CPLN HW 2

Ajjit Narayanan

Bill Cohen

Introduction

In this paper, we try to estimate the effect of various socioeconomic factors on median household value at the block group level in Philadelphia using spatial regression techniques. In a previous paper, we used ordinary least squares (OLS) regression to model the effect of four predictor variables on median house value: the percentage of vacant lots, the number of people in poverty, the percentage of people with bachelors degrees, and percent of single house units. While OLS regression is a useful and relevant tool, it is not the most appropriate method for modeling data with a spatial component; its assumptions of that observed values and model residuals are independent are violated by data with a spatial relationship.

Here, we model the same Philadelphia data set using three spatial regression methods: spatial lag, spatial error, and geographically weighted regression, and compare them to our results from our previous OLS regression. We expect these spatial analyses will better model the relationships between our predictor and response variables because they account for the spatial dependencies that violate the assumptions of OLS.

Methods

A Description of the Concept of Spatial Autocorrelation

Spatial regression seeks to describe how spatial distributions influence the relationships between variables, called spatial correlation, and those between values within a single variable, called spatial autocorrelation. The nature of these spatial dependencies can be summarized by Waldo Tobler's first law of geography, which states, "everything is related to everything else, but near things are more related than distant things". In Philadelphia, or at the city level generally, we can apply this concept through spatial analysis to better understand unique neighborhood characteristics. For example, we intuitively understand that the market value of a home in a new housing development is likely to be higher if it is next to a posh neighborhood full of expensive

homes (positive spatial autocorrelation), than if it is next to a few blocks of vacant properties (negative spatial correlation).

In the OLS regression presented in the previous paper, we made predictions about spatial autocorrelation by looking at choropleth maps of each variable. From that visual assessment, we noted areas of block groups sharing similar values, indicating the need to test for spatial autocorrelation. To do so, we use the Moran's I test which compares a variable's value at each observation to the value(s) of nearby observations. These "nearest neighbors" are selected by a user-defined weight matrix.

Moran's I is defined as:

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

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

where $\overline{X}$ is the global mean of variable X, X_i is the value of the variable at location i, X_j is the value of the variable at location j, W_{ij} is the weight of j relative to i, and n is the total number of observations in the data set.

For each observation n, an n x n weight matrix defines the value of the pairwise spatial relationship with the remaining $n - 1$ observations. Throughout our primary analysis we will be using a Queen weight matrix. Observations are said to have queen contiguity with neighbors that share a segment or vertex, and Queen neighbors will have a pairwise value of 1 in the matrix. Observations with no shared features will have a pairwise value of 0. For our data set, the Queen matrix identifies those census blocks that share a border or corner. For comparison, we use a secondary distance-weighted matrix with a fixed distance of 5,000 ft. Census block groups whose centroids are within 5,000 ft. of the centroid of a given observation are considered neighbors and given a pairwise matrix value of 1. All others outside the radius of 5,000 ft. are given a matrix value of 0.

There are several other types of weight matrices and methods of neighborhood evaluation that could be considered in spatial analysis. The Moran's I test requires a symmetric weight matrix for its analysis, meaning the pairwise weight of any observation j on another observation i is the same as the weight of i on j. A k-nearest neighbor weight matrix is not symmetric and therefore cannot be used with the Moran's I test.

The values of Moran's I typically range from -1, indicating a strong negative autocorrelation, to 1, indicating a strong positive autocorrelation. In some cases the value of Moran's I can extend beyond these bounds. To test for significance, we randomly shuffle the values for all observations, (i.e. move the value of the variable for a given observation to another randomly selected location, then move that value to another location, and so on for all n observations), then re-run the Moran's I test. If we run this random permutation and calculate Moran's I many times, we should expect the results to be normally distributed with a mean near 0, indicating no spatial autocorrelation due to the randomness of the permutations. In our analysis, we calculate Moran's I for 999 random permutations.

To test for significance, we rank the results from highest to lowest for all 1,000 tests, including our real data set and all random permutations. The rank of our observed Moran's I is divided by the total number of tests ranked (in this case 1,000) to get a pseudo p-value, indicating the likelihood of randomly getting a Moran's I value at least as large as the real observed value. We can use the pseudo p-value to test the null hypotheses, H_a, which states that there is no spatial autocorrelation (i.e. a random pattern). If there is truly no spatial autocorrelation, the terms in the equation above will cancel such that E(I) = -1/(n-1). If 0.05 < pseudo p-value < 0.95, we fail to reject to null hypothesis and do not observe significant spatial autocorrelation. If the pseudo p-value < 0.05, we can reject H_0 for our first alternative hypothesis H_{a1}, indicating significant observed positive spatial autocorrelation. If the pseudo p-value > 0.95, we reject both H_0 and H_{a1}, leading to $H_{a2 }$ of significant observed negative spatial autocorrelation.

A Review of OLS Regression and Assumptions

In the previous assignment, we utilized multiple ordinary least squares (OLS) regression analysis to estimate and compare the strength of the relationship between median house value and the four predictor variables defined in this report. OLS regression models the linear relationship between a response variable and one or more predictor variables by fitting an equation that minimizes the sum of squared deviations between predicted and observed values of the response variable. Five key assumptions characterize a linear regression model: a linear relationship between the response variable and each predictor (linearity), proximity of observations in time or space does not influence values (independence of observations), model residuals should be normally distributed (normality of residuals) and have a constant variance (homoscedasticity), and no correlation between predictors (no multicollinearity). For a detailed review of OLS regression and its application for this data set, please see our previous assignment.

The assumptions that OLS residual errors are random and independent are often violated by data with a spatial nature, such as census block group data. Using Moran's I, we can test the OLS residuals for spatial autocorrelation to determine if this assumption is violated. Alternatively, we can check this assumption by regressing the residuals for each observation on the "lagged" residuals, which are mean values of neighboring residuals as defined by a weight matrix. The resulting regression coefficient for lagged residuals is known as ρ rho, or lambda in GeoDa. It will indicate the degree and direction of spatial autocorrelation ranging from -1 to 1. In either test, if the null hypothesis is rejected, there is a significant spatial association between residuals and the assumption is violated, indicating the parameters determined by the OLS model do not fully account for the relationship between dependent and independent variables.

The software package GeoDa is used to conduct many of the tests used in this analysis. It offers methods to test for other OLS assumptions as well. The Breusch-Pagan test, the Koenker-Bassett test, and the White test are three diagnostics that test for heteroscedasticity. The null hypothesis for these three tests is that the data is homoscedastic and the assumption for OLS holds true, meaning the residuals are random with a constant variance. If the p-value is <0.05, the null hypothesis is rejected for the alternative hypothesis of hetereoscedastcity, meaning the residuals are not random and the assumption is violated. The Jarque-Bera test in GeoDa tests for another OLS assumption-normality of residuals. The null hypothesis is that the assumption holds and the residuals have a normal distribution. If the p-value is <0.05, the null hypothesis is rejected and the assumption is violated.

Spatial Lag and Spatial Error Regression

The first two spatial regressions performed in this paper are spatial lag and spatial error regression. GeoDa will be used to run these analyses. Both methods attempt to account for spatial autocorrelation in the data by adding one or more spatially lagged variables as predictor terms in the OLS equation.

In our spatial lag model, we assume that the value of the dependent variable for a given observation is associated with the values of neighboring locations, with neighbors defined by a queen weight matrix. To measure this association, a lagged variable term, $\rho * Wy$, is added to the OLS regression equation, where Wy is the lagged dependent variable and ρ is the coefficient of the y-lag variable. The value of the lagged dependent variable Wy is calculated as the mean dependent variable value of queen neighbors. And ρ can take on values between -1 and and is used as a measure of how pendent the dependent variable is on the lagged value of the dependent variable.

Spatial lag regression uses the following equation:

$$LNMEDHVAL = \rho W_{LNMEDHVAL} + \beta_0 + \beta_1(LNNBELPOV100) + \beta_2(PCTBACHMOR) + \beta_3(PCTSINGLE) + \beta_4(PCTVACANT) + \varepsilon$$

where $W_{LNMEDHVAL}$ is the lagged variable of LNMEDHVAL with coefficient ρ (-1 < ρ < 1). The remaining parameters, defined in OLS regression, are as follows:

$\beta_0 = E(\hat{\beta}_0)$ = population value of the y-intercept; $\hat{\beta}_0$ = sample statistic estimating the population y-intercept

$\beta_i = E(\hat{\beta}_i)$ = population value of the slope coefficient for predictor i.

ε is a random error term such that $\sigma^2 = E(\hat{\sigma}^2)$ = population variance of residuals ε where $\varepsilon \sim N(0, \sigma^2)$

In a spatial error model, we assume that the value of the residual for a given observation is associated with the values of the OLS residuals of neighboring locations, with neighbors defined by the weights matrix. To measure this association, the OLS regression is first run on its own without any augmentation. The error term ε is then regressed on the lagged residuals of its queen neighbors using a coefficient parameter λ, lagged residual term W_ε and a random noise term u. This new set of terms is then used to replace the original epsilon term of the OLS model equation, to provide the full spatial error model equation:

$$LNMEDHVAL = \beta_0 + \beta_1(LNNBELPOV100) + \beta_2(PCTBACHMOR) + \beta_3(PCTSINGLE) + \beta_4(PCTVACANT) + \lambda W\varepsilon + u$$

where $W\varepsilon$ is the lagged variable of the OLS residual term ε with coefficient λ (-1 < λ < 1) and u is a random error term, defined by the equation $\varepsilon = \lambda W\varepsilon + u$

The remaining parameters, defined in OLS regression, are:

$\beta_0 = E(\hat{\beta}_0)$ = population value of the y-intercept; $\hat{\beta}_0$ = sample statistic estimating the population y-intercept

$\beta_i = E(\hat{\beta}_i)$ = population value of the slope coefficient for predictor i.

