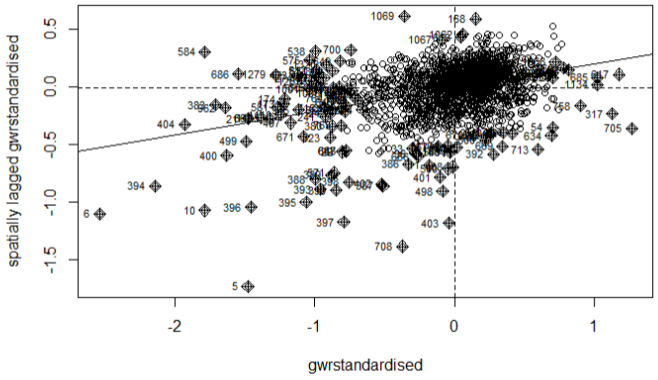
Succeeded at 2022年11月20日 12:50:31 (Elapsed Time: 12.75 seconds)

It can be found that the results of R are slightly different from those of arcGISpro, because the gold search method of arcGISpro is different from the AIC optimization algorithm in arcMap and R. However, other indicators, such as the values of R-squared, are basically consistent, indicating that the results of arcGISpro still have a certain reference.

Firstly, regression results of adaptive bandwidth are shown. In Table 4, comparing with the OLS regression, we can find the difference between the residuals and squares of GWR and OLS. The R-squared of GWR (overall) is 0.848, while the R-squared of OLS is 0.662. This indicates that the geographically weighted regression method is more suitable for the prediction of this data set, because the R-squared of GWR is larger, which can better explain the variance of the dependent variable.

Akaike Information Criteria (AIC) is used to compare GWR and OLS, spatial lag regression and spatial error regression. The Akaike Information Criteria for GWR is 308, and the Akaike Information Criteria for spatial lag regression is 523. The Akaike Information Criteria for spatial error regression is 755. We know that the smaller Akaike Information Criteria are, the better the fit of regression is. Therefore, in comparison, GWR has the smallest AIC, indicating that GWR is more suitable for models with higher complexity and can better fit the observed data.

The following gallery shows Moran's I scatterplot for GWR, as well as OLS, spatial lag regression, and spatial error regression. It can be seen that GWR's Moran's I scatterplot and OLS's Moran's I scatterplot indicate that in addition to the positive correlation of the data, the autocorrelation of the former is slightly lower than that of the latter, but compared with spatial lag regression and spatial error regression, GWR's Moran's I scatterplot showed stronger autocorrelation, and both spatial lag regression and spatial error regression showed negative correlation.



