Haobing Liu, Jack Rummler, Kate Tanabe

MUSA 500

9 November 2022

Spatial Lag, Spatial Error, and Geographically Weighted Regression in Philadelphia's block group average household income

Introduction

In this report, we are continuing the investigation of the relationship between median home values and several neighborhood characteristics in Philadelphia. We are specifically looking at owner-occupied housing units in Philadelphia and using data at the census block group level. Our dependent variable is median value of all owner-occupied housing units, and our predictors are the proportion of residents with at least a bachelor's degree, the proportion of housing units that are vacant, the precent of housing units that are detached single family houses, and the number of households with incomes below 100% of the poverty level within each block group.

Our last report used Ordinary Least Square (OLS) regression to examine this relationship, but this approach is often inappropriate for data sets like ours that include spatial components. To account for this, we will run spatial lag, spatial error, and geographically weighted regression on our data to examine if these methods can account for the spatial autocorrelation that exists with OLS residuals.

Methods

a) Description of the Concept of Spatial Autocorrelation

A key premise in spatial statistics is the First Law of Geography. This law describes the concept of spatial autocorrelation, which Waldo Tobler states is the theory that "everything is related to everything else, but near things are more related than distant things." We will use Moran's I to understand if spatial autocorrelation is present without our data. Moran's I is a method for testing for spatial autocorrelation which looks at the covariance of the variables at nearby locations and standardizes this value by the variance in the variable. The equation for Moran's I is presented below where $\bar{X}$ is the mean of variable X, X_i is the value of variable X at a particular location *i,* X_j is the variable value at another location *j,* W_{jj} is the weight indexing location of *i* relative to *j,* and *n* is the number of observations.

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

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

As mentioned above, the equation for Moran's I includes a weight indexing location which is based on the weight matrix. This allows us to identify and account for spatial proximity of block groups. We will use the queen weight matrix, a contiguity-based measure of proximity, throughout this report. The queen neighbor approach identifies neighbors as those that intersect at either a point or a segment. In our case,

this means that neighbors are considered block groups that share a point or segment of a border with another block group. Some statisticians like to use and try more than one spatial weight matrix in analyses as it helps to ensure that results are not solely determined by the matrix choice.

We will identify if our spatial autocorrelation values are significant by testing the following three hypotheses:

$$H_0: No\ spatial\ autocorrelation$$

$$H_{a1}:\ Positive\ spatial\ autocorrelation$$

$$H_{a2}:\ Negative\ spatial\ autocorrelation$$

First, we will calculate Moran's I values on our variable for each block group, shuffle the values around on our map of Philadelphia block groups 999 times, then organize all 1,000 Moran's I values in descending order (or ascending order if the Moran's I value is negative). We will then identify where the original Moran's I value is within the entire list of all permutations. We will then calculate the pseudo p-value by dividing the rank of the original Moran's I by 1,000. If this value is below 0.05, we will reject the null hypothesis, H_0, that spatial autocorrelation is not present.

Another aspect of spatial dependencies is local spatial autocorrelation, which aims to determine if and to what extent values at places in vicinity of the location *i* are associated with *i.* We can use both local and global Moran's I to understand if spatial autocorrelation is present in our dataset.

b) Review of OLS Regression and Assumptions

As part of our last report, we used OLS regression, which examines the strength and direction relationships between variables and breaks down the goodness of fit model. OLS regression calculates the amount by which the dependent variable changes when a predictor changes by one unit while holding any other predictors constant. As part of OLS, we make several model assumptions. We first assume there is a linear relationship between the dependent variable and each of the predictors. We also assume independence of our observations and residuals. Additionally, we assume our residuals are normally distributed. We also assume homoscedasticity of our residuals. The final assumption is that there is no multicollinearity.

However, when the data has spatial elements, as our dataset does, we often cannot assume that the errors are independent. We can test this assumption by examining the spatial autocorrelation of residuals using Moran's I. Another way to test OLS residuals for spatial autocorrelation is to regress the errors on nearby errors. In our case, we would regress the residual on residuals of neighboring block groups as defined by the queen matrix. The neighboring residual is known as the lagged residual or ρ (rho), also known as λ (lambda).

We are using GeoDa to run our regression. GeoDa has methods of testing the other regression assumptions. GeoDa has three ways to test for heteroscedasticity. These include the Breusch-Pagan Test, the Koenker-Bassett Test, and the White Test. Each of these tests use the hypotheses listed below. If the p-value is less than 0.05, then we will reject the null hypothesis, H_0, that there is no heteroscedasticity.

$$H_0: No\ heteroscedasticity\ present$$

$$H_a:\ Heteroscedasticity\ present$$

GeoDa also has the Jarque-Bera test which tests for normality of errors. This test examines the hypotheses presented below. Again, if the p-value is below 0.05 then we will reject the null hypothesis, H_0, that errors are normal.

$$H_0: Errors\ are\ normally\ distributed$$

$$H_a:\ Errors\ are\ not\ normally\ distributed$$

c) Spatial Lag and Spatial Error Regression

We will use GeoDa to also run the spatial lag and spatial error regressions. The spatial lag model assumes that the value of the dependent variable in one location is associated with the values of the dependent variable in locations nearby. Nearby locations are defined by the weight matrix, which in our case is the queen weight matrix. The model will include the spatial lag, known as ρ (rho), of the dependent variable as a predictor. The equation for the spatial lag model is presented below, where *ρ* is the coefficient of *Wy* which is the lag of the variable y, β_o is the intercept of the dependent variable, $\beta_1 X_1 \ldots \beta_n X_n$ represents the predictor variables and their coefficients, and ε is the errors.