Spatial lag and spatial error regression account for the spatial dependency issues that would otherwise violate the assumption of independence of observations in OLS regression. However, all of the other OLS assumptions discussed above still apply to these methods.

After running these two spatial analyses, we compare their results to the OLS model developed in the previous paper. We assess the performance of these three

models based on a variety of criteria, including the Akaike Information Criterion and Schwartz Criterion, the Log Likelihood, and the Likelihood Ratio Test.

The results of the Akaike Information Criterion (AIC) and the Schwartz Criterion (SC) offer a way to compare the relative goodness of fit of two or more models, and can be used to compare both our spatial regression and OLS results. The model with the lowest AIC and SC result will be the best performing, meaning it loses the least amount of information.

The Log Likelihood method is used to assess the relative strength of nested models and comes from estimating the parameters using Maximum Likelihood Estimation. In this case, there is nesting between the OLS regression and each of the spatial models. The spatial lag and spatial error equations are based on the OLS model equation and both contain all of the OLS model terms, so we can say the OLS model is a special case of of spatial lag and spatial error. However we cannot use Log Likelihood to compare spatial lag and spatial error because there is no nesting in either direction. Unlike the AIC and SC comparisons, the model with highest Log Likelihood value (or least negative) is the strongest.

The Likelihood ratio test is only used to compare OLS to either spatial lag or spatial error regression models. The test assumes that both are just as good and no one is better than the other as the null hypothesis H_0. If the results of the LRT test show a p-value of <0.05, the null hypothesis is rejected for the alternative hypothesis that the model with the higher (less negative) Log Likelihood is a better model.

In addition to running these diagnostic comparisons, we can use the Moran's I test on the residuals to look at the severity of spatial autocorrelation present in the regression residuals for each model. The model with the lowest Moran's I value has the least spatial autocorrelation among residuals.

Geographically Weighted Regression

Using ArcGIS (ArcMap 10.5) we conduct a third method of spatial analysis-geographically weighted regression-which attempts to deal with localized variation in spatial correlation and autocorrelation.

Both spatial lag and spatial error regression assume spatial stationary across the data set. For the data to be characterized by spatial stationary, the relationships described by the regression model have to be uniform throughout the study area. In reality, this is not the case. In the Philadelphia region for instance, housing prices in a suburban neighborhood will be determined by a different set of conditions than housing prices in Center City or South Philadelphia. Said differently, the weight of

predictors on a response variable in the study area can differ significantly at the local level. This is known as spatial non-stationary.

Simpson's paradox states this phenomenon for statistical data more generally. In broad terms, Simpson's paradox acknowledges that trends identified for a sample as a whole may aggregate and overlook significantly different relationships in subsets of the sample data. GWR attempts to address this problem by running distance-weighted local regressions for each observation such that for each model, near observations assert a stronger influence than distant observations. This results in n unique local models which can be examined individually, but more usefully, can be compared and combined using some global regression diagnostics.

The equation for our geographically weighted regression model is:

$$LNMEDHVAL_i = \beta_{i0} + \beta_{i1}(LNNBELPOV100) + \beta_{i2}(PCTBACHMOR) + \beta_{i3}(PCTSINGLE) + \beta_{i4}(PCTVACANT) + \varepsilon_i$$

where $i = 1 \ldots n$ for n observations such that a unique model exists for each location i. The parameters β_{i0},β_{i1}, β_{i2}, β_{i3}, β_{i4}, and ε_i are determined for each observation using a distance weighting function described below. They can be interpreted as estimating the values of the local statistics for each observation:

$\beta_{i0} = E(\hat{\beta}_{i0})$ = local value of the y-intercept centered at observation i; $\hat{\beta}_{i0}$ = sample statistic estimating the local y-intercept for the neighborhood surrounding observation i.

$\beta_i = E(\hat{\beta}_i)$ = population value of the slope coefficient for predictor i.

Rather than contiguity determining neighborhood weights as used in spatial lag and spatial error regression, the weights used in GWR are determined by a kernel decay function of bandwidth, where the radius and strength of influence at a given location can be set to a fixed or variable distance.

If a fixed-distance bandwidth is chosen, the kernel weights for each observation are calculated by the following decay function:

$$w_{ij} = \begin{cases} e^{-0.5(\frac{distance_{ij}}{h})^2}, & \text{if distance}_{ij} \leq h \\ 0, & \text{otherwise} \end{cases}$$

where weight is purely a function of distance and any observations outside the radius of influence are not considered.

If adaptive bandwidth is chosen, the number of observations considered influential will be a constant k for each local regression. The distance used in the kernel function will vary for each location such that k nearest neighbors are included.

$$w_{ij} = \begin{cases} [1 - (\frac{distance_{ij}}{h})^2]^2, & \text{if distance}_{ij} \leq h \\ 0, & \text{otherwise} \end{cases}$$

It is crucial to select a bandwidth that will accurately account for the spatial distribution of the data. Fixed-distance bandwidth is more appropriate for observations with a constant density, while adaptive bandwidth is more appropriate for unevenly clustered data. In the GWR analysis presented here, we will use adaptive bandwidth to account for the wide range of census block group sizes and the disparity of their distribution. (Large blocks tend to be clustered with other large blocks, while small blocks are clustered near other small blocks.) So if we used a fixed bandwidth, the models for small, tightly clustered census blocks would incorporate a high number of observations while those for large spread out blocks would have very few observations.

As with spatial lag and spatial error regression, GWR attempts to account for the spatial dependence in the data, which violates the OLS assumption of independence of observations and residuals. A valid GWR model still must meet the other OLS assumptions of normality of residuals, homoscedasticity, and no multicollinearity. Additional, GWR requires a large number of observations.

Multicollinearity, when two or more predictor variables exhibit a linear association, is identified on a feature by feature basis using the Condition Number in the attribute table of ArcMap's GWR regression output. In our OLS model, we looked for global multicollinearity using the pairwise Pearson correlation matrix and the VIF test. However in GWR, strong predictor correlations may only be clustered in certain sub-regions throughout the study area, resulting in over or under prediction in those areas in a global model. In GWR, the condition number provides a local diagnostic for multicollinearity at each observation. Null values, values greater than 30, or a value equal to -1.7976931348623158e + 308, indicate issues with local multicollinearity for a given local model.

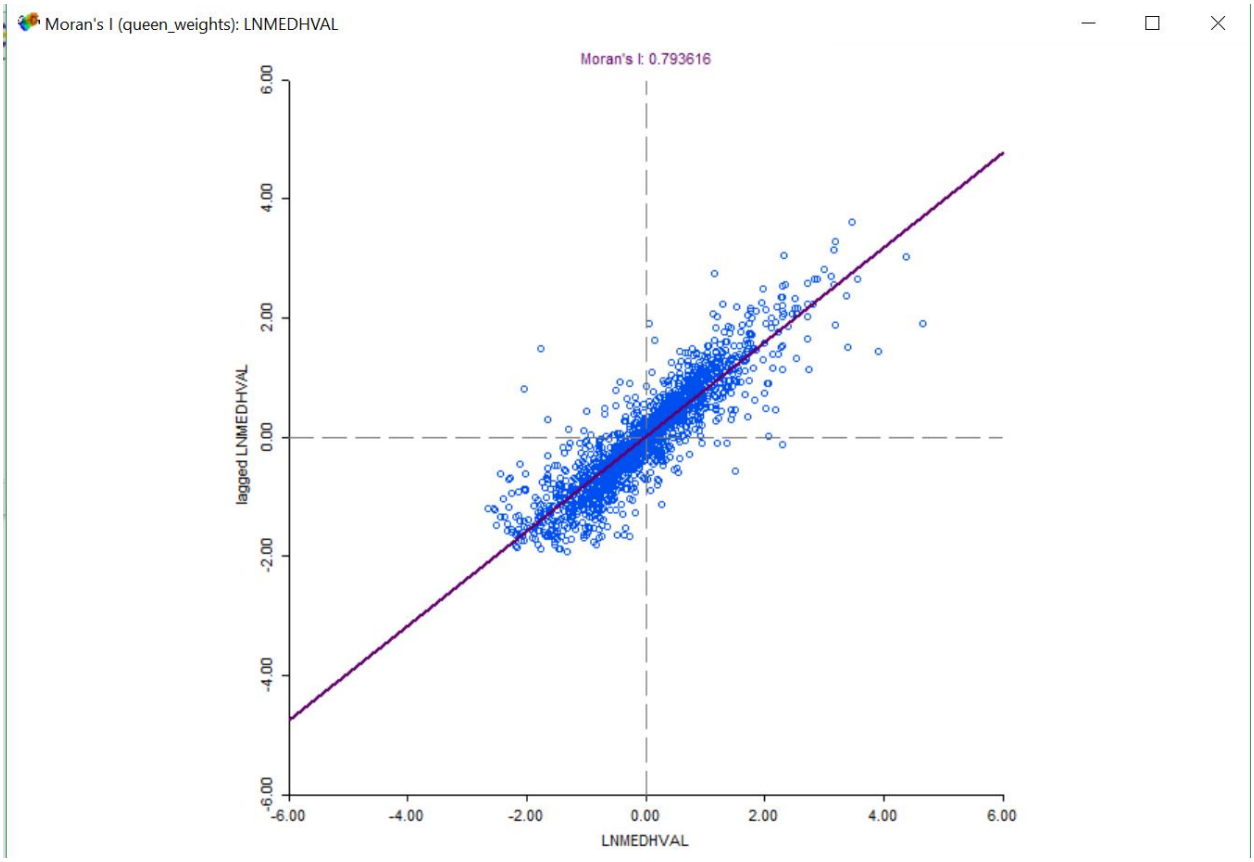
When assessing the strength of OLS models, a t-test is used for each parameter and the corresponding p-value indicates its significance. In GWR, there is a local model run for each observation, resulting in unique parameters for each model. The challenge of testing and interpreting p-values for each unique parameter for each local model makes this type of significance testing very inefficient. Instead, we compare the beta coefficient values (β) to their standard errors (SE) for each local regression (β/SE). Resulting values between -2 and 2 typically indicate non-significance, similar to p-