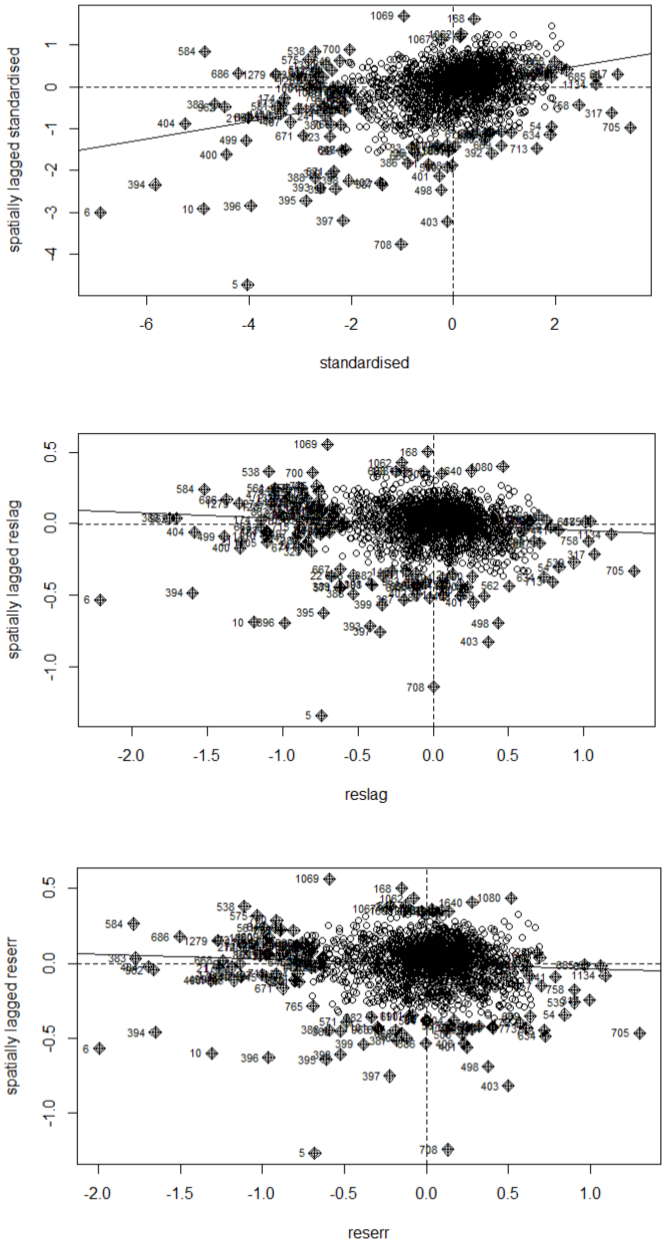
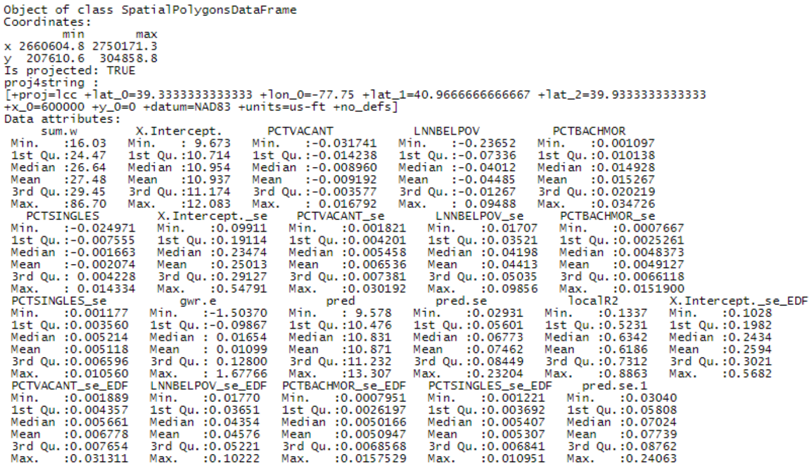
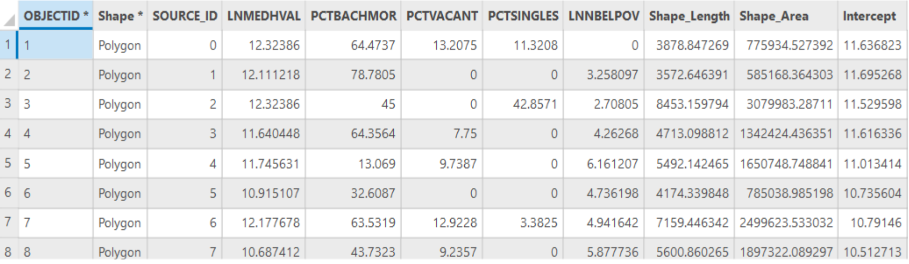