$$Spatial\ Lag\ = \rho Wy + \beta_0 + \beta_1 PCBACHMORE + \beta_2 PCTVACANT + \beta_3 PCTSINGLES + \beta_4 NBELPOV100 + \beta_5 MEDHHINC + \varepsilon$$

The spatial error model assumes that the error in one location is associated with the errors at nearby locations. Again, nearby is defined by the weight matrix in use, which in this case is the queen weight matrix. To run a spatial error model, we first must run an OLS regression where we regress the dependent variable on the predictors then we regress the residuals on the nearest neighbor residuals. This will separate the residuals into two groups: one with the spatial component, *λWε,* and one which is random noise, u. The equation for the spatial error model is presented below, where β_o is the intercept of the dependent variable, $\beta_1 X_1 \ldots \beta_n X_n$ represents the predictor variables and coefficients, and *λ is the coefficient of Wε*, the spatially lagged residuals, and u is random noise.

$$Spatial\ Error\ = \lambda W\varepsilon\ + \beta_0 + \beta_1 PCBACHMORE + \beta_2 PCTVACANT + \beta_3 PCTSINGLES + \beta_4 NBELPOV100 + \beta_5 MEDHHINC\ + u$$

The previously mentioned assumptions needed for OLS are still needed for both spatial lag and spatial error regression, except for the assumption of spatial independence of observations. By using spatial lag and spatial error regression instead of OLS, we hope that the residuals are not spatially autocorrelated and less heteroscedastic. We will compare the results of both the spatial lag and spatial error regressions with those of OLS. We will look at several criteria to determine whether the spatial models perform better than OLS. These criteria include the Akaike Information Criterion (AIC), the Schwarz Criterion, the log likelihood, and the likelihood ratio test.

The AIC and Schwarz Criterion both measure the quality of a model. Both criteria require the use of at least two models as they compare the value of one model to that of another. In both cases, a better model

would have a smaller AIC and Schwarz Criterion value. The log likelihood identifies the maximum likelihood method of fitting regression models to the data. When looking at log likelihood, a better model would have a higher log likelihood value. Finally, the likelihood ratio test compares the OLS model with the spatial model. The likelihood ratio test uses the hypotheses presented below. If the p-value is less than 0.05, we will reject the null hypothesis, H_0, that the spatial lag model is not better than the OLS model.

$$H_0: Spatial\ lag\ and\ spatial\ error\ models\ are\ not\ better\ than\ the\ OLS\ model$$

$$H_a: Spatial\ lag\ and\ spatial\ error\ models\ are\ better\ than\ the\ OLS\ model$$

Another way to compare the OLS results with the spatial lag and spatial error results is to look at the Moran's I of regression residuals. For Moran's I, we look to see if we can fail to reject the null hypothesis, H_0, which is that there is no spatial autocorrelation present. We can only fail to reject the null hypothesis if the p-value is above a certain alpha level, which is typically 0.05. When comparing multiple models, we can decide if one is better if we can fail to reject the null hypothesis or if the p-value is notably greater than p-values of the other models.

d) Geographically Weighted Regression

We will conduct our geographically weighted regression (GWR) using ArcGIS. GWR builds on OLS and takes spatial non-stationarity into account. GWR allows us to run multiple local regressions instead of running a single, global regression. A local regression allows us to understand the relationships between dependent and predictor variables across space. It is applied to each of the locations in a dataset, which in this case is block groups. GWR allows us to disaggregate our data and look at it regionally, rather than globally, which helps us determine relationships on a regional level and understand if Simpson's Paradox is occurring. Simpson's Paradox is a phenomenon that can occur within datasets where relationships appear or change at different scales. For example, there may be a negative relationship between two variables when looking at the entire city but when we disaggregate the data into groups or regions, we may see that some areas have positive relationships.

We present the equation for GWR below, where β_{i0} represents the intercept term of the regression equation at location *i,* $\beta_{i1}X_{i1} \ldots \beta_{in}X_{in}$ represents the predictor variables at location *i,* and ε_i represents the errors at location *i.* It is critical to include the subscript *i* because it indicates that the regression model refers to the relationship between the predictors and the dependent variable only at the specific location, *i.*

$$y_i = \beta_{i0} + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{im}x_{im} + \varepsilon_i$$

$$= \beta_{i0} + \sum_{k=1}^{m} \beta_{ik}\, x_{ik} + \varepsilon_i$$

Local regression requires several observations to run. In this case, the observations are the block group locations. This is necessary because GWR uses other observations in the dataset when running its regression. GWR assigns weights to the other observations, which vary based on the given location *i.* An

observation closer to location *i* will have a greater weight in the regression and therefore a stronger influence on the estimation of the parameters for location *i.*

To identify which observations are close to a given location *i*, we will choose an appropriate bandwidth, or distance measure. Bandwidth allows us to determine if another observation should be considered a neighbor of the observation in question. There are two options when using bandwidth: fixed bandwidth and variable bandwidth. Fixed bandwidth uses a constant distance from the observation and assigns all the observations within that certain distance as a close neighbor. Fixed bandwidth may result in some observations having a greater or lesser number of neighbors than other observations based on the distribution of observations in space. Adaptive bandwidth uses a fixed number of observations, but the distance will not always be the same. This is because the distance necessary to travel from each observation to find the fixed number of neighbors may change based on the distribution of observation in space. In this way, adaptive bandwidth prioritizes having the same number of neighbors for each observation. We will use adaptive bandwidth in our analysis. This is more appropriate than fixed bandwidth because our observations are heterogeneously shaped polygons and are varied across space. If our dataset had a more even distribution of observations across space, then we would use a fixed bandwidth.

