Study 2: Median Housing Values in Philadelphia

1 INTRODUCTION

This study is another analysis using aggregated areal data and serves to lay the groundwork for a ***spatial regression analysis*** of median housing values in Philadelphia. The data set used in this study contains some demographic information based on census surveys for 367 census tracts of Philadelphia in 1990, including median housing values, number of vacant house units, total population and population of white residents. Notice that there are 14 tracts with median housing values = 0, together with other missing values, which will be eliminated from this analysis, so that the number of tracts after data cleaning remains 353.

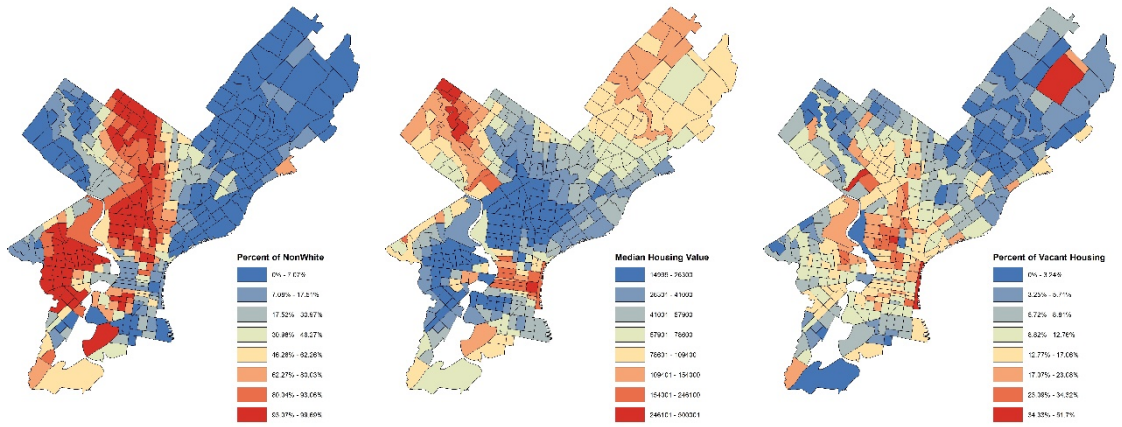
It's natural to assume that ***physical and social environment*** of a particular neighborhood may affect the housing prices there, especially for Philadelphia that is always regarded as a "city of neighborhoods". In the article "Revitalizing Neighborhoods" that briefly discusses about housing market and demographic trends in Philadelphia, it's mentioned that many neighborhoods are suffering from long-term social and economic decline, associated with half-century lasting population loss. As a result, an unprecedented level of abandoned and vacant housing has emerged and prevailed as an important indicator of physical blight, which substantially influences the housing prices. Also, in practice, the racial composition of neighborhood is somewhat related to education levels, median household income, and etc., which can be reflected in housing prices (i.e., neighborhoods with higher education levels and median household income are expected to exhibit higher housing prices. Given such observations and assumptions, we attempt to regress ***median housing values*** on ***percent of vacant housing*** and the ***percent of nonwhite population*** in each census tract.

2 METHODS & RESULTS

2.1 Exploratory Analysis

This analysis starts with construction of two possible explanatory variables (predictors), percent of vacant housing and percent of nonwhite population mentioned above. For comparison, we then create choropleth maps of these two explanatory variables and the dependent variable, median housing values, to observe if there exist similar spatial distribution patterns (correlations), as shown in Figure 1 below. It can be observed that median housing values at central Philadelphia is the lowest, and gradually increase as going outward to suburban areas. In contrast, the other two

explanatory variables exhibit relatively opposite distribution patterns, that is to say, higher percentages of vacant house units and nonwhite residents are more concentrated in the center city than outer zones. Though there also exists some degree of differences between the explanatory variables that percent of nonwhite is more clustered in central, north and west Philadelphia, while percent of vacant housing is relatively dispersed, the general similarity indicates ***negative correlations*** of each explanatory variable with the dependent variable. To sum up, the values of each variable for census tracts are obviously related to their close neighbors in space, which is exact the sign of ***spatial autocorrelation***.



Percent of NonWhite **Median Housing Values** **Percent of Vacant Housing**

Figure 1. Comparison of Variables

Then we will examine the distributions of each variable in JMP. The distribution of median housing value (denoted as MV) are highly skewed, which can be removed by transforming these values to logs (denoted as lnMV), as shown in Figure 2 & 3 below. The logged values are more normally distributed than original data since the respective normal quantile plot approximates a more diagonal (i.e., 45 degree) straight line.