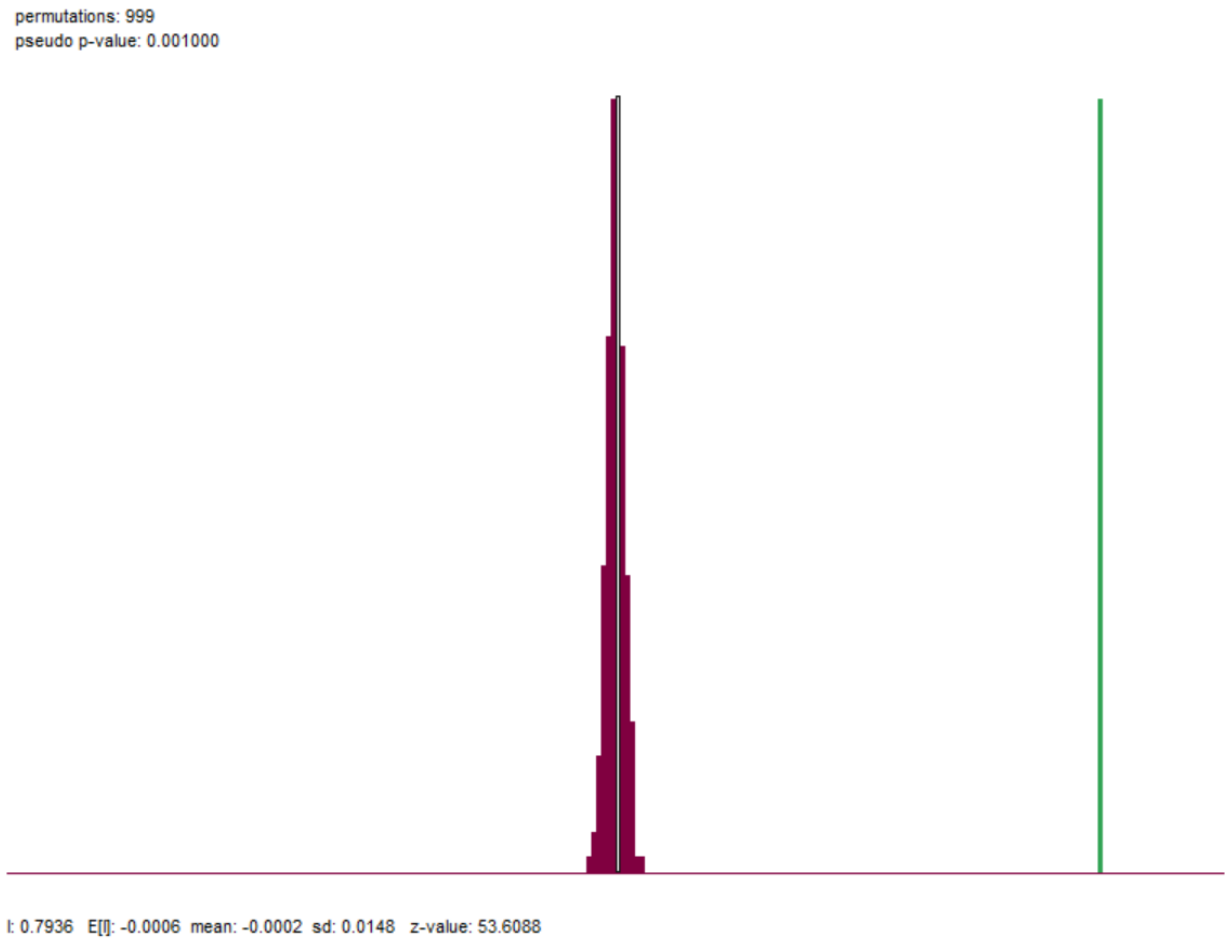
values > 0.05. While this information will be useful for local interpretation, this study focuses on the global interpretation and diagnostics for GWR. We will look at the Moran's I test for the residuals of the GWR regression, and use the Aikake Information Criteria (AIC) to compare the GWR results to our other regression models.

Results

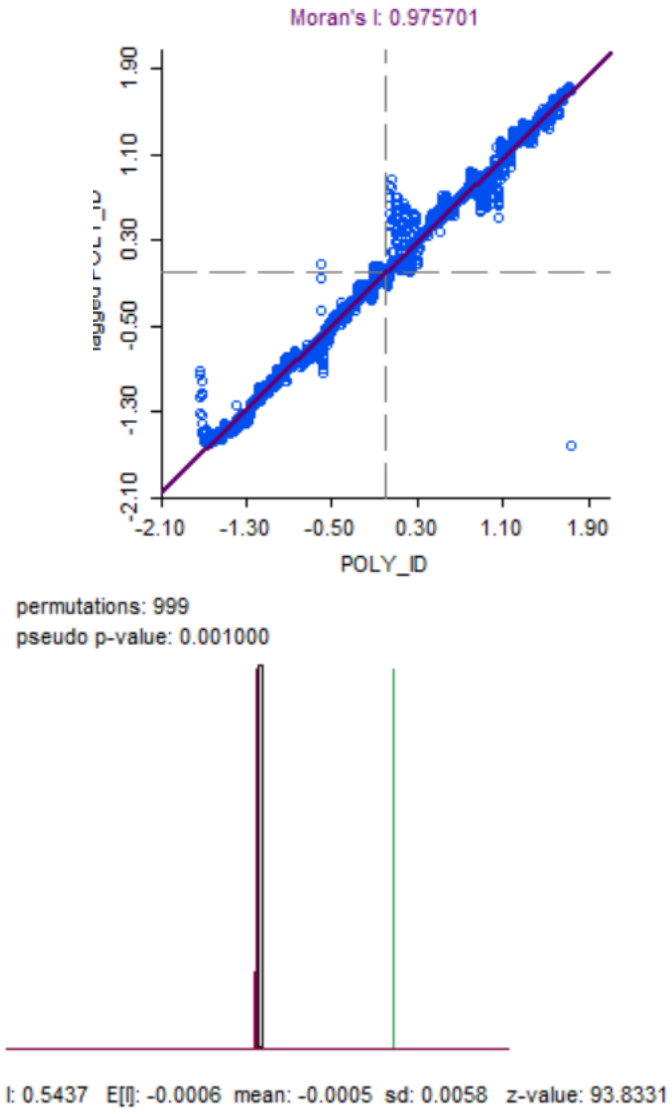
Spatial Autocorrelation

To assess the need for spatial analysis, we first look for spatial autocorrelation in the dependent variable LNMEDHVAL. To do so, we run a Moran's I test with a queen weight matrix and test for significance using 999 random permutations. Below are the results from both tests.