Many of the assumptions we use in OLS still hold in GWR except for the assumption of global multicollinearity. Because GWR uses local regression for each observation and feature in the dataset, the value of explanatory variables is often substantially spatially clustered, and we run into a problem with multicollinearity. This is also a problem when two or more of the predictor variables have similar clustered patterns. We will use a condition number, which measures the amount of multicollinearity. A condition number of over 30 or equal to zero means there is multicollinearity present.

GWR differs from other regression models like OLS because it does not produce a p-value as an output. GWR has one set of parameters and one set of standard errors for each regression point. This leads to a very large number of tests that would be required to determine significance. If we were to run significance tests on each of the regression points, we would expect a certain number of significant results based simply on chance. There are methods to conduct multiple testing, but they are not currently available in ArcGIS.

Methods

a) Spatial Autocorrelation

To determine spatial autocorrelation, we first look at the Global Moran's I. Below, a scatterplot of Global Moran's I using queen weight matrix is presented.

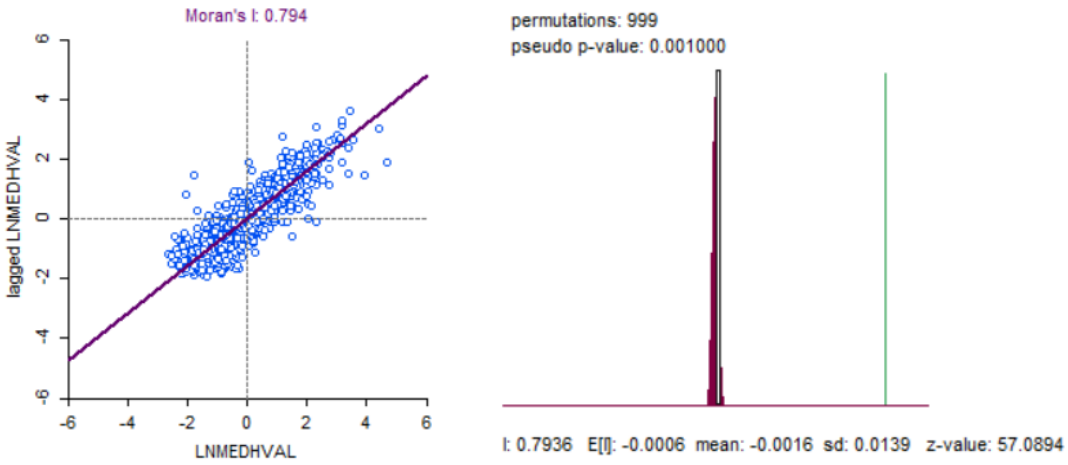

Figure 1: Global Moran's I and Significance Test (LNMEDHVAL)

We interpret this scatter plot as how many standard deviations a block group is from LNMEDHVAL from its queen neighbors. The Global Moran's I value is 0.794, thus we can reject H_0 in favor of H_{a1}, that there is positive spatial autocorrelation.

Next, we completed significance testing using 999 permutations. The histogram below shows the 999 permutations, where the green line represents our Moran's I value of 0.794. The pseudo p-value is calculated by ranking the 1,000 Moran's I values in descending order, identifying the rank of the actual Moran's I value and dividing it by 1,000 (Rank of 1 / 1000 = 0.001). Because our pseudo p-value is < 0.05, we reject H_0 that there is no spatial autocorrelation.

Local Moran's I is another way to determine spatial autocorrelation in which we use Local Indices of Spatial Autocorrelation (LISA). Local Moran's I is a measure of how similar location (i) is to its queen neighbors (j), or more simply stated, a measure of similarity to nearby observations.

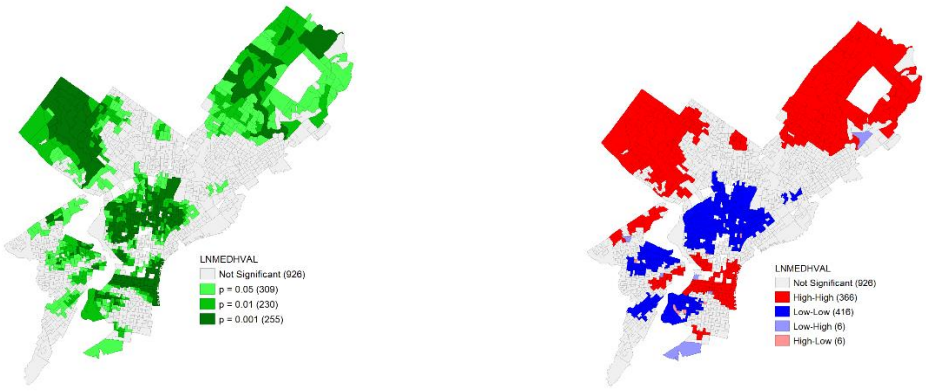

Figure 2: Local Moran's I Outputs

The outcomes for our LISA statistic can either be classified as (1) High-High (High X_i and High X_j with positive spatial autocorrelation); Low-Low (Low X_i and Low X_j with positive spatial autocorrelation); (3)

Low-High (Low X_i and High X_j with negative spatial autocorrelation); (4) High-Low (High X_i and Low X_j with negative spatial autocorrelation); (5) not significant. Given this is the median household value, the clusters make sense. Our high-high clusters are in Northwestern and Northeastern Philadelphia approaching the suburbs, as well as the Center City, Old City, and University City areas. Our low-low clusters are in North Philadelphia and West Philadelphia. These make sense given the economic conditions of these regions. Our low-high and high-low areas are more sporadic but are found mostly in Center City, South Philadelphia, and West Philadelphia, where there is more diversity in economic conditions.

b) A Review of OLS Regression and Assumptions: Results

We revisited OLS regression from our previous report, where we regressed LNMEDHVAL on PCTBACHMORE, PCTSINGLES, PCTVACANT, and LNBELPOV100. The table of our findings is presented below.

