ASSIGNMENT 6

Study: Median Housing Values in Philadelphia

1 INTRODUCTION

Recall that in last report [A5], the primary differences of areal data from continuous data has been given in the "Blood Group A" study [S1, A5], including ***centroid*** distance and *spatial weight matrix*. We have also laid the groundwork of ***spatial autoregressive model*** that is more appropriate to capture the spatial autocorrelation among variables of interest in "Median Housing Values in Philadelphia" case [S2, A5]. Besides, we have used three alternative ***statistics (Moran, rho, and corr)*** to test the magnitude of such spatial relations, associated with ***sac-perm test*** for p-values. The test results yielded significant spatial correlation among residuals in OLS model remained to be solved in this report.

The dataset and hence associated variables used here is the same to [S2, A5], where ***logged median housing values*** would be again regressed on ***percent of nonwhite*** and ***percent of vacant house units*** at each census tracts, by two most fundamental ***spatial regression models*** (14 tracts with meaningless demographic information has been removed already). Here, the dependent variable of interest and the two independent variables are respectively denoted as lnMV, %NW, and %Vac.

2 METHODS & RESULTS

2.1 Boundary-Share Weight Matrix

Before introducing the two new spatial models to replace OLS model, we will construct a new spatial weight matrix for Philadelphia tracts that relies on ***boundary shares***, designated as ***boundary-share weight matrix***. Unlike exponential matrix used in [S2, A5] that are easily computed based on centroid distance, this new matrix models the spatial relations of areal units according to pairwise shared boundaries that performs better in identifying ***contiguities***. Recall that the census tract in North Philadelphia we discussed is surrounded by 5 neighboring tracts but only 4 neighbors are included in the ***first four nearest-neighbor matrix*** due to smaller centroid distances than the excluded one. This neighbor definitely has some degree of "spatial influence" on the target tract and should have been included, since they are indeed contiguous.

Hence, if we define the set of boundary points of unit R_i as $bnd(i)$, l_i as the ***total boundary length***

of bnd(i) that is shared with other spatial units, and the shared boundary length of i with any particular unit j as l_{ij}, then the fraction of this length shared yields a potentially relevant shared boundary weights, given by:

$$w_{ij} = \frac{l_{ij}}{l_i} = \frac{l_{ij}}{\sum_{k \neq i} l_{ik}}$$

We then use the Intersect tool in ArcMap to calculate the shared lengths of each pairwise contiguous tracts and then import the results to MATLAB to build such boundary-shared weight matrix, each row of which would hence be normalized to yield a unit sum of 1. For example, if we denote each of these 353 tracts with a unique ID, it can be observed in Figure 1 below that the tract with ID = 284 (denoted in light red) is contiguous with three neighboring tracts (denoted in light blue), with respective IDs as 269, 273 and 282, whose normalized spatial influences on this target tract are also given respectively.