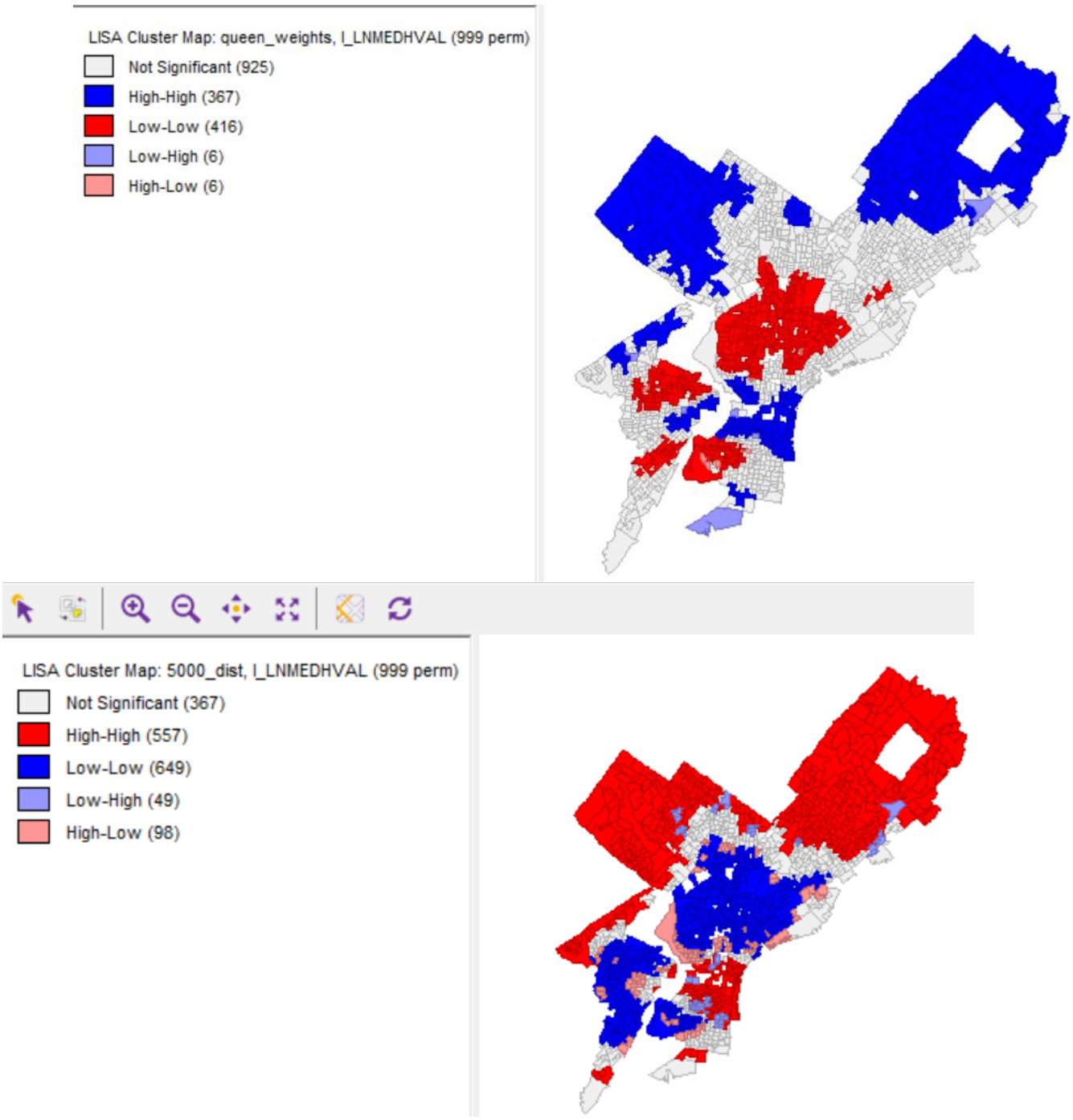

Moran's I (queen_weights): LNMEDHVAL
Moran's I: 0.793616
lagged LNMEDHVAL
LNMEDHVAL
6.00
4.00
2.00
0.00
-2.00
-4.00
-6.00
-6.00
-4.00
-2.00
0.00
2.00
4.00
6.00



The Moran's I value of 0.79 indicates strong positive spatial autocorrelation. Because the pseudo p-value is less than 0.05, we know this value is statistically significant and can reject the null hypothesis. We replicate these tests using a distance-based weights matrix for comparison. Below are those results and they confirm that spatial autocorrelation is a problem in our predicted variable.



We also perform a local Moran's I (LISA) test so we can analyze the positive or negative neighborhood autocorrelation for median house value. We present the results of the LISA tests using both weight matrices here (Top: Queen weights; Bottom: Distance weights):



Both maps indicate significant high-high areas (i.e. where house values are high and positively correlated) in four clusters: far northeast and northwest, Center City, and University City. Significant low-low clustering occurs in the center of the Philadelphia region and in small pockets in the southwest. The second map, which uses a distance weights matrix has many more low-high and high-low areas on the perimeter of the low-low and high-high clusters. However the queens weight matrix map has very few of these, so any conclusions about the low-high and the high-low areas can't be made.

A Review of OLS Regression and Assumptions: Results

Here are the results of running the OLS regression in GeoDa.

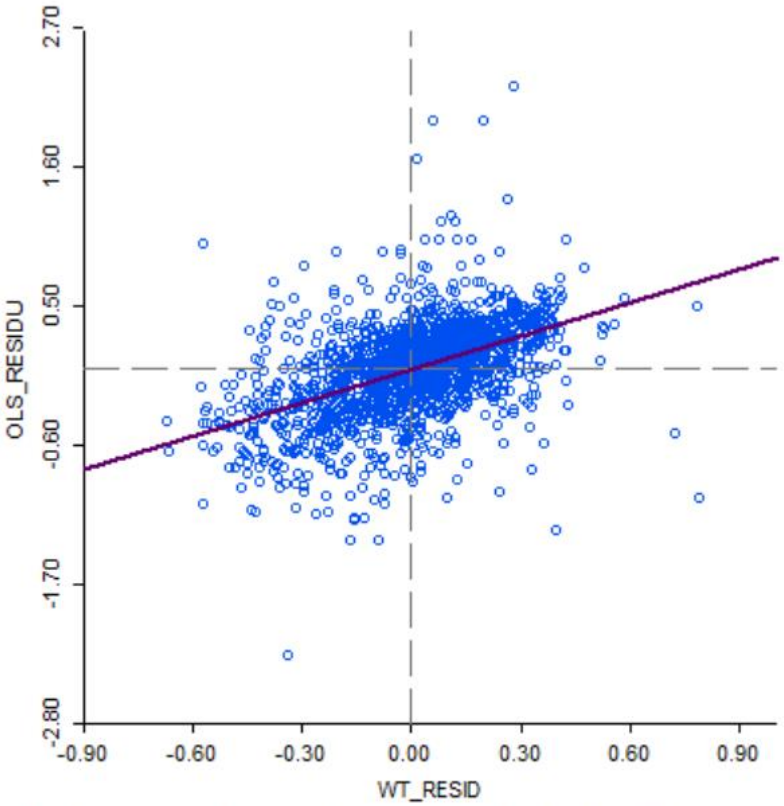
```
# SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
# Data set            :  Regression Data
# Dependent Variable  :   LNMEDHVAL  Number of Observations: 1720
# Mean dependent var  :      10.882  Number of Variables   :    5
# S.D. dependent var  :     0.62972  Degrees of Freedom    : 1715
#
# 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
#         PCTBACHMOR     0.0209095    0.000543184        38.4944     0.00000
#          PCTVACANT    -0.0191563    0.000977851       -19.5902     0.00000
#         PCTSINGLES    0.00297695    0.000703155        4.23371     0.00002
#          LNNBELPOV    -0.0789035      0.0084567        -9.3303     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)              0.3131        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
```

These results indicate that all four of our predictors---percentage of population with at least a bachelor's degree, percentage of vacant lots, percentage of single housing units, and the log of the number of people living below the poverty line---are all significant with the first two predictors having positive effects and the last two having negative effects. Together, these predictors account for approximately 66% of the variation in the log of median house values as given by the R^2 and adjusted R^2 in the model. The three tests for heteroscedasticity are consistent with each other and each return back a p-value of 0, indicating that we can reject the null hypothesis of homoscedasticity and have reason to believe that there is heteroscedasticity in the residuals.

The Jarque-Bera test for normality also shows a significant p-value of 0, which tests the joint null hypothesis that skewness is 0 and the kurtosis is 3, leading us to reject the null hypothesis and indicating that the residuals are not normally distributed.