Table 1: OLS Regression Table

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

All four of our indicators are highly significant. LNNBELPOV and PCTVACANT are negatively correlated with LNMEDHVAL whereas PCTBACHMOR and PCTSINGLES are positively correlated with LNMEDHVAL. Our R-squared is 0.66, indicating that 66% of the variance in LNMEDHVAL has been explained by the model.

We have several assumptions when completing OLS regression. First, we assume homoscedasticity, that the variance of the residuals is constant. Next, we assume the normality of our residuals. Third, we assume observations are independent of each other, meaning there should be no spatial dependencies within the data.

We check for homoscedasticity through the Breusch-Pagan test, the Koenker-Bassett test, and the White test. If the p-value < 0.05, then we reject H_0 that there is no heteroscedasticity. We use the Jarque-Bara test to examine normality of our errors. If the p-value < 0.05, then we reject H_0 that there is normality of our errors. When testing spatial autocorrelation, we will again use the Moran's I test and the significance test.

We first tested for homoscedasticity, where the Breusch-Pagan, the Koenker-Bassett, and the White tests all produced p-values of 0.00. All three tests have a p-value < 0.05, thus we reject H_0 that there is no

heteroscedasticity. Heteroscedasticity is important because our residuals should just be random noise. This is the first problem we encounter with OLS.

Next, we used the Jarque-Bera test to examine the normality of the distribution of errors. The result of the test was a p-value < 0.001, meaning we reject H_0 that errors are normal, thus indicating an issue with normality of residuals.

After, we analyze our OLS residuals against our weighted residuals (the queen neighbors from earlier). The scatterplot of this is presented below.

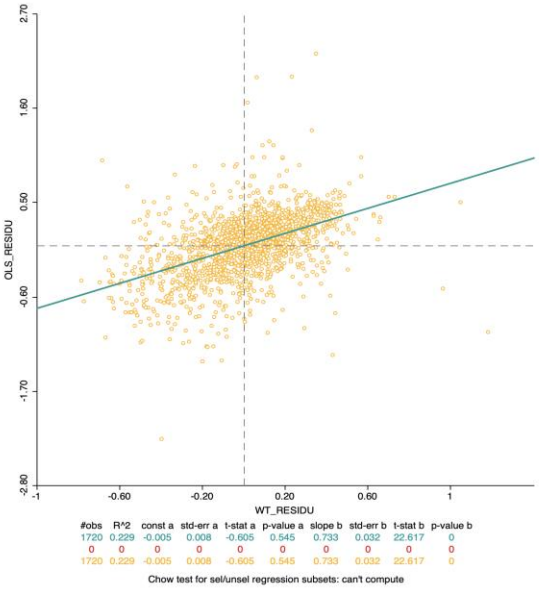

Figure 3: OLS Residuals and Weighted Residuals Scatterplot

The scatterplot has a slope-b of 0.733; we can interpret this as an increase of 1 unit in the lagged residual will change the residual by 0.733 units. However, this is problematic; if there is a relationship between OLS residuals and weighted residuals, spatial dependencies exist which violates the assumption of residual independency.

We also ran a Moran's I test and a significance test.

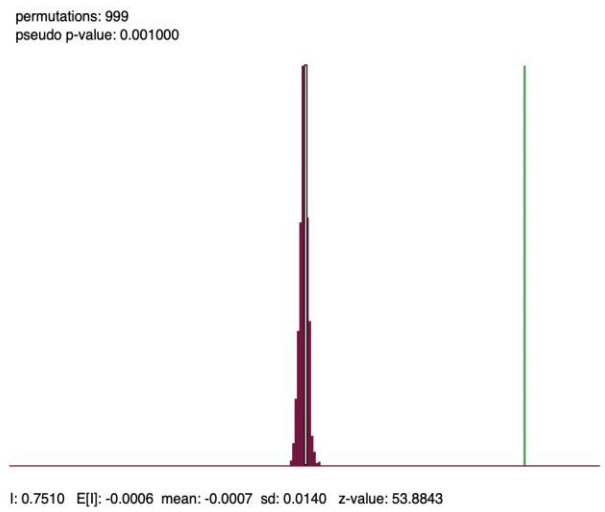

Figure 4: Significance Test (OLS Regression)

The Moran's I value is 0.791, thus we can reject H_0 in favor of H_{a1}, because there is a positive correlation with a value close to 1. When running the significance test, again, our pseudo p-value is equal to 0.001. Since 0.001 < 0.05, we reject H_0 that there is no spatial autocorrelation. This again exemplifies the problematic nature of our model, as our residuals should not be spatially autocorrelated.

The OLS model of Philadelphia's block groups indicates heteroscedasticity, non-normality of residuals, and spatial dependency in the residuals, which go against the several assumptions we have when running an OLS regression. Because of the spatial components of our data, OLS regression is likely not a good fit. Our next two sections will look at spatial error, spatial lag, and geographically weighted regression, all models that account for spatial components, to see if they have a better goodness of fit.

c) Spatial Lag and Spatial Error Regression: Results

First, we will look at the spatial lag model. As aforementioned, the spatial lag model uses the spatial lag of LNMEDHVAL as a predictor, where ρ is indicative of the lag of Wy. Presented below are the results of our spatial lag regression.