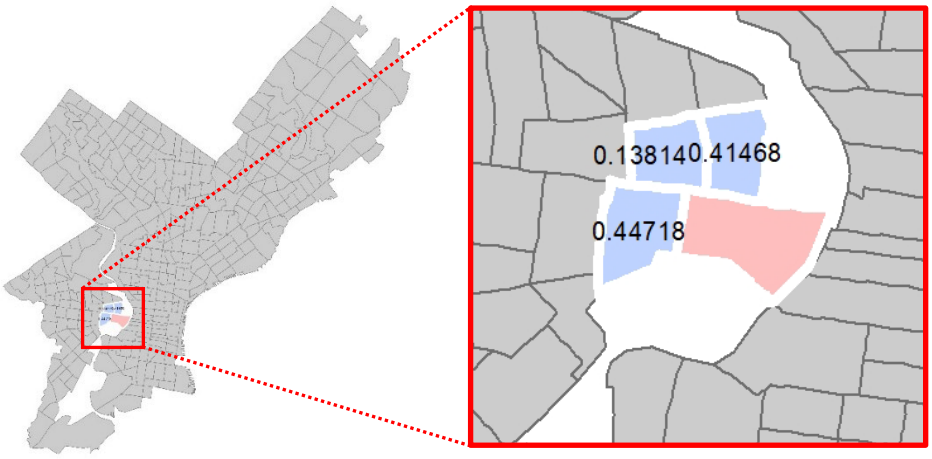


Figure 1. Target Tract

It's noticed that such weights need not always be the most "reasonable" ones. In Figure 2 below, by observing the tract locating at the southern end of Philadelphia (with ID = 353 and denoted in light red), we can find out that it's considered spatially influenced by ***only one*** neighboring tract (with ID = 352 and denoted in light blue) and the respective weight is designated as 1, because the connections to its northeast neighbors (denoted in light green) at the east side of Schuylkill River are excluded due to the removed tracts. However, though this target tract is ***geographically separated*** by the river, it could be somewhat considered to adjoin to tracts on the east side of the river in some cases that such geographic borders are very narrow and throw merely little influence of separation.

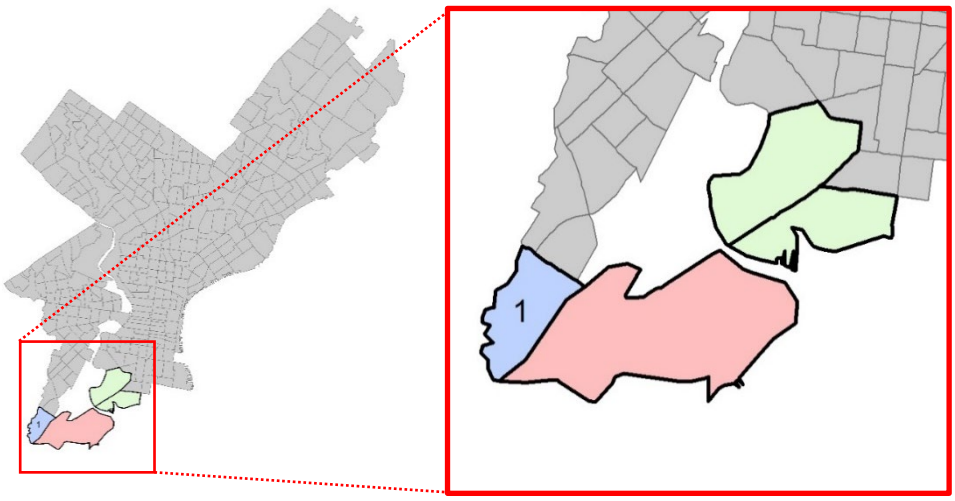


Figure 2. Tracts Separate by the River

Based on this boundary-share weight matrix, we then recheck the presence of spatial autocorrelation in the OLS residuals using the same ***sac-perm test*** in MATLAB as we have done in [S2, A5]. The results are given in Figure 3 below:

RANGE OF RANDOM-PERMUTATION INDEX VALUES:

INDEX	Moran	corr	rho
MAX	0.1289	0.2429	0.4575
MIN	-0.1060	-0.2028	-0.4143

TABLE OF SIGNIFICANCE LEVELS:

INDEX	VALUE	SIGNIF
Moran	0.7584	0.0010
corr	0.8683	0.0010
rho	0.9941	0.0010

Figure 3. SAC-Perm Test Results for OLS Residuals

Similar to the test using exponential weight matrix, notice that the ***significance levels*** for the three test statistics (Moran, corr, rho) are also ***maximally significant***, indicated by p-values (denoted as SIGNIF) all designated as 0.001. Furthermore, they are still ***higher*** and even ***much higher*** than the range of values represented by the MAX and MIN obtained in the 999 random permutations. For example, the most widely used and powerful statistic to test spatial autocorrelation, Moran's I (denoted as Moran), yields a value of 0.7584 that is well above the range of values, -0.1060 to 0.1289. Same conclusion can be drawn from the other two statistics. Hence, in a conclusion, this test agrees with the finding of [S2, A5] that the OLS model yields significant spatial autocorrelation

in residuals.

2.2 Spatial Error Model (SEM)

The first model we will perform is the ***spatial error model (SEM),*** which postulates that dependencies among the regression residuals (errors) u_i at each areal unit (i.e., census tract) *i* are captured by the spatial autoregressive model by:

$$u_i = \rho \sum_{j \neq i} w_{ij} u_j + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, ...n$$

For some specific spatial weight matrix, $W = (w_{ij} : i, j = 1, ..., n)$, i.e., boundary-share weight matrix, the spatial error model in the form of matrix can be written as:

$$Y = X\beta + u \quad , \quad u = \rho W u + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \quad ,$$

where it's hypothesized that all spatial dependences are modeled among the unobserved errors in the term $\rho W u$.