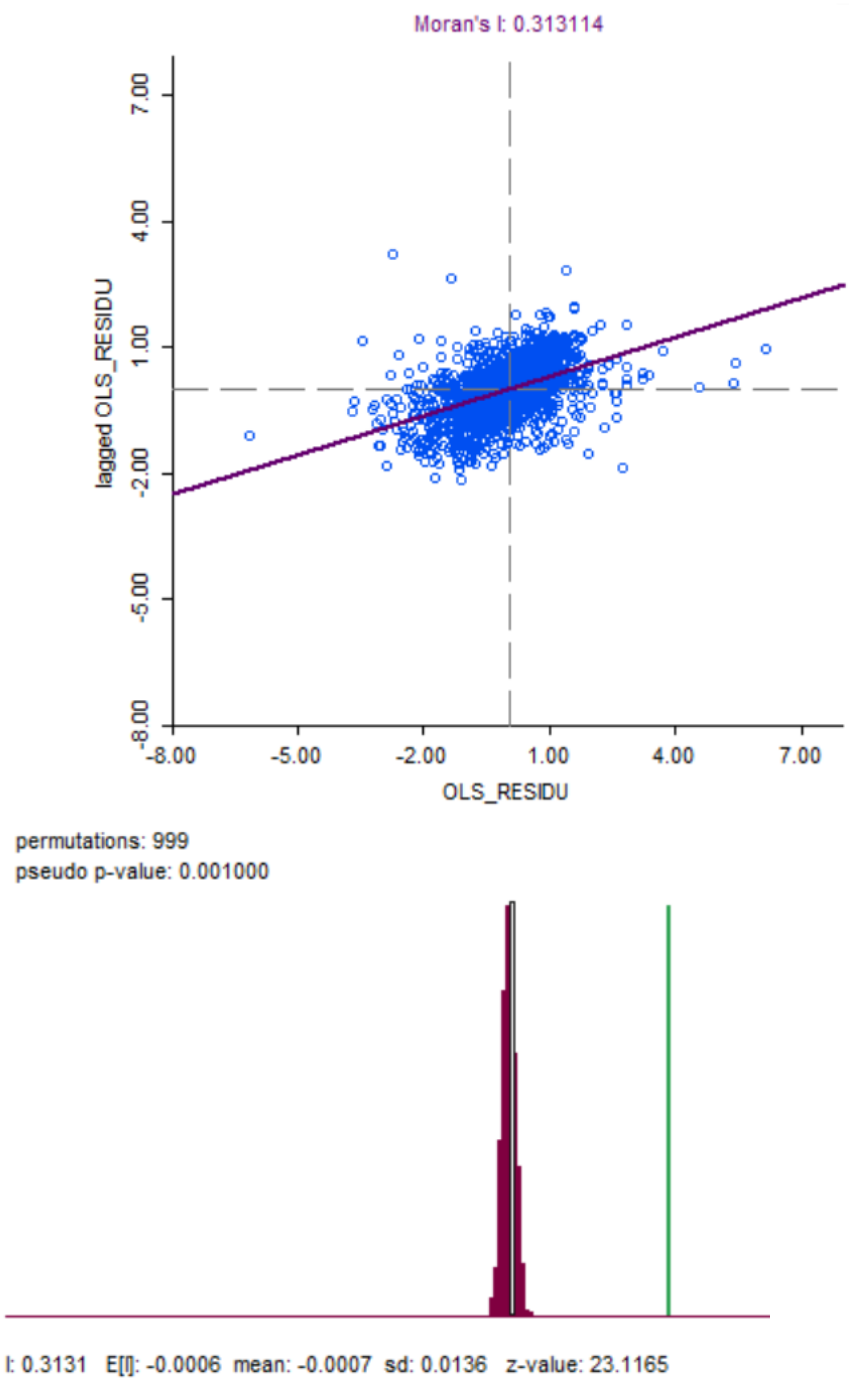
Next we analyze the spatial autocorrelation between residuals by creating scatterplots of the OLS residuals vs. the weighted residuals using both weight matrices



#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	0.234	-0.004	0.008	-0.545	0.586	0.883	0.039	22.888	0.000
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1720	0.234	-0.004	0.008	-0.545	0.586	0.883	0.039	22.888	0.000

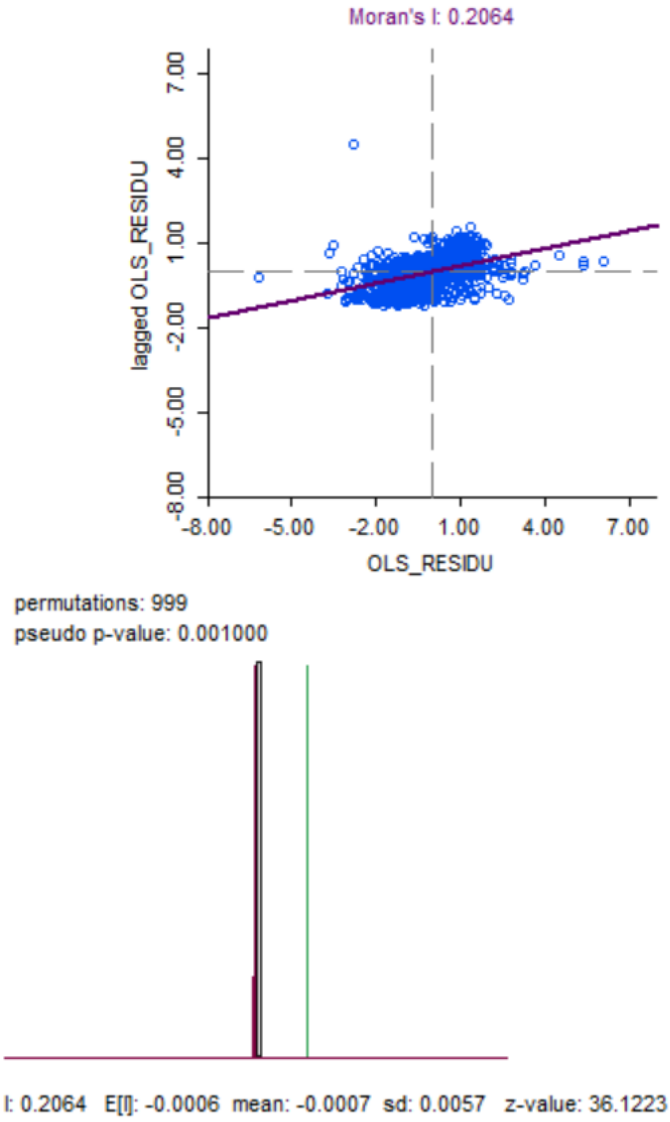
Chow test for sel/unsel regression subsets: need two valid regressions

We see that ρ (slope b in the above graph) has a value of 0.88 and is significant with an associated p-value of essentially 0. This indicates significant spatial autocorrelation. To confirm this, we also present the Moran's I scatterplots and random permutation tests of the OLS Residuals using both weights matrices.



We see that the Moran's I value is 0.31 and the permutation test returns a pseudo p-value of .001, meaning we reject the null of no spatial autocorrelation. All of the above graphs show us significant problems with spatial autocorrelation. This is problematic because one of the assumption of OLS regression, namely the independence of observations, has been violated. The value of LNMEDHVAL in one block group relies on the value of LNMEDHVAL in nearby block groups.

As a final confirmation, we re-run the tests for spatial autocorrelation in OLS residuals using our alternate weight matrix. Below are the scatterplots of OLS residuals vs Weighted residuals as well as the Moran's I plots using our distance weight matrix.



So the Moran's I score is slightly lower at 0.206 but the pseudo p-value of the permutation test is still 0.001, which confirms our results that the null hypothesis of no spatial autocorrelation is rejected.

Spatial Lag Regression Results

Below are the results of running a spatial lag model on our data using a queen weight matrix.

```
# SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
# Data set            : Regression Data
```

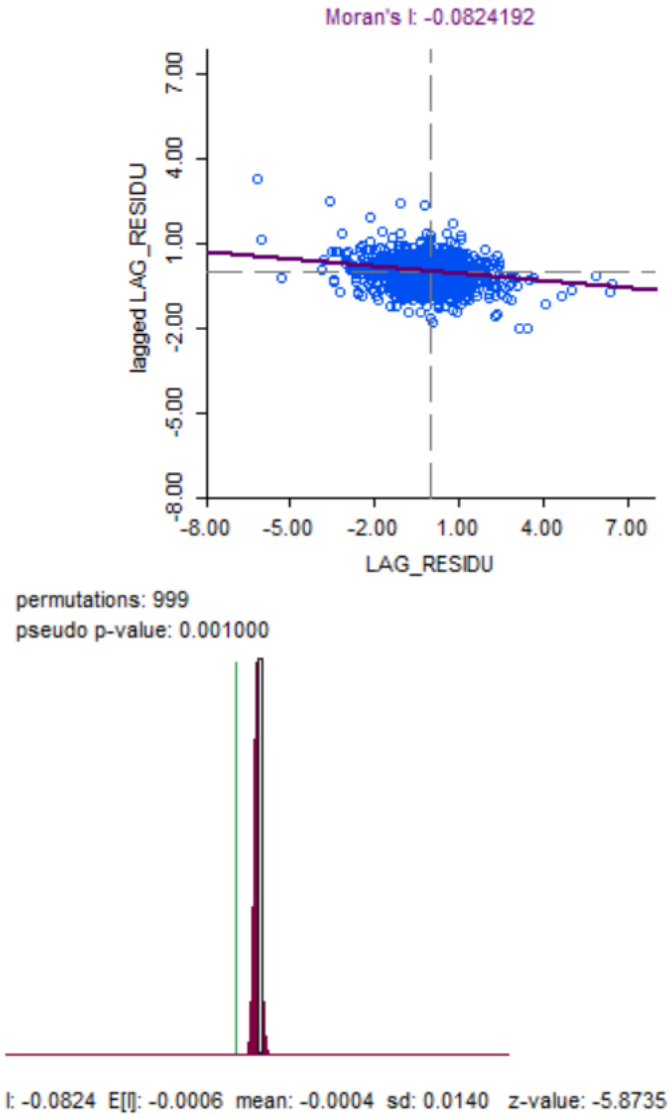
```
# Spatial Weight      : queen_weights
# Dependent Variable  :   LNMEDHVAL  Number of Observations: 1720
# Mean dependent var  :      10.882  Number of Variables   :    6
# S.D. dependent var  :     0.62972  Degrees of Freedom    : 1714
# Lag coeff.   (Rho)  :    0.651107
#
# R-squared           :    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
#         PCTBACHMOR    0.00851569     0.00052192        16.3161     0.00000
#          PCTVACANT   -0.00852676     0.00074357       -11.4673     0.00000
#         PCTSINGLES    0.00202905     0.00051571        3.93448     0.00008
#          LNNBELPOV    -0.0340632     0.00629222       -5.41355     0.00000
# ------------------------------------------------------------------------
--
#
# REGRESSION DIAGNOSTICS
# DIAGNOSTICS FOR HETEROSKEDASTICITY
# RANDOM COEFFICIENTS
# TEST                                    DF      VALUE        PROB
# Breusch-Pagan test                      4       220.5298     0.00000
#
# DIAGNOSTICS FOR SPATIAL DEPENDENCE
# SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : queen_weights
# TEST                                    DF      VALUE        PROB
# Likelihood Ratio Test                   1       911.8633     0.00000
```

We see that the coefficient W_lnmedhval, which is the average weighted value of LNMEDHVAL among a block groups neighbors, has a very high value of 0.65, the highest of all predictors. It also has an associated p-value of 0, meaning its statistically significant. This means that for every unit increase in the LNMEDHVAL of surrounding block groups, the LNMEDHVAL for that block group goes up by 0.65 units.