Figure 8

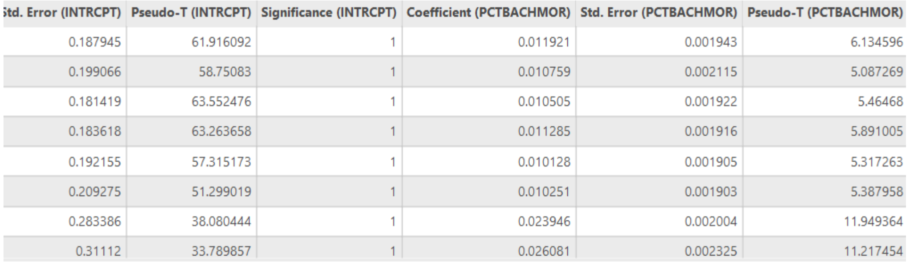
You can call "SDF" in the geographically weighted regression to view the attribute table of the local regression results, as shown in the figure below. I also present the attribute table of the arcGISpro result, which contains the same variable index. The R-square values of the observation results are 0.1337~0.8863, all of which are positive, indicating that the output results are credible, which is different from the conclusion tried in ArcGIS pro. According to the attribute table, the minimum value of local R-square is 0.1337. In fact, the R-square of many observed values in the data set is actually very low, which can basically be regarded as 0. According to the intercept, we can find that the coefficients of PCTVACANT, LNNBELPOV and PCTSINGLES are negative, indicating that the relationship between these three prediction variables and the dependent variables is negatively correlated, while the coefficient of PCTBACHMOR is positive, indicating that it is positively correlated with the dependent variables. This table means that when the PCTVACANT is added by one unit, the dependent variable will decrease by around 0.0092 units, and when LNNBELPOV is added by one unit, the dependent variable will decrease by 0.045 units, when PCTBACHMOR is added by one unit, the dependent variable will increase by 0.015 units, when PCTSINGLES is added by one unit, the dependent variable will decrease by 0.002 units. We can also see that the precision error of the coefficient, the lower the standard error of 0 relative to the coefficient, the more reliable the coefficient will be. When the standard error is equal to or greater than half of the coefficient, it is very unreliable and most likely the result of chance.



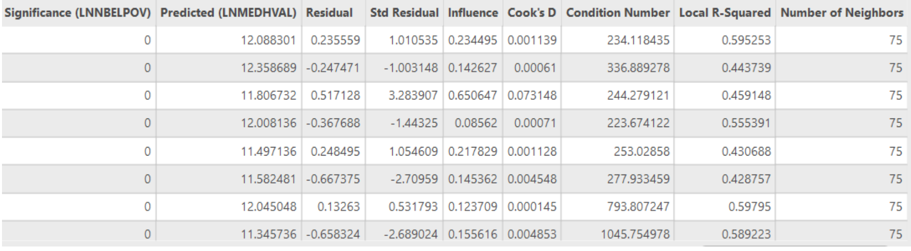
```
Object of class SpatialPolygonsDataFrame
Coordinates:
        min       max
x 2660604.8 2750171.3
y  207610.6  304858.8
Is projected: TRUE
proj4string :
[+proj=lcc +lat_0=39.3333333333333 +lon_0=-77.75 +lat_1=40.9666666666667 +lat_2=39.9333333333333
+x_0=600000 +y_0=0 +datum=NAD83 +units=us-ft +no_defs]
Data attributes:
     sum.w         X.Intercept.       PCTVACANT             LNNBELPOV           PCTBACHMOR
 Min.   :16.03   Min.   : 9.673   Min.   :-0.031741   Min.   :-0.23652   Min.   :0.001097
 1st Qu.:24.47   1st Qu.:10.714   1st Qu.:-0.014238   1st Qu.:-0.07336   1st Qu.:0.010138
 Median :26.64   Median :10.954   Median :-0.008960   Median :-0.04012   Median :0.014928
 Mean   :27.48   Mean   :10.937   Mean   :-0.009192   Mean   :-0.04485   Mean   :0.015267
 3rd Qu.:29.45   3rd Qu.:11.174   3rd Qu.:-0.003577   3rd Qu.:-0.01267   3rd Qu.:0.020219
 Max.   :86.70   Max.   :12.083   Max.   : 0.016792   Max.   : 0.09488   Max.   :0.034726
   PCTSINGLES          X.Intercept._se    PCTVACANT_se        LNNBELPOV_se       PCTBACHMOR_se
 Min.   :-0.024971   Min.   :0.09911   Min.   :0.001821   Min.   :0.01707   Min.   :0.0007667
 1st Qu.:-0.007555   1st Qu.:0.19114   1st Qu.:0.004201   1st Qu.:0.03521   1st Qu.:0.0025261
 Median :-0.001663   Median :0.23474   Median :0.005458   Median :0.04198   Median :0.0048373
 Mean   :-0.002074   Mean   :0.25013   Mean   :0.006536   Mean   :0.04413   Mean   :0.0049127
 3rd Qu.: 0.004228   3rd Qu.:0.29127   3rd Qu.:0.007381   3rd Qu.:0.05035   3rd Qu.:0.0066118
 Max.   : 0.014334   Max.   :0.54791   Max.   :0.030192   Max.   :0.09856   Max.   :0.0151900
 PCTSINGLES_se          gwr.e                pred            pred.se            localR2        X.Intercept._se_EDF
 Min.   :0.001177   Min.   :-1.50370   Min.   : 9.578   Min.   :0.02931   Min.   :0.1337   Min.   :0.1028
 1st Qu.:0.003560   1st Qu.:-0.09867   1st Qu.:10.476   1st Qu.:0.05601   1st Qu.:0.5231   1st Qu.:0.1982
 Median :0.005214   Median : 0.01654   Median :10.831   Median :0.06773   Median :0.6342   Median :0.2434
 Mean   :0.005118   Mean   : 0.01099   Mean   :10.871   Mean   :0.07462   Mean   :0.6186   Mean   :0.2594
 3rd Qu.:0.006596   3rd Qu.: 0.12800   3rd Qu.:11.232   3rd Qu.:0.08449   3rd Qu.:0.7312   3rd Qu.:0.3021
 Max.   :0.010560   Max.   : 1.67766   Max.   :13.307   Max.   :0.23204   Max.   :0.8863   Max.   :0.5682
 PCTVACANT_se_EDF   LNNBELPOV_se_EDF  PCTBACHMOR_se_EDF   PCTSINGLES_se_EDF    pred.se.1
 Min.   :0.001889   Min.   :0.01770   Min.   :0.0007951   Min.   :0.001221   Min.   :0.03040
 1st Qu.:0.004357   1st Qu.:0.03651   1st Qu.:0.0026197   1st Qu.:0.003692   1st Qu.:0.05808
 Median :0.005661   Median :0.04354   Median :0.0050166   Median :0.005407   Median :0.07024
 Mean   :0.006778   Mean   :0.04576   Mean   :0.0050947   Mean   :0.005307   Mean   :0.07739
 3rd Qu.:0.007654   3rd Qu.:0.05221   3rd Qu.:0.0068568   3rd Qu.:0.006841   3rd Qu.:0.08762
 Max.   :0.031311   Max.   :0.10222   Max.   :0.0157529   Max.   :0.010951   Max.   :0.24063
```