In MATLAB, we regress lnMV on %Vac and %NW using the SEM based on boundary-share matrix constructed in section 2.1. The regression results will be presented later for comparison with the other spatial regression model.

Recall we have rechecked and confirmed the presence of spatial correlations among OLS residuals based on the p-values and simulated value intervals of each statistic in section 2.1 above. Here, the SEM residuals will be examined using the same sac-perm test for 999 random permutations.

RANGE OF RANDOM-PERMUTATION INDEX VALUES:

INDEX	Moran	corr	rho
MAX	0.1563	0.2723	0.4743
MIN	-0.1146	-0.2214	-0.4275

TABLE OF SIGNIFICANCE LEVELS:

INDEX	VALUE	SIGNIF
Moran	-0.0804	0.9800
corr	-0.1416	0.9680
rho	-0.2494	0.9540

Figure 4. SAC-Perm Test Results for SEM Residuals

The results of sac-perm test are given in Figure 4 above. Compared to test results of simple OLS residuals, it turn out that ***positive*** spatial autocorrelation is ***not present*** but ***negative*** correlations are

detected, since all three statistics' p-values (denoted as SIGNIF) turn out to be above 0.95, which indicates that the null hypothesis of spatial independence is rejected for an alternative hypothesis that there is significantly ***negative*** spatial autocorrelation among residuals of predicted housing prices. Besides, the evaluated ***common scale factor*** $\hat{\rho}$ of ρ in ρWu is designated as 0.8064, which refers to the magnitude of ***positive*** spatial dependences of OLS residuals on their respective neighboring residuals (defined by the boundary-share matrix). To sum up, the SEM indeed effectively removes the positive spatial autocorrelation in OLS residuals but yields unexpected negative spatial autocorrelation in SEM residuals. Hence, we will try another spatial regression model to examine its performance.

2.3 The Spatial Lag Model (SLM)

The alternative linear spatial autoregressive model is obtained based on assumption that these autoregressive relations are among the ***dependent variables themselves***:

$$Y_i = \beta_0 + \rho \sum_h W_{ih} Y_h + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \quad , \quad i = 1, ..., n$$

In the case of median housing values in Philadelphia, it's assumed that median housing prices at each census tract Y_i is influenced not only by other housing attributes (i.e., in this context, %Vac and %NW), but prices in surrounding tracts Y_h. Such relations are typically called ***spatial lag*** relations, which can be governed by the other model named ***spatial lag model (SLM)***. Next, we run the SLM using same boundary-share matrix and apply the sac-perm test for residuals in SLM. For the purpose of comparison, testing and regression results of SLM and SEM are given side by side in Figure 5 & 6 in the next section below.

2.4 Comparison of SEM and SLM

RANGE OF RANDOM-PERMUTATION INDEX VALUES:

INDEX	Moran	corr	rho
MAX	0.1563	0.2723	0.4743
MIN	-0.1146	-0.2214	-0.4275

TABLE OF SIGNIFICANCE LEVELS:

INDEX	VALUE	SIGNIF
Moran	-0.0804	0.9800
corr	-0.1416	0.9680
rho	-0.2494	0.9540

RANGE OF RANDOM-PERMUTATION INDEX VALUES:

INDEX	Moran	corr	rho
MAX	0.1396	0.2382	0.4261
MIN	-0.1306	-0.2445	-0.4868

TABLE OF SIGNIFICANCE LEVELS:

INDEX	VALUE	SIGNIF
Moran	0.0219	0.2620
corr	0.0382	0.2730
rho	0.0665	0.2830

Figure 5. Comparison of SAC-Perm Test Results for SEM & SLM Residuals

It's noticed from Figure 5 above that with resepct to SEM, all p-values (denoted as SIGNIF) of each testing statistic for SLM exhibit statitstical insignificance, since none of they fall in either the low (say $\alpha \le 0.05$) or high ($\alpha \ge 0.95$) levels, which respectively refer to positive and negative spatial autocorrelation. Taking Moran as an example since it's the most useful one, the observed value is designated as 0.0219, which lies in the interval of simulated values indicated by MAX = 0.1396 and Min = -0.1306. In addition, the p-value of Moran = 0.262 indicates that under the results of 999 permutations, the chance of observing this large a Moran's I is 262/1000 if in fact there is indeed no spatial autocorrelation. The chance is to some extent robust enough, so that it's concluded the null hypothesis of spatial independence ***cannot be rejected***. It turns out that SLM does effectively remove the spatial autocorrelation in OLS residuals and yield non-spatially-related residuals.