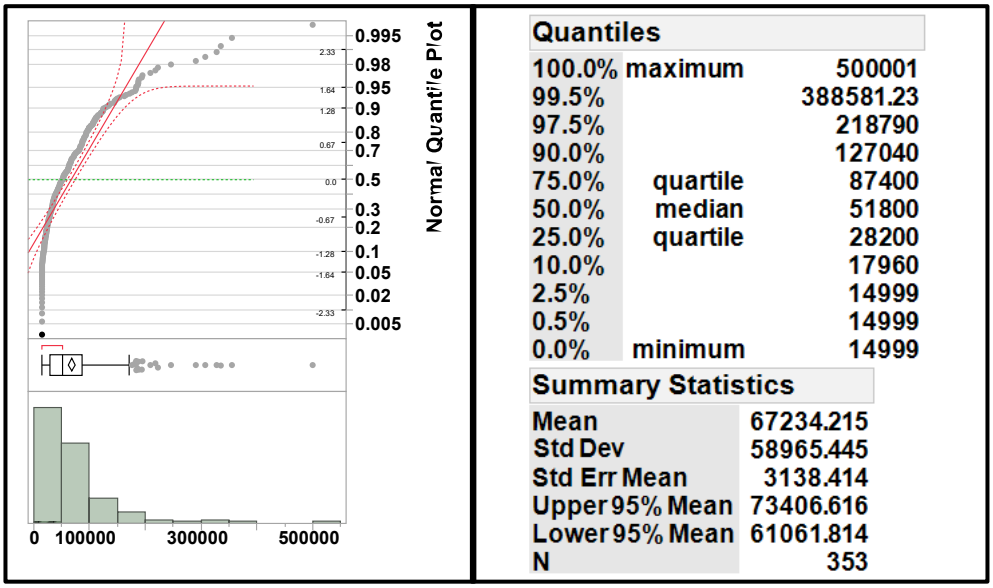


Figure 2. Distribution of Median Housing Values

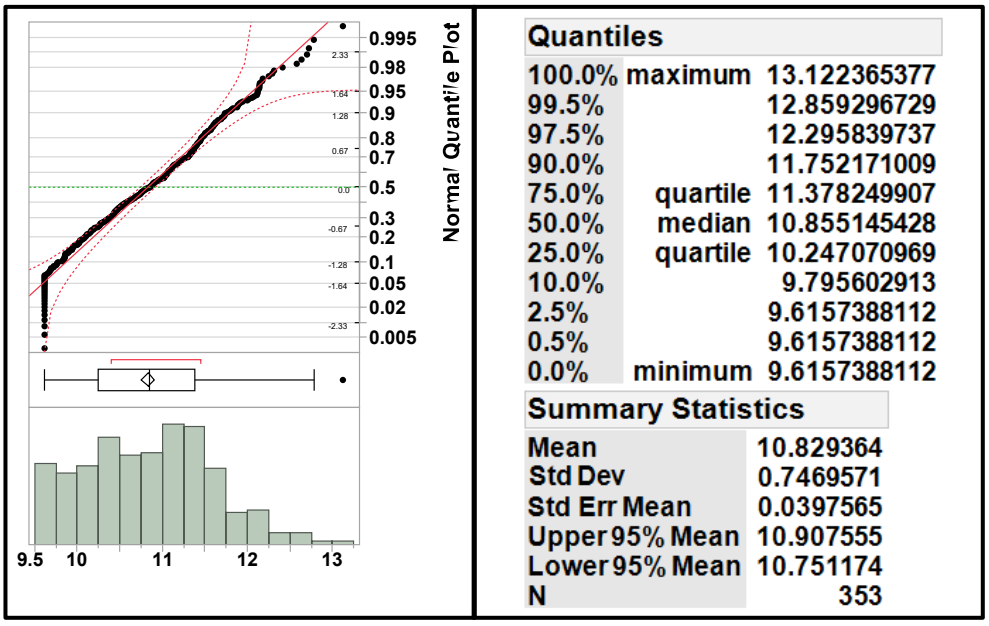
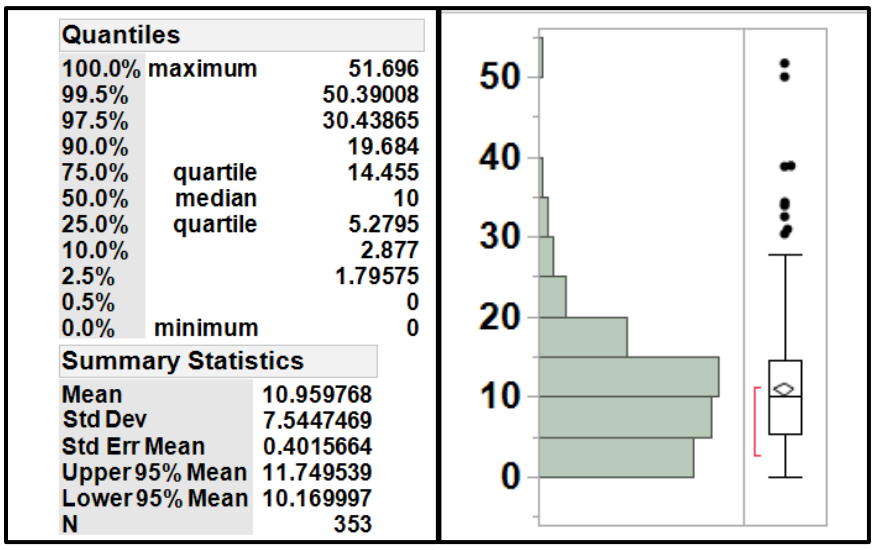