	OBJECTID *	Shape *	SOURCE_ID	LNMEDHVAL	PCTBACHMOR	PCTVACANT	PCTSINGLES	LNNBELPOV	Shape_Length	Shape_Area	Intercept
1	1	Polygon	0	12.32386	64.4737	13.2075	11.3208	0	3878.847269	775934.527392	11.636823
2	2	Polygon	1	12.111218	78.7805	0	0	3.258097	3572.646391	585168.364303	11.695268
3	3	Polygon	2	12.32386	45	0	42.8571	2.70805	8453.159794	3079983.28711	11.529598
4	4	Polygon	3	11.640448	64.3564	7.75	0	4.26268	4713.098812	1342424.436351	11.616336
5	5	Polygon	4	11.745631	13.069	9.7387	0	6.161207	5492.142465	1650748.748841	11.013414
6	6	Polygon	5	10.915107	32.6087	0	0	4.736198	4174.339848	785038.985198	10.735604
7	7	Polygon	6	12.177678	63.5319	12.9228	3.3825	4.941642	7159.446342	2499623.533032	10.79146
8	8	Polygon	7	10.687412	43.7323	9.2357	0	5.877736	5600.860265	1897322.089297	10.512713



Std. Error (INTRCPT)	Pseudo-T (INTRCPT)	Significance (INTRCPT)	Coefficient (PCTBACHMOR)	Std. Error (PCTBACHMOR)	Pseudo-T (PCTBACHMOR)
0.187945	61.916092	1	0.011921	0.001943	6.134596
0.199066	58.75083	1	0.010759	0.002115	5.087269
0.181419	63.552476	1	0.010505	0.001922	5.46468
0.183618	63.263658	1	0.011285	0.001916	5.891005
0.192155	57.315173	1	0.010128	0.001905	5.317263
0.209275	51.299019	1	0.010251	0.001903	5.387958
0.283386	38.080444	1	0.023946	0.002004	11.949364
0.31112	33.789857	1	0.026081	0.002325	11.217454