We can also see that the remaining predictors are still significant at a 5 percent confidence level. Comparing the beta coefficients in the spatial regression to the OLS regression, we see that the Betas are now all smaller in absolute value. For example, the beta coefficient on PCTVACANT went from -0.019 to -0.008. In essence, more of the

variation has been explained by the lagged LNMEDHVAL term and so the impact of other regressors on our predicted variable LNMEDHVAL has diminished.

In terms of evaluating heteroscedasticity, we can look at the results of the Bresuch-Pagan test on the spatial lag regression residuals. That test returns a p-value of essentially 0, meaning we reject the null hypothesis of homoscedasticity at any reasonable confidence level. However, it is clear that the spatial lag model is much better than the simple OLS model when it comes to in sample fit. The Akaike Information Criterion and the Schwarz Criterion are lower by approximately 1000 for the spatial lag model, indicating much better in sample fit. The spatial lag Log Likelihood is higher (less negative) than the OLS Log Likelihood, again indicating a better model fit. The Likelihood Ratio test also returns a test statistic of 911.86 and an associated p-value of 0, which means we reject the null hypothesis that the spatial lag model does just as good as the simple OLS model and conclude that the spatial lag model is doing a better job than the OLS model. Below is the Moran's I scatterplot and permutation test on the spatial lag regression residuals.

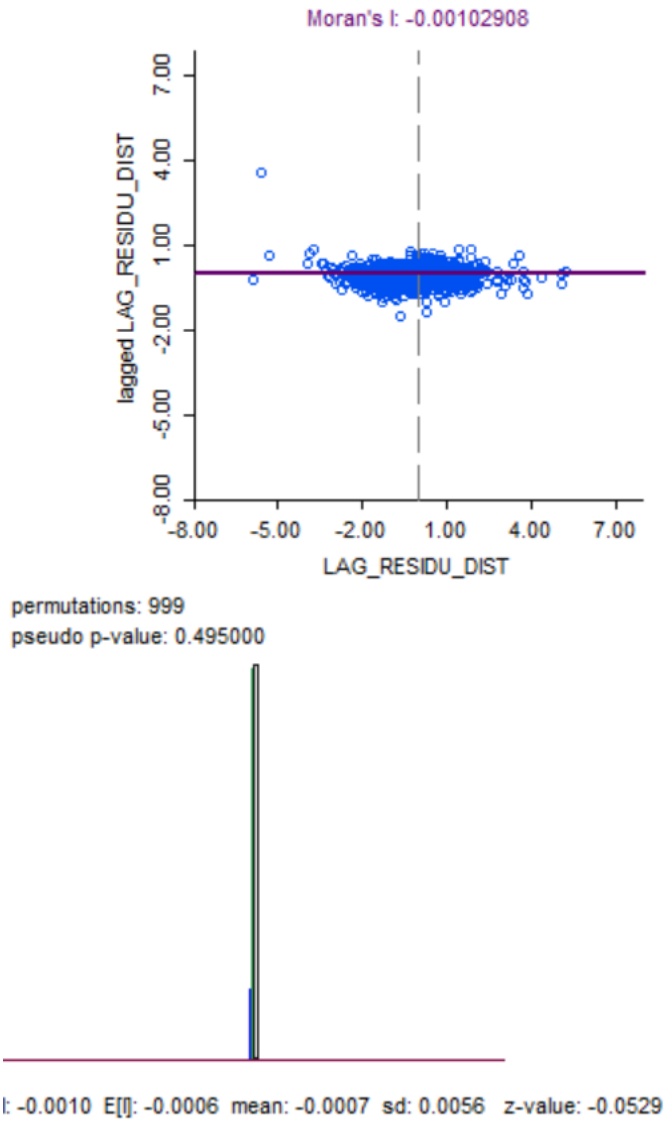


As we can see, there is much less spatial autocorrelation present in this model with a Moran's I value of only 0.008. The permutation test still returns a pseudo p-value of 0.001, suggesting that the null hypothesis of no spatial autocorrelation in the residuals is rejected. So based on all of the criteria including the AIC, the SIC, log likelihood, likelihood ratio test, and the Moran's I value, the spatial lag model is much better than the OLS model.

Additionally, we replicate all the results using our distance based weights matrix. Below are the Regression outputs, and the Moran's I of spatial lag regression using the alternate weight matrix.

```
# SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
# Data set            : Regression Data1
# Spatial Weight      : 5000_dist
# 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)  :    0.640801
#
# R-squared           :    0.772188  Log likelihood        :     -386.19
# Sq. Correlation     : -            Akaike info criterion :      784.38
# Sigma-square        :   0.0903383  Schwarz criterion     :     817.081
# S.E of regression   :    0.300563
#
# ------------------------------------------------------------------------
--
#        Variable       Coefficient     Std.Error       z-value
Probability
# ------------------------------------------------------------------------
--
#        W_LNMEDHVAL      0.640801      0.0213072        30.0744     0.00000
#           CONSTANT       4.01011       0.238102         16.842     0.00000
#          PCTBACHMOR     0.0131799    0.000539605         24.425     0.00000
#           PCTVACANT    -0.0110148    0.000823039        -13.383     0.00000
#          PCTSINGLES   0.000866634    0.000583582         1.48503     0.13754
#           LNNBELPOV    -0.0380291     0.00708702       -5.36603     0.00000
# ------------------------------------------------------------------------
--
#
# REGRESSION DIAGNOSTICS
# DIAGNOSTICS FOR HETEROSKEDASTICITY
# RANDOM COEFFICIENTS
# TEST                                    DF      VALUE        PROB
# Breusch-Pagan test                      4       253.8165     0.00000
#
# DIAGNOSTICS FOR SPATIAL DEPENDENCE
# SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : 5000_dist
# TEST                                    DF      VALUE        PROB
# Likelihood Ratio Test                   1       650.6065     0.00000
```



These results for the beta coefficients are mostly consistent with the other weight matrix. The only tangible difference between the two spatial lag regression models is that the beta coefficient on PCTSINGLES is now no longer statistically significant with a p-value of 0.22416. Furthermore, the effect of the beta's seem to have shifted slightly but the signs on all of them remain the same. Where things differ between the spatial lag models is that this one with the distance weight matrix actually has a much lower Moran's I score of -0.001 (compared to -0.08) and a pseudo p-value 0.49, suggesting that the null hypothesis is accepted and there is no longer spatial autocorrelation among the residuals!

Spatial Error Regression Results

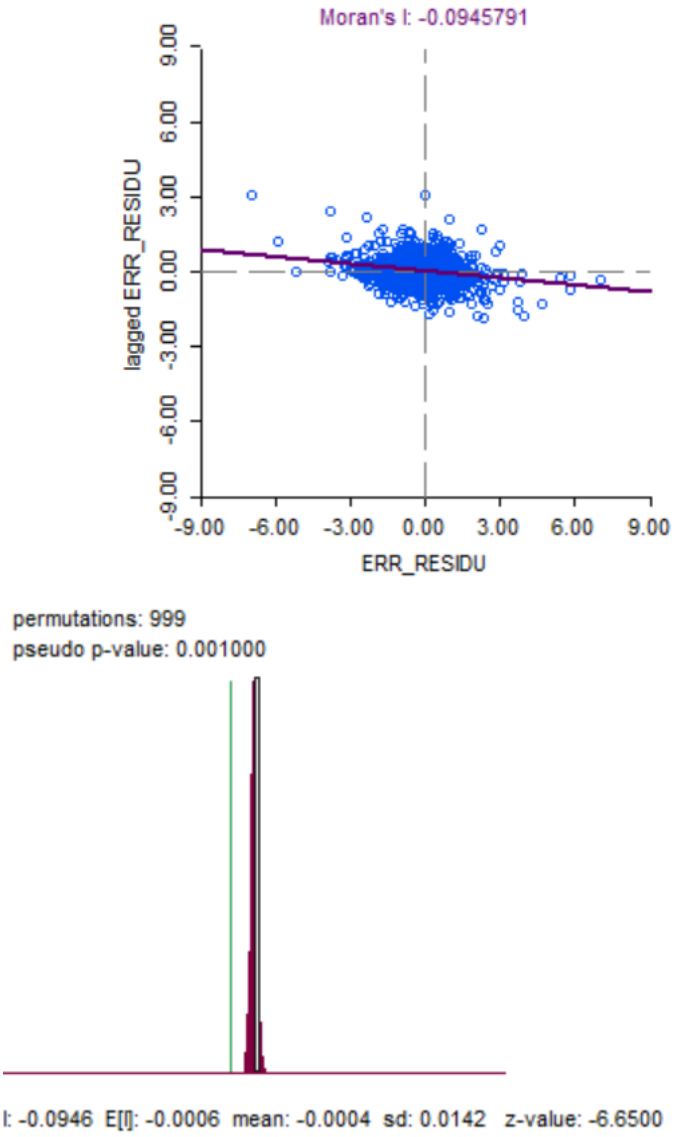
Now we present the results of the spatial error Regression. Below is the regression output using the queens weight matrix.

```
# SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
# Data set            : Regression Data1
# 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) :    0.814872
#
# 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
#         PCTSINGLES    0.00266586    0.000620803        4.29421     0.00002
#          PCTVACANT   -0.00577991    0.000886626         -6.519     0.00000
#         PCTBACHMOR    0.00982427    0.000728944        13.4774     0.00000
#          LNNBELPOV    -0.0345369     0.00708851       -4.87224     0.00000
#             LAMBDA      0.814872      0.0163744         49.765     0.00000
# ------------------------------------------------------------------------
--
#
# REGRESSION DIAGNOSTICS
# DIAGNOSTICS FOR HETEROSKEDASTICITY
# RANDOM COEFFICIENTS
# TEST                                    DF      VALUE        PROB
# Breusch-Pagan test                      4       211.1640     0.00000
#
# DIAGNOSTICS FOR SPATIAL DEPENDENCE
# SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : queen_weights
# TEST                                    DF      VALUE        PROB
# Likelihood Ratio Test                   1       678.0016     0.00000
```