Figure 3. Distribution of Logged Median Housing Values

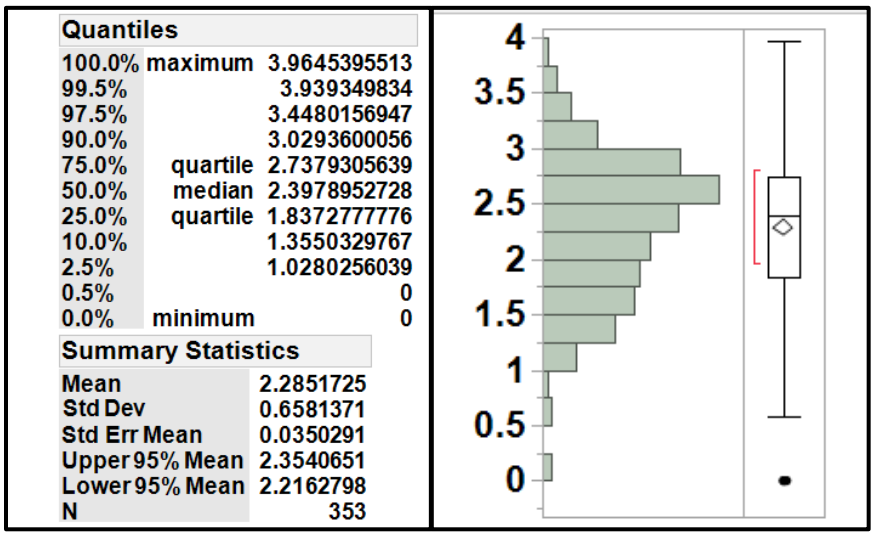
Likewise, we then examine the frequency distributions of the two explanatory variables. For convenience, percent of vacant house units and percent of nonwhite are respectively denoted as %Vac and %NW here. Notice that %Vac also exhibits some degree of skewness but could be solved using the log-transformation [log(1+%Vac), denoted as logit%Vac], as shown in Figure 4 and 5 below.



Quantiles		
100.0%	maximum	51.696
99.5%		50.39008
97.5%		30.43865
90.0%		19.684
75.0%	quartile	14.455
50.0%	median	10
25.0%	quartile	5.2795
10.0%		2.877
2.5%		1.79575
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	10.959768
Std Dev	7.5447469
Std Err Mean	0.4015664
Upper 95% Mean	11.749539
Lower 95% Mean	10.169997
N	353

Figure 4. Distribution of Percent of Vacant Housing



Quantiles		
100.0%	maximum	3.9645395513
99.5%		3.939349834
97.5%		3.4480156947
90.0%		3.0293600056
75.0%	quartile	2.7379305639
50.0%	median	2.3978952728
25.0%	quartile	1.8372777776
10.0%		1.3550329767
2.5%		1.0280256039
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	2.2851725
Std Dev	0.6581371
Std Err Mean	0.0350291
Upper 95% Mean	2.3540651
Lower 95% Mean	2.2162798
N	353

Figure 5. Distribution of Logged Percent of Vacant Housing

Next, it's observed that there is a concentration of %NW values at both ends of the distribution. Such bimodal pattern can be "flattened" through logit transformation, log{%NW/(1-%NW)}, as

shown in Figure 6 and 7 below.