Furthermore, the autoregressive results of both SEM and SLM are also presented for comparison below:

```
FINAL REGRESSION RESULTS:

VAR          COEFF        Z-VAL          PROB
const     11.202002    93.134155      0.000000
%Vac      -0.005577    -2.065597      0.038867
%NW       -0.008143    -8.651840      0.000000

Variance = 0.084511

AUTOCORRELATION RESULTS:

             VAL         Z-VAL          PROB
rho      0.860369     33.781485      0.000000


GOODNESS-OF-FIT RESULTS:

Extended R-Square      = 0.35737
Extended R-Square Adj  = 0.35369
Squared_Correlation  = 0.39166
Log Likelihood Value = -113.0032
AIC = 236.0065
AIC_corrected = 236.1794
BIC = 255.3388

TESTS OF SEM MODEL:

TEST            VAL          PROB
LR       392.992913      0.000000
Com-LR     6.575770      0.037333

MORAN z-score and p-value = (-1.9366, 0.9736)
```

```
FINAL REGRESSION RESULTS:

VAR          COEFF        Z-VAL          PROB
const      2.023761     6.722147      0.000000
%Vac      -0.004215    -1.779236      0.075201
%NW       -0.003170    -5.599376      0.000000

Variance = 0.090844

AUTOCORRELATION RESULTS:

             VAL         Z-VAL          PROB
rho      0.829982     31.889387      0.000000


GOODNESS-OF-FIT RESULTS:

Extended R-Square      = 0.22428
Extended R-Square Adj  = 0.21985
Squared_Correlation  = 0.3191
Log Likelihood Value = -120.7901
AIC = 251.5803
AIC_corrected = 251.7532
BIC = 270.9126

TEST OF SLM MODEL:

TEST           VAL          PROB
LR      377.419117      0.000000

MORAN z-score and p-value = (0.73811, 0.23022)
```

Figure 5. Comparison of Regression Results for SEM & SLM

We can find out that both regression models yield similar formats to OLS regression in the sense that significance levels (p-values) for each independent variable and the constant intercept are reported, together with the evaluated common scale factor $\hat{\rho}$ of ρ in ρWu and various measures of "goodness of fit". It's noticable that estimation of parameters and associated p-values in SEM and SLM are obatined based on ***maximum-likelihood estimation***, designated as a primary departure from OLS due to the special autoregressive structure (more detailed explanations are given in [Notebook, III.7 & 8]).

The effects (i.e., coefficicents denoted as COEFF) of %Vac and %NW are ***consistently negative*** in SEM and SLM, that is to say, ***one-unit increase*** in percent of nonwhite or percent of vacant house units at each census tract would result in the ***decrese*** of median house values with units indicated by ***resepective coefficients***. In addition, %NW are extremely significant in both models,

while %Vac in SLM is a little bit less significant with respect to SEM (with a p-value of 0.0752 vs. 0.0387).

Notice that the levels of spatial autocorrelation, $\hat{\rho}$, are respectively designated as 0.8604 and 0.8300 in SEM and SLM. Such similar magnitude of postive spatial autocorrelation again supports the findings of sac-perm test for OLS residuals above.

2.5 Goodness of Fit

Among the measures of "goodness of fit", we in particular focus on the ***Extended R-Squared*** value, which is in many ways the most meaningful. Next, the interpretation of Extended R-Squared starts with a brief discussion of R-Squared and Adjusted R-Squared in OLS models.

2.5.1 The R-Squared Measure for OLS

Recall that we have used them to describe the percent of variance explained by the model. To be more specific, the solution to β parameters vector (coefficients) for each independent variable serves to ***minimize*** the ***sum of squared deviations*** of predicted values from actual values. For each observation of a specific dependent variable *y*, if we denote each of the actual value and the average as y_i and $\overline{y}$ respectively, the ***total variance*** of *y* can be defined as $\sum_i (y_i - \overline{y})^2$. Then if we again denote the predicted value by OLS as $\hat{y}_i$, it turns out that the total variance can be decomposed into two parts, designated as sum of squared deviations of both $y_i - \hat{y}_i$ and $\hat{y}_i - \overline{y}$, i.e., that

$$\sum_i (y_i - \overline{y})^2 = \sum_i (\hat{y}_i - \overline{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \quad ,$$

where on the right side of the formula, the form term refers to the ***model variance***, and the latter form is designated as ***residual variance***. Usually, the residual variance is also denoted as $\sum_i \hat{e}_i^{\,2}$.