We can see that the lambda in the regression is statistically significant and has a value of 0.81. What this means that the OLS residuals can be decomposed into two parts: 81% have a spatial pattern and 29% is actually white noise. We can say that as the spatially lagged residuals increase by one unit, the predicted variable LNMEDHVAL increases by one unit. (ASK EUGENE QUESTION ABOUT THIS- can you put this as %?

Is this accurate?) The remaining terms are all significant as well and follow the same pattern as the OLS regression results, albeit with a little bit smaller effects in absolute value. As an example, the beta coefficient on PCTVACANT went from -0.019 to -0.009. So we're seeing the same effect as in the spatial lag model where more of the variation is being explained by the Lambda term thus reducing the absolute value of the beta coefficients on other predictors.

After looking at the results of the Breusch-Pagan test, it is clear that heteroscedasticity in the spatial error regression residuals is still a problem. The test gives us a test statistic of 211.16 and an associated p-value of 0, leading us to reject the null of homoscedasticity. However, it is clear that the spatial error model has much better in sample fit. The AIC and the SIC are 754 and 784 respectively, which is much lower than the 1432 and 1460 under our OLS regression. The log-likelihood is this model is -372, which is less negative than the log-likelihood under OLS which was -711.493. And finally, the likelihood ratio test returns a p-value of 0, meaning we reject the null hypothesis that the spatial error model does just as good as the simple OLS model and conclude that the spatial error model is actually doing a better job than the OLS model. We finally take a look at the Moran's I scatterplot and permutation test of spatial error regression residuals.

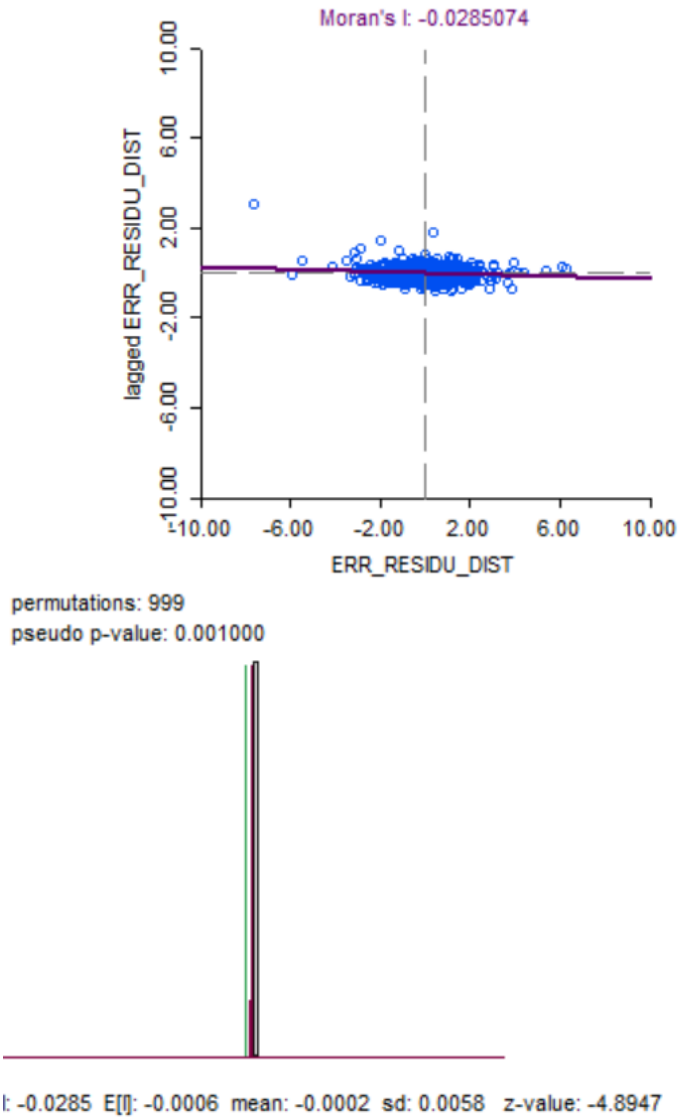


The Moran's I value is -0.09, which is fairly small. The pseudo p-value on the permutation test is 0.001, which means we still reject the null of no spatial autocorrelation in the spatial error residuals. Compared to the Moran's I value of 0.31 for the OLS residuals, this is much closer to 0. Thus spatial autocorrelation is smaller in the spatial error model, but still statistically significant. Overall, the spatial error model outperforms the OLS model when comparing the AIC, SIC, log-likelihood, LRT, and Moran's I value.

Now we compare the spatial lag and the spatial error models. Only the AIC and SIC are valid diagnostics, as neither method is a special subtype of the other. In the spatial error model the AIC is 754.985 while the SIC is 782.235. In our spatial lag model, the AIC is 523.123 and the SIC is 555. 824. Thus, the spatial lag model seems to have better in sample fit than the spatial error model.

And finally to confirm our results, we use the alternate weight matrix and replicate the spatial error Regression and re-run the tests for spatial autocorrelation in the residuals. The results are below.

```
# SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
# Data set            : Regression Data1
# Spatial Weight      : 5000_dist
# 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) :    0.915990
#
# R-squared           :    0.771445  R-squared (BUSE)      : -
# Sq. Correlation     : -            Log likelihood        : -414.175389
# Sigma-square        :   0.0906329  Akaike info criterion :     838.351
# S.E of regression   :    0.301053  Schwarz criterion     :     865.601
#
# ------------------------------------------------------------------------
--
#        Variable       Coefficient     Std.Error       z-value
Probability
# ------------------------------------------------------------------------
--
#           CONSTANT       11.0013      0.0956271        115.044     0.00000
#         PCTSINGLES    0.00134956    0.000660583        2.04298     0.04105
#          PCTVACANT   -0.00788087    0.000943771       -8.35041     0.00000
#         PCTBACHMOR     0.0166038    0.000653451        25.4093     0.00000
#          LNNBELPOV    -0.0451571     0.00751176       -6.01151     0.00000
#             LAMBDA       0.91599      0.0196697        46.5686     0.00000
# ------------------------------------------------------------------------
--
#
# REGRESSION DIAGNOSTICS
# DIAGNOSTICS FOR HETEROSKEDASTICITY
# RANDOM COEFFICIENTS
# TEST                                    DF      VALUE        PROB
# Breusch-Pagan test                      4       195.4631     0.00000
#
# DIAGNOSTICS FOR SPATIAL DEPENDENCE
# SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : 5000_dist
# TEST                                    DF      VALUE        PROB
# Likelihood Ratio Test                   1       594.6359     0.00000
```



While the results are mostly consistent, there are a few differences when compared to the Queen weight matrix. In particular, the lambda increased from 0.81 to 0.91 and the coefficients on the beta's moved around a little bit. The in sample fit metrics are also a little worse using this distance weights matrix with and AIC and SIC that is approximately bigger by +75 when compared to the spatial error model with the queen weight matrix. The Moran's I score decreases from -0.09 to -0.02, but the pseudo p-value on the permutation test is still 0.01 meaning we still reject the null of homoscedasticity in the spatial error regression residuals.