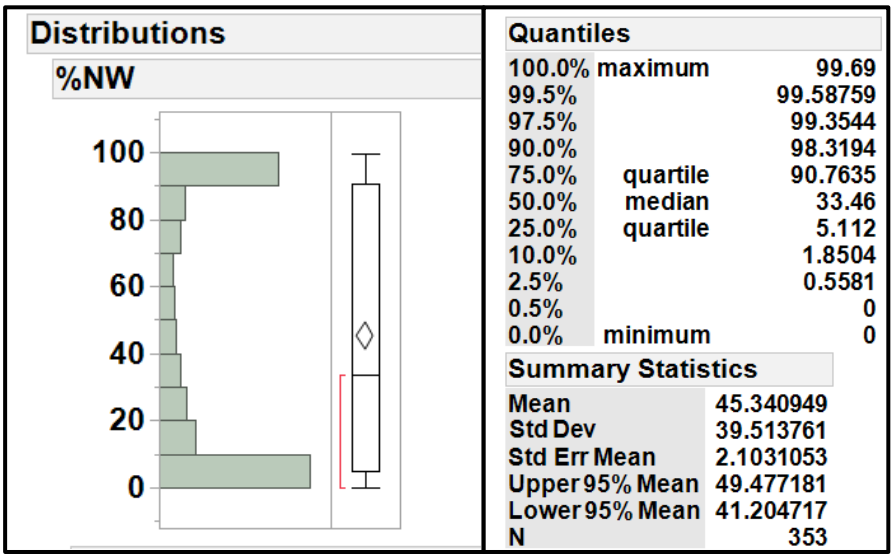


Figure 6. Distribution of Percent of Nonwhite

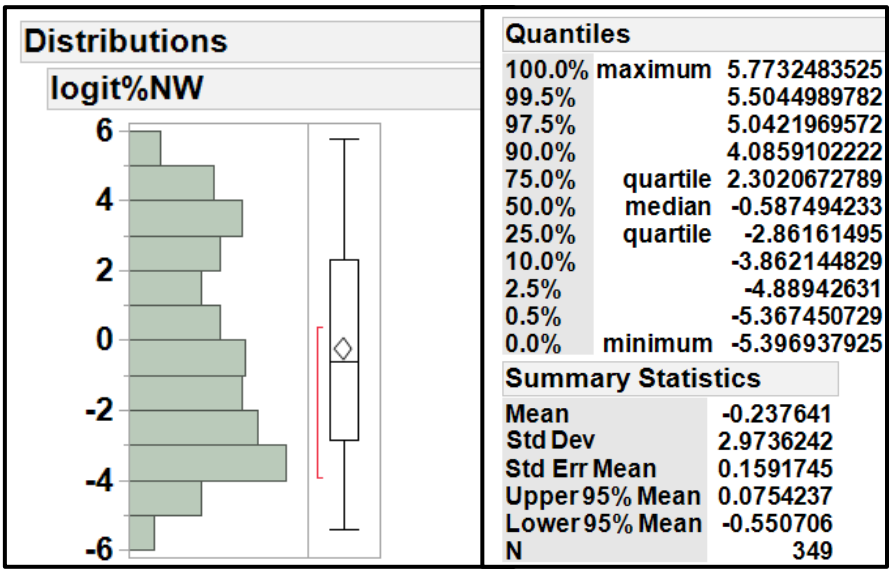


Figure 7. Distribution of Logit Percent of Nonwhite

2.2 Multiple Regression Model

Recall that we have assumed the existence of spatial autocorrelation among variables through visual observations between census tracts, so that we need a mildly different model to capture such spatial structure, called ***spatial autoregressive model***:

(2.2.1) $Y_i = \beta_0 + \sum_{j=1}^{k} \beta_0 x_{ij} + u_i \quad , \quad i = 1,...,n \quad ,$

where Y_i represents values of a random variable at each areal unit *i*, and $(x_{ij} : j = 1,...,k)$ represents a set of "explanatory" attributes of *i* that are postulated to have spatial influence on Y_i .In this case, x_{i1} and x_{i2} correspond to percent of vacant house units and percent of nonwhite respectively, and Y_i is the median housing values at each census tract *i* of 353 in total. Besides the key difference of such model from those used in previous studies (A[1-4]) that the variables is in terms of ***spatial areal units*** rather than ***point locations***, another main distinguishableness is the representation of residuals designated as u_i instead of ε_i . In this setting, each unobserved residual u_i at areal unit *i* (i.e., census tract *i*) is reasonably postulated to be influenced by u_j at neighboring areal unit *j*, where the positive spatial influence of u_i on u_j is described by spatial weights w_{ij} (explained in study 1). Hence, these residuals can be modeled as:

(2.2.2) $u_i = \sum_{j \neq i} \alpha(w_{ij}) u_j + \varepsilon_i$

where $\alpha(w_{ij})$ is some appropriate "influence" function dependent on w_{ij} , and ε_i represents the part of residual u_i that is not influenced by neighboring areal units (i.e., ε_i is designated as non-spatially-correlated residuals of spatially-correlated residuals u_i).