Table 2: Spatial Lag Regression

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : Regression Data
Spatial Weight      : Regression Data
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       3.89845       0.201114        19.3843     0.00000
        LNNBELPOV    -0.0340547     0.00629287       -5.41163     0.00000
       PCTBACHMOR    0.00851381    0.000521935         16.312     0.00000
       PCTSINGLES    0.00203342     0.00051577         3.9425     0.00008
        PCTVACANT    -0.0085294    0.000743667       -11.4694     0.00000
------------------------------------------------------------------------
```

All the variables from the OLS regression are presented, with the addition of "W_LNMEDHVAL" which is the spatial lag of LNMEDHVAL, also known as the coefficient of the parameter ρ. With a p-value < 0.001, it is a significant indicator in our model. Since W_LNMEDHVAL is significant, we can assume that logged median house value at one location is associated with logged median house values that are queen neighbors. Comparatively to OLS, all our other predictors remain with p-values < 0.05, indicating they are all still significant.

We completed the Breusch-Pagan test on the spatial lag regression and got a p-value < 0.001. Since the p-value < 0.05, we reject H_0 that there is no heteroscedasticity, indicating there is still a problem with heteroscedasticity even with our new model.

Next, we compared the spatial lag regression to the OLS regression based on the Akaike Information Criterion (AIC), the Schwarz criterion, the log likelihood, and the likelihood ratio test. These tools all help compare our models and indicate their goodness of fit. The results are presented in the table below.

Table 3: OLS v. Spatial Lag

Test	**OLS**	**Spatial Lag**
AIC	1432.99	523.48
Schwarz	1460.24	556.18
Log Likelihood	-711.49	-255.74
Likelihood Ratio		p-value = 0.00

Since the AIC of the spatial lag model is smaller by more than a difference of 3, we can assume the spatial lag has a better goodness of fit than OLS. This outcome is seen again the Schwarz criteria. Since the spatial lag's log likelihood is larger than the OLS' log likelihood, we can again assume the spatial lag has a better goodness of fit than OLS.

Finally, we look at Moran's I and significance testing to see if spatial dependencies exist.

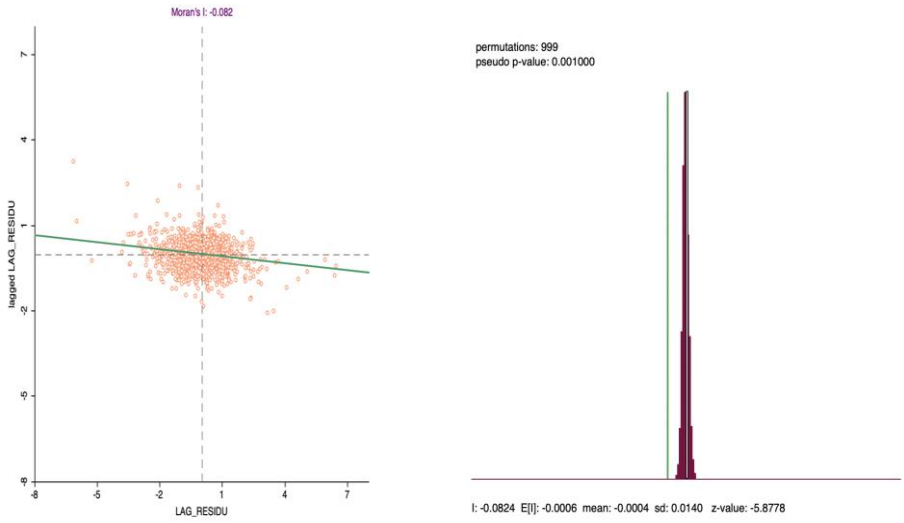

Figure 5: Moran's I and Significance Test (Spatial Lag)

Our Moran's I value is -0.082 and our pseudo p-value is 0.001. Because our Moran's I value is closer to 0, we see there is significantly lower spatial autocorrelation. However, our pseudo p-value < 0.05, so we reject H_0 that there is no spatial autocorrelation. While there is still a presence of heteroscedasticity and spatial dependencies, our previous tests indicate that the spatial lag is a better specification than the OLS model, and thus we can reject H_0 that OLS regression is doing a better job than spatial lag model.

We repeat these same steps but now using the spatial error model. The spatial error model assumes that residuals at one location are associated with residuals at nearby locations. Again, the spatial error model follows the same equation as OLS, but account for spatially lagged residuals, $\lambda W\varepsilon$, and random noise, u. Presented below is the output of our spatial error regression.

Table 4: Spatial Error Model

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : Regression Data
Spatial Weight      : Regression Data
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
      PCTSINGLES    0.00267792    0.000620832        4.31343     0.00002
       PCTVACANT   -0.00578308    0.000886701       -6.52201     0.00000
          LAMBDA      0.814918       0.016373        49.7719     0.00000
------------------------------------------------------------------------
```

The lag coefficient, λ, is equal to 0.81. With a p-value < 0.05, we can assume that it is a significant indicator in our model. Since λ falls on a -1 to 1 scale, we see there is a quite strong positive correlation. Comparatively to OLS, our previous four predictors all also maintain p-values < 0.05, so they are also significant indicators in our model.

We completed the Breusch-Pagan test on the spatial error regression and got a p-value < 0.001. Since the p-value < 0.05, we reject H_0 that there is no heteroscedasticity, indicating there is still a problem with heteroscedasticity.

Next, we use the same four tests as we did in spatial lag to see which model has a stronger goodness of fit. The model comparison interpretations remain the same for spatial error.