In this setting, the desired R-Squared is taken to be the fraction of total variation accounted for by model variation, i.e.,

$$R^2_{OLS} = \frac{model\ variation}{total\ variance} = \frac{\sum_i (\hat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} \quad \text{or} \quad \overline{R}^2_{OLS} = \frac{residual\ variance}{total\ variance} = 1 - \frac{\sum_i \hat{e}_i^{\,2}}{\sum_i (y_i - \overline{y})^2}$$

2.5.2 Adjusted R-Squared

While the R^2_{OLS} is intuitively very appealing as a measure of goodness of fit, it suffers from some limitations. For example, the most important one is that it in fact ***keeps increasing*** as ***more***

explanatory variable are ***added*** to the model, which would inevitably lead to the problem of "overfitting". Indeed, for an OLS model with *k* predictors, the R^2_{OLS} can be perfectly increased to 1 if we include $n-1$ explanatory variables in the model. The solution to this problem is to use another measure that could reflect the number of explanatory variables, that is, the R^2_{OLS} will be accordingly turned down in value, designated as adjusted R-Squared, defined by:

$$\bar{R}^2_{OLS} = 1\text{-}(\frac{n\text{-}1}{n-k})(1-R^2_{OLS})$$

2.5.3 Extended R-Squared Measures for GLS

When extending R^2_{OLS} and $\bar{R}^2_{OLS}$ to GLS, there is a problem that there is no unambiguous way to define the "fraction of variation explained" by the given GLS model, because the fundamental decomposition of total variation no longer holds (as more specifically explained in [Notebook, III.9-4-9, 11-12]). However, the residual variance $\sum_i \hat{e}_i^{\,2}$ still captures the sum of squared deviations of each y_i from their predicted values $\hat{y}_i$. Moreover, since total variance $\sum_i (y_i-\bar{y})^2$ still represents the sum of *y* deviations from their least-squares prediction $\bar{y}$, under the null model. So that it's reasonable to modify the R^2_{OLS} to its extended form for GLS model by comparing its ***mean square error*** and that under the null model:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \quad \text{and} \quad MSE_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

The schematic graphic of this comparison is given by the Figure 6 and 7 below:

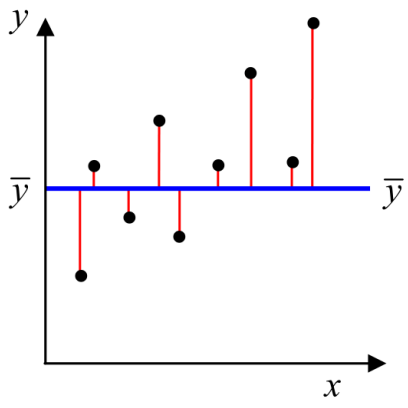


Figure 6. Null Deviations

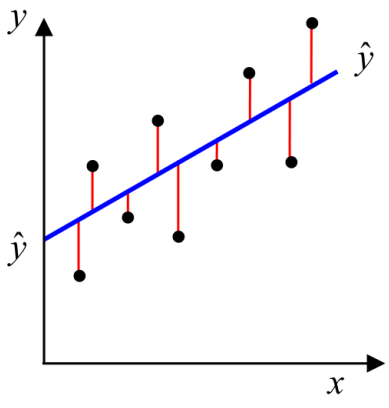


Figure 7. Model Deviations

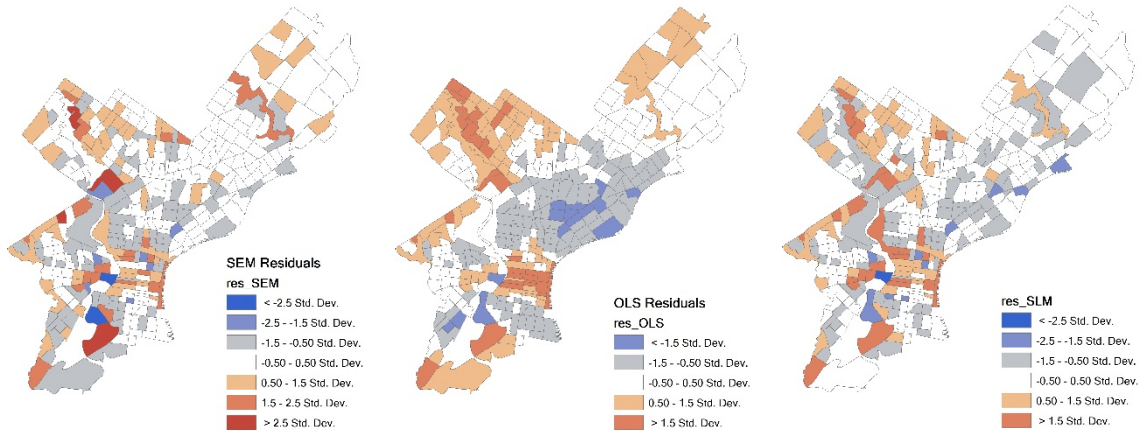
Hence, the extended R-Squared and adjusted extended R-Squared for GLS is given by:

$$R^2_{GLS} = 1 - \frac{MSE}{MSE_0} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{and} \quad \bar{R}^2_{GLS} = 1\text{-}(\frac{n\text{-}1}{n-k})(1 - R^2_{GLS})$$

Given the overview of extended R-Squared for GLS models, we can then compare these values obtained in SEM and SLM. It's observed that though SLM performs better in removing spatial autocorrelation in residuals than SEM, its extended R-Squared designated as 0.2243 is well smaller than that in SEM as 0.3574. In other words, SEM accounts for about 13 more percent of variance in median housing values with respect to SLM.

2.6 Visual Comparison of Residuals in Different Models

Finally, it's of interest to compare the residuals for these models we have used (OLS, SEM and SLM) spatially to see whether the relative patterns of spatial residuals agree with our statistical findings above. Residuals of different models are side by side aligned in Figure 8 below in the form of ***standard deviation***, so that they could be unified in comparative scale with corresponding colors.



SEM Residuals **OLS Residuals** **SLM Residuals**

Figure 8. Comparison of Residuals of SEM, OLS & SLM

It's observed that OLS residuals exhibit the most obvious pattern of ***clustering*** compared to the other two models, that is, residuals in similar magnitude (denoted by color ramp) are closer to each other in space as expected. Such sign of positive spatial autocorrelation is ***not present*** in maps of residuals for SEM and SLM, where residuals are more ***randomly*** distributed spatially. However, recall that based on the statistical findings above, SLM has indeed removed spatial correlation successfully, while SEM yield unexpected significantly ***negative*** spatial relations among residuals, which is somewhat hard to determine through visual observation on these maps. Actually, spatial residuals for SEM and SLM looks somewhat similar to each other visually.

3 DISCUSSION

In this report, we have extended the application of spatial autoregressive models in replacement of simple OLS models. In the case of median housing values in Philadelphia, it turns out that these two models (SEM and SLM) perform much better in removing positive spatial autocorrelation among residuals than OLS. Though SEM yields better performance than SLM from the perspective of goodness of fit (e.g., obviously larger R-Squared), we still consider SLM as the relatively ***best*** model in this case since the primary objective is to remove spatial autocorrelation, where SLM works well while residuals in SEM yield some degree of ***negative spatial relations***.

It should be addressed that any kind of models are just mathematically ***simplified simulations*** of the reality. Thus, it's inappropriate to expect a ***totally perfect fitting*** of median housing values in Philadelphia in this context and it's held even in any other statistical cases. But we can still ***approach*** the "fact" (i.e., fitting of perfection) by ***modifying*** our statistical models, which is designated as a process of ***trial-and-error***. Here, for the purpose of better fitting of housing prices, an alternative solution is to include several more reasonable explanatory variables, e.g., median household income, education levels, and other environmental factors like the number of service facilities and infrastructures at each census tract.

4 REFERENCE

[Notebook] T.E.Smith, "Notebook for ESE 502"

[A5] J.Wu (2019), "Assignment 5: Studies for ESE 502"