Significance (LNNBELPOV)	Predicted (LNMEDHVAL)	Residual	Std Residual	Influence	Cook's D	Condition Number	Local R-Squared	Number of Neighbors
0	12.088301	0.235559	1.010535	0.234495	0.001139	234.118435	0.595253	75
0	12.358689	-0.247471	-1.003148	0.142627	0.00061	336.889278	0.443739	75
0	11.806732	0.517128	3.283907	0.650647	0.073148	244.279121	0.459148	75
0	12.008136	-0.367688	-1.44325	0.08562	0.00071	223.674122	0.555391	75
0	11.497136	0.248495	1.054609	0.217829	0.001128	253.02858	0.430688	75
0	11.582481	-0.667375	-2.70959	0.145362	0.004548	277.933459	0.428757	75
0	12.045048	0.13263	0.531793	0.123709	0.000145	793.807247	0.59795	75
0	11.345736	-0.658324	-2.689024	0.155616	0.004853	1045.754978	0.589223	75

Table 5

Below is the map distribution of local R-squared in the GWR results. It can be seen that the R-squared differentiation of Philadelphia is very different, with some regions showing very high R-squared, indicating that there is a significant relationship between each predictor and dependent variable. For example, the GWR model fits poorly in the central city and north Philadelphia with low local R-squared values.

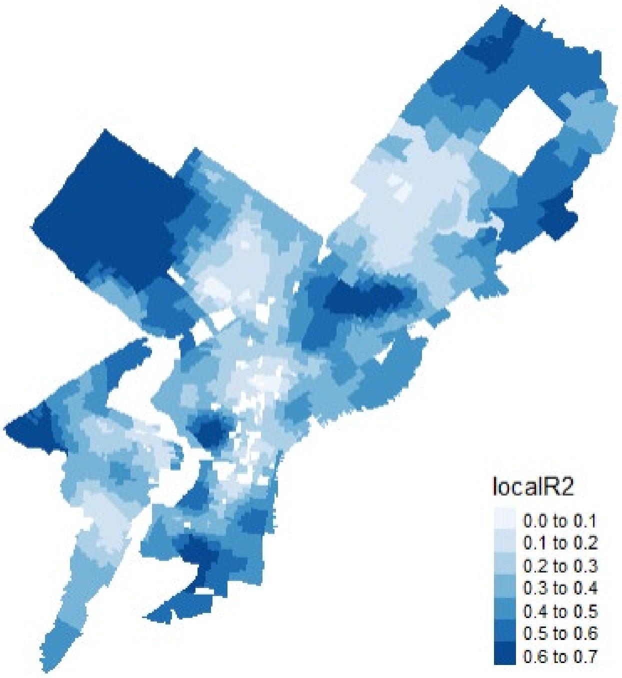


Figure 9

4. Discussion

After the regression of four models, we get different results and indicators. By comparing the four models, we can say that GWR is the most suitable regression model for this data set.

First of all, GWR is more suitable for data with complex geographic distribution than OLS, which cannot take into account the potential spatial instability. Moreover, by comparing the results of R-squared, GWR is higher (0.848>0.662), indicating that GWR produces better fitting. GWR with different specifications still produces lower AIC than OLS. However, in fact, we should not compare GWR AIC with OLS AIC because there is no spatial autocorrelation in GWR residuals, but there is spatial autocorrelation in OLS. Compared with spatial lag regression and spatial error regression, GWR is more suitable. By comparing the AIC value, it can be found that the AIC value of GWR is the smallest.

There are still some limitations for all these models. Firstly, the results of Breush-Pagan Test for all the models are statistically significant with a p value < 0.05, which means all the models have a problem with heteroscedasticity. Secondly, although the GWR model performed best, there is still a problem with spatial autocorrelation compared with Spatial Lag regression model and Spatial Error model.