Table 5: OLS v. Spatial Error

Test	OLS	Spatial Error
AIC	1432.99	755.38
Schwarz	1460.24	782.631
Log Likelihood	-711.49	-372.69
Likelihood Ratio		p-value = 0.00

Since the AIC of the spatial error model is smaller by more than a difference of 3, we can assume the spatial lag has a better goodness of fit than OLS. This outcome is seen again the Schwarz criteria. Since the spatial error's log likelihood is larger than the OLS' log likelihood, we can again assume the spatial error has a better goodness of fit than OLS. With a likelihood ratio p-value < 0.001, we reject H_0 that OLS regression is doing a better job than spatial error.

The Moran's I and significant tests for spatial error regression are presented below.

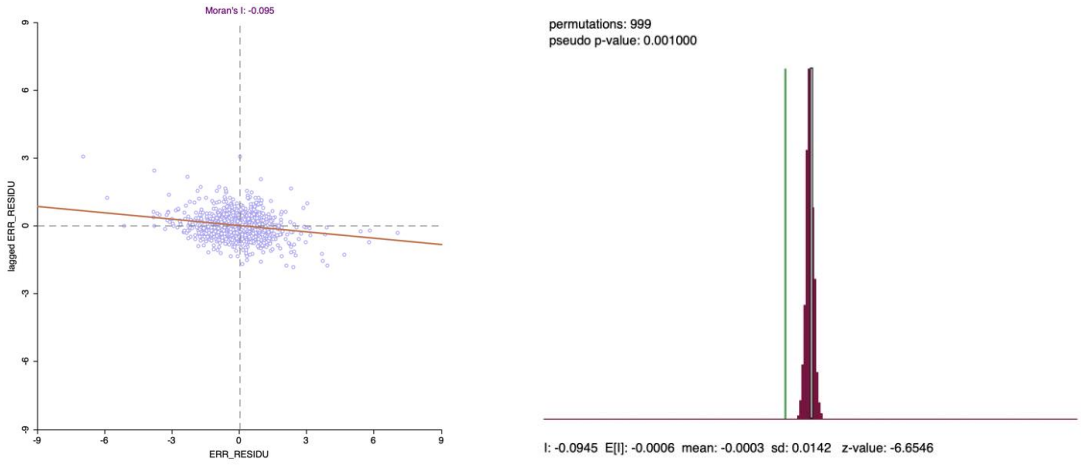

Figure 6: Moran's I and Significance Test (Spatial Error)

Our Moran's I value is -0.095 and our pseudo p-value is 0.001. Because our Moran's I value is closer to 0, we see there is significantly less correlation. However, our pseudo p-value < 0.05, so we reject H_0 that there is no spatial autocorrelation. While there is still a presence of heteroscedasticity and spatial dependencies, our previous tests indicate that the spatial error is a better specification than the OLS model, and thus we can reject H_0 that OLS regression is doing a better job than spatial error model.

Overall, in comparison to OLS, both spatial lag and spatial error have a better goodness of fit to our data. Despite continual problems with heteroscedasticity and spatial dependency, we find that through our different tests and Moran's I plot, they are still doing better than OLS. When analyzing which is better

between spatial lag and spatial error, we only use AIC and Schwarz. For both tests, we are looking at which one has a lower value. The AIC and Schwarz values for both models are in the table below.

Table 6: Spatial Lag v. Spatial Error

Test	Spatial Lag	Spatial Error
AIC	523.48	755.38
Schwarz	556.18	782.631

Spatial lag has lower AIC and Schwarz values in comparison to spatial error, thus we would assume that spatial lag model has a better goodness of fit than spatial error.

d) Geographically Weighted Regression: Results

Finally, we complete a geographic weighted regression as our final model method.

Table 7: GWR Diagnostics

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

We did not use R^2 for the spatial lag and spatial regression models because the interpretation of R^2 is different compared to OLS. However, we look at the $R^{2\ value}$ to determine which model does the best job at explaining the variance in LNMEDHVAL. In our OLS model, the R^2 is equal to 0.66 whereas in GWR, it is equal to 0.86. With a higher $R^{2\ closer}$ to 1, GWR does a better job at explaining the variance in the dependent variable.

The AIC of our GWR model is equal to 582.15, in comparison to 1432.99, 523.48, and 755.38 for OLS, spatial lag, and spatial error, respectively. We are looking for the smallest value for goodness of fit, thus when using AIC, spatial lag would be the best, followed by GWR, spatial error, and OLS, respectively.

Now, we present the Moran's I value and significance test for GWR.

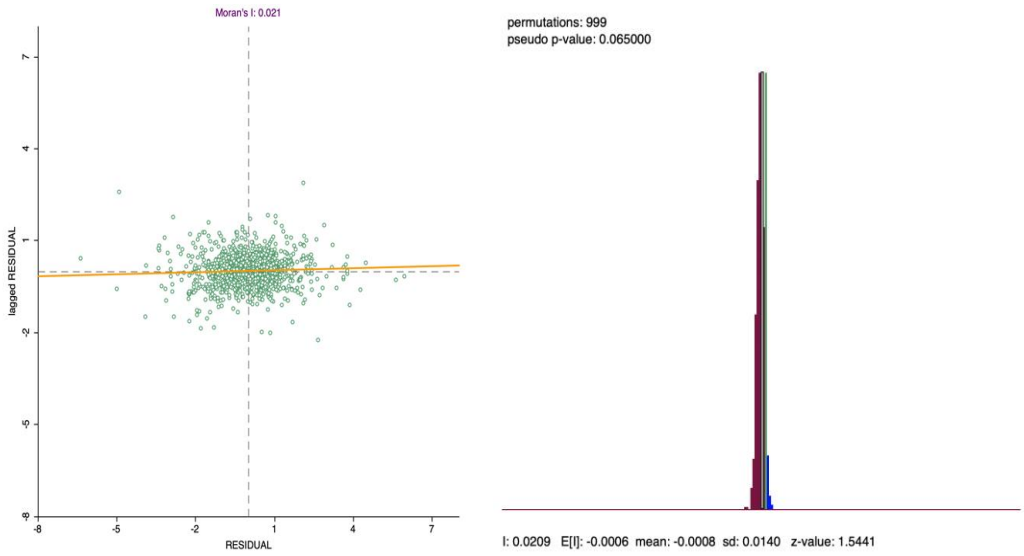

Figure 7: Moran's I and Significance Test (GWR)