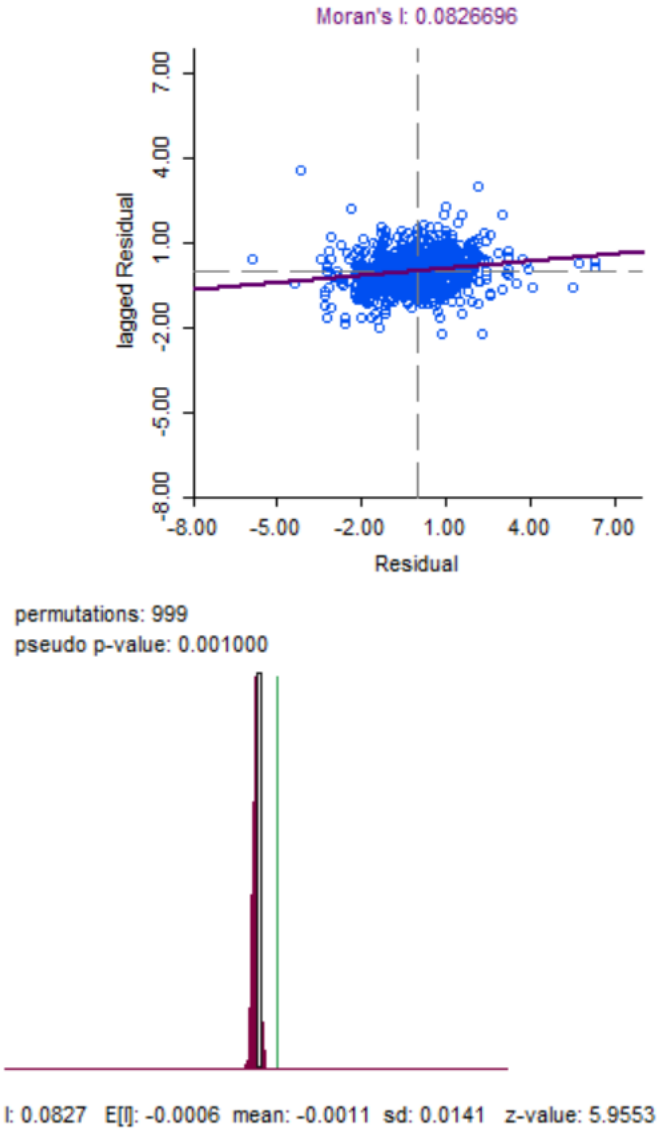
Geographically Weighted Regression Results

Below is the supplementary table from ArcGIS for the geographically weighted regression.

GWR_regression_data_1_supp

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

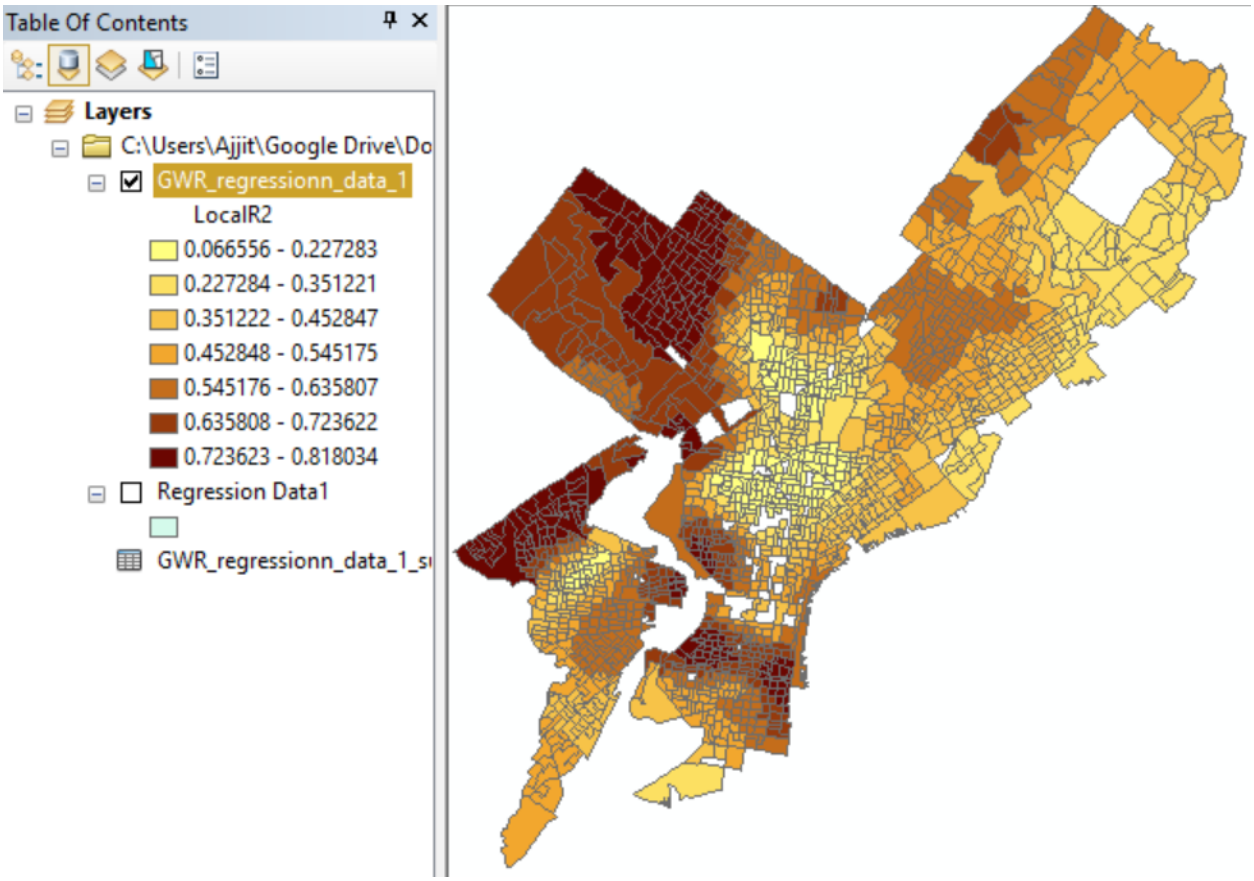
The overall R^2 is smaller in the GWR model as compared to the OLS model (0.81 vs 0.66). Thus, the GWR model is doing a better job of explaining the variance in the predicted variable. Furthermore, the SIC of the GWR model is 668, which means it comes in second place when compared to all our other models. In particular, only the spatial lag model with a queen weight matrix has a lower AIC and SIC of 523 (That model also has the same R^2 of 0.81). Below we also provide the Moran's I scatterplot and permutation test of GWR residuals.



So we see that the Moran's I value is 0.08, which indicates much less spatial autocorrelation in the residuals than OLS, which had a Moran's I value of around 0.3. The pseudo p-value of the permutation test is still 0.001 so the null of no spatial autocorrelation is still rejected.

Next we compare the Moran's I value of this GWR regression to the other spatial regressions that used queen matrices. In comparison to the spatial lag regression, the GWR had marginally higher Moran's I in absolute value (+0.08266 vs -0.0824. In comparison to the spatial error model, the GWR model had a marginally lower Moran's I in absolute value (+0.08266 vs -0.0945). Interestingly enough, the GWR regression is the only case where there is a positive Moran's I value, indicating positive spatial autocorrelation between residuals.

Finally, we present a choropleth map of the local R-squared values for each block group given to us in the GWR regression.



This map shows us that in places like Northwest Philly, South Philly, University City, the GWR model does a really good job and can accurately predict a lot of the variation in our predicted variable LNMEDHVAL. These areas have R^2 values between 0.6 and 0.81. In other parts of Philly such as Northeast and North Philly, our model falls short and doesn't do as good a job, having R^2 values between 0.06 and 0.5.

Discussion

In this study, we regressed the natural log of median house value (LNMEDHVAL) on four predictors using OLS, three methods of spatial analysis discussed above, and compare the results. The predictors are: percentage of vacant lots (PCTVACANT), the natural log of the number of people in poverty (LNNBELPOV), percentage of people with bachelors degrees (PCTBACHMOR), and percentage of single housing units (PCTSINGLES). After running all 4 models with queen weight matrices, we see that spatial autocorrelation is much smaller in the spatial regression models as compared to the OLS models. The GWR and Spatial Lag regression have the

smallest Moran's I values of -0.08 and +0.08 respectively. However, the Spatial Log model has an AIC of 523.123 which is the lowest among all models. Thus the Spatial Lag model has the best in sample fit and one of the lowest degrees of spatial autocorrelation in the residuals, making it the best model.

These findings should be taken with a grain of salt as all models still suffer from statistically significant spatial autocorrelation in the residuals, meaning the assumption of independence of observations has been violated. Furthermore, all of the models had non-normal distribution of residuals as determined by the Jarque-Bera test, suggesting that the assumption of normality of regression residuals had been violated. However in comparison to the naive OLS model, the Spatial Lag model outperforms it on every metric and is better able to explain the variation in the LNMEDHVAL variable. In particular, we still see that the percent of vacant lots and the log of the number of people living in poverty have negative effects and the percent of single unit homes and percent of people with bachelors degrees have positive effects on the log of the median house value for a block group. However after introducing a spatially lagged LNMEDHVAL variable in the Spatial Lag Regression, the β's have gone down in absolute value. In effect, the lagged LNMEDHVAL variable has soaked up some of the variation in the LNMEDHVAL variable and have reduced the impact of the other coefficients. In the future, it would be helpful to find more predictors that would allow us to explain away the spatial autocorrelation that still exists.