Our Moran's I value is equal to 0.021. With a value close to 0, we can assume there are no spatial dependencies. Our pseudo p-value in the significance test is equal to 0.065. Since the value > 0.05, we fail to reject H_0 that there is no spatial autocorrelation. GWR is the only model to have no spatial dependencies and has the Moran's I that is closest to 0. Since we uphold a lot of the same assumptions in GWR as we do with OLS, our model seems to be doing a significantly better job at upholding those assumptions. Since GWR has the closest Moran's I to 0 and the only model where we fail to reject H_0, we believe it is a better choice than OLS, spatial lag, and spatial error.

Next, we want to understand the local dynamics that are captured within GWR modelling; specifically, we want to understand relationships between the dependent variable and predictors across space and how well our multiple regressions capture this. First, we look at local R^2 for all census blocks.

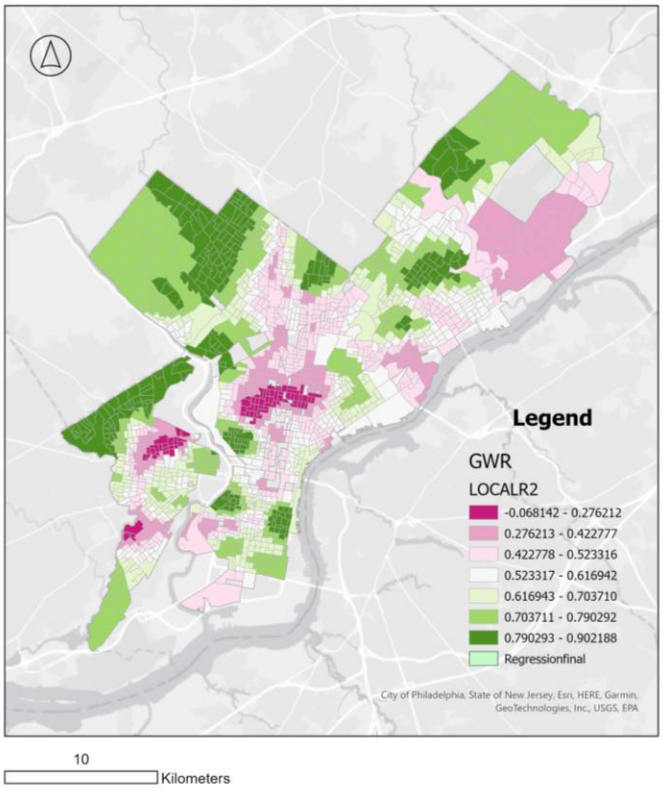

Figure 8: Global R^2 Choropleth Map

An R^2 closer to 1 indicates a strong fit whereas an R^2 closer to 0 indicates a weak fit. Something to note is that ArcGIS Pro has an implementation error that causes some of our outputs to be negative values. Regardless, we see throughout several parts of the city, we have strong fits, but there is great variation across R^2. There are several census blocks where the model has a weaker fit, but the R^2 for the entirety of our study area is 0.86, which means that our model is doing a great job at explaining the variance in LNMEDHVAL.

Now, we look at coefficient estimates across our four indicators. Here is what are scale indicates.

$SE \leq -2$: negative relationship with DV, possibly significant (dark blue)
$-2 < SE \leq 0$: negative relationship with DV, possibly insignificant (light blue)
$0 < SE < 2$: positive relationship with DV, possibly insignificant (light red)
$SE \geq 2$: positive relationship with LNMEDHVAL, possibly significant (red)

Our four indicator maps are attached, where ratio1 is PCTBACHMOR, ratio2 is PCTVACANT, ratio3 is PCTSINGLES, and ratio4 is LNNBELOWPOV.

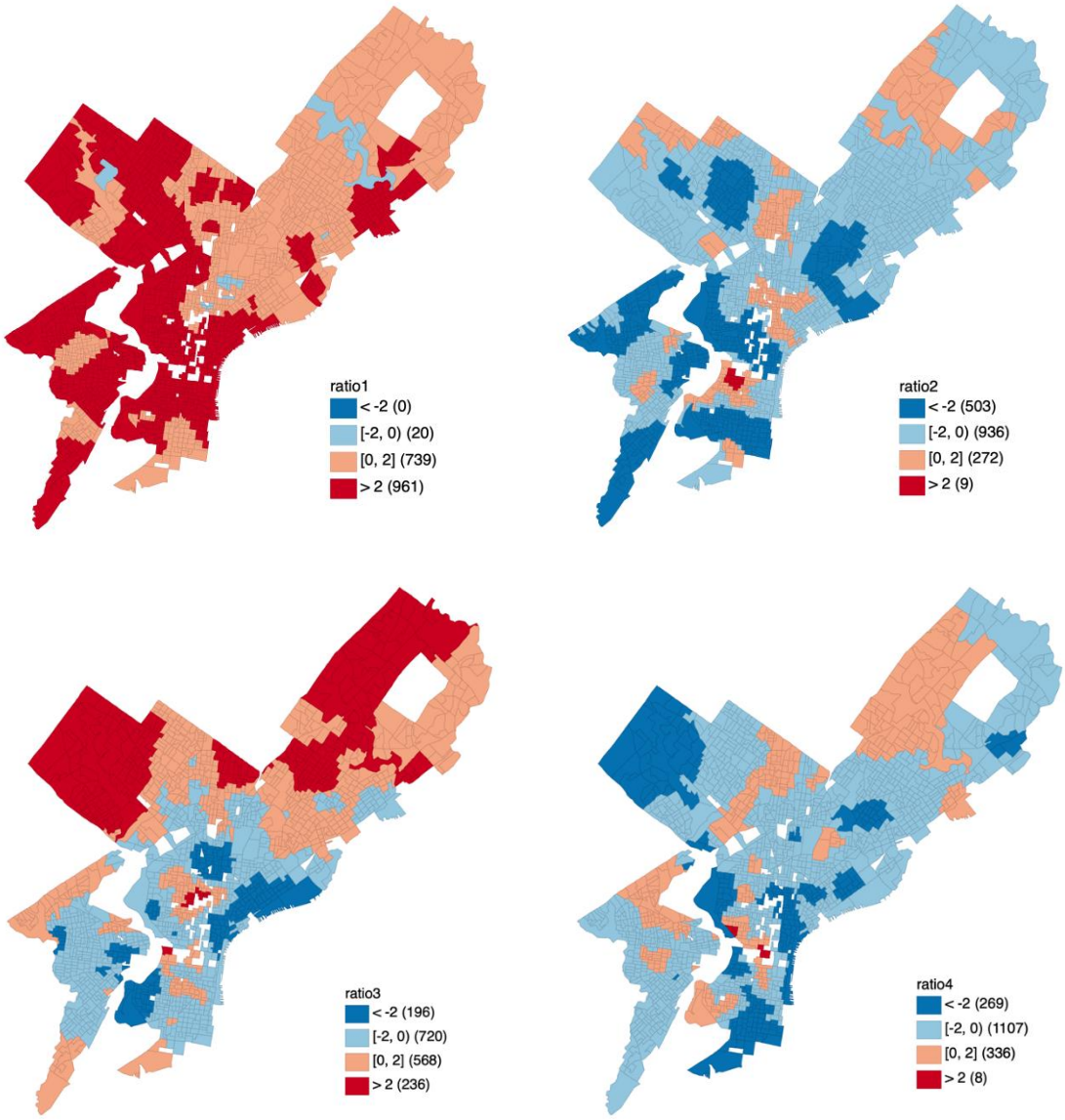

Figure 9: Spatial variation maps of indicators

These maps examine spatial stationary relationships across LNMEDHVAL and the respective predictors. We see quite strong regional variation across all four predictors. For PCTBACHMOR, we see minimal clustering and positive correlation across all census blocks, with possible significance especially in parts in South, West, and Northwest Philadelphia. For PCTVACANT, we see significant negative relationships in South Philadelphia, University City, and around Fairmount Park, whereas only a few blocks have a significant positive correlation. For PCTSINGLES, it appears the further outside downtown Philadelphia the block is, the more likely there is to be a positive and significant relationship. LNNBELOWPOV seems to have the most global variation, with significant negative relationships in parts of Northwest, South and downtown Philadelphia, and very minimal significant positive relationships. If we look at the counts of each metric on our scale, we can assume that for PCTBACHMOR, most of our census blocks have significant positive correlations; for PCTVACANT, we mostly have negative correlations with some being likely significant and most being unlikely significant; for PCTSINGLES, we see significant variation in

correlation and significance; and for LNNBELOWPOV, we see mostly negative correlations, with the majority being likely insignificant.

Results

In this paper, we revisited our previous study on the relationship between median household values and several neighborhood predictors in Philadelphia at the census block group level. We preserved our same four indicators as before. Previously, we used an Ordinary Least Squares regression, but we found several limitations to the method, specifically with the several violations of the model which included heteroscedasticity, multicollinearity, non-normality of residuals, and spatial dependencies. This is because OLS is not generally a go-to model choice when dealing with spatial data. Thus, in this analysis, we focused on other models that accounted for different spatial dynamics to our data to reduce heteroscedasticity and normalize our residuals.

We used Akaike Information Criterion, Schwarz Information Criterion, Log Likelihood, and Likelihood Ratio tests to determine goodness of fit. We then used Breusch-Pagan test to determine heteroscedasticity, and finally Moran's I and significance testing to determine spatial autocorrelation. Overall, we found that spatial lag, spatial error and GWR all had lower AIC and Schwarz than OLS, indicating better goodness of fit. Spatial lag overall had the best goodness of fit based on AIC and Schwarz. However, when we looked at Moran's I and significance tests, we found that spatial lag and spatial error still violated errors associated with heteroscedasticity and spatial autocorrelation residuals. GWR produced the Moran's I closest to 0, and we could also fail to reject that there is no spatial autocorrelation. Thus, we would state that GWR is the best model based on our results.

Despite spatial lag, spatial error, and GWR all being better fits than OLS, there were still limitations to this study. As aforementioned, spatial lag and spatial error produced p-values of 0.00 after the Breusch-Pagan test, indicating heteroscedasticity. Both models also produced pseudo p-values of 0.001 after completing significance tests, meaning we still saw spatial autocorrelation. While GWR did not have a stronger AIC score than spatial lag, it did not violate any of our assumptions.