In particular, the simplest strategy for constructing the influence function $\alpha(w_{ij})$ is to assume it as a ***common scale (constant)*** factor ρ , so that the expression (2.2.2) is reduced to:

(2.2.3) $u_i = \rho \sum_{j \neq i} w_{ij} u_j + \varepsilon_i$

Hence if we denote the spatial weight matrix w_{ij} as *W*, the matrix form of (2.2.3) is given by:

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

Therefore, the spatial autoregressive model in terms of matrix can be written as:

(2.2.5) $Y = X\beta + u \quad , \quad u \sim N(0, V) \quad ,$

with both unknown parameter vector β and covariance matrix V as in Ordinary Kriging discussed in [A4]. Similarly, as we have figured out the solution to Ordinary Kriging, it's assumed that $V = \sigma^2 I_n$, and then the OLS estimate of β and associate OLS estimate of residuals u can be

obtained by:

(2.2.6) $\hat{\beta} = (X'X)^{-1}X'y$ and hence $\hat{u} = y - X\hat{\beta}$

Next, to apply the spatial autoregressive model, we continue to use the ***linear*** model below (with respective denotations of variables mentioned above, lnMV, %Vac, and %NW) to predict logged median housing prices according to percent of vacant house units and percent of white residents at each census tract *i*:

(2.2.7) $\ln MV(i) = \beta_0 + \beta_1 \%Vac(i) + \beta_2 \%NW(i) + u(i)$, $i = 1,...353$

The regression results are shown in Figure 8 below. The adjusted R-Square denoted as RSquare Adj = 0.3888 indicates that about 38% variances of logged median housing prices are accounted for by this model. Both p-values (denoted as Prob>|t|) of the two explanatory variables are smaller than 0.05, meaning the two variables %NW and %Vac are statistically significant, among which %NW (p-value < 0.0001) is a little bit more significant than %Vac (p-value = 0.0246).

Response lnMV

Effect Summary

Source	LogWorth	PValue
%NW	29.128	0.00000
%Vac	1.609	0.02461

Summary of Fit

RSquare	0.392259
RSquare Adj	0.388786
Root Mean Square Error	0.583972
Mean of Response	10.82936
Observations (or Sum Wgts)	353

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	77.03827	38.5191	112.9514
Error	350	119.35836	0.3410	Prob > F
C. Total	352	196.39663		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.434975	0.057812	197.79	<.0001*
%Vac	-0.010289	0.004558	-2.26	0.0246*
%NW	-0.01087	0.00087	-12.49	<.0001*

Figure 8. Results of Multiple Linear Regression Model

Next, we examine the relation of predicted logged median housing values and residuals, and find out that there are two clustered point groups of predictions at both ends as shown in Figure 9 below. In other words, the model yields more predictions of ***relatively higher and lower*** median housing values than the median value range.

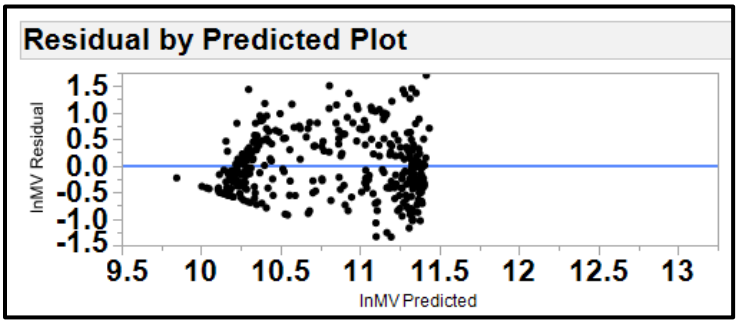


Figure 9. Predictions against Residuals

By plotting the predicted values against %NW in Figure 10 below, we can observe similar ***concentration*** patterns at both ends, which means there are more census tracts with r***elatively high*** percentages of white or nonwhite residents than census tracts with more ***evenly distributed*** racial compositions, as a result of the common social phenomenon that population from same cultures are more likely to settle in near neighborhoods and hence geographic population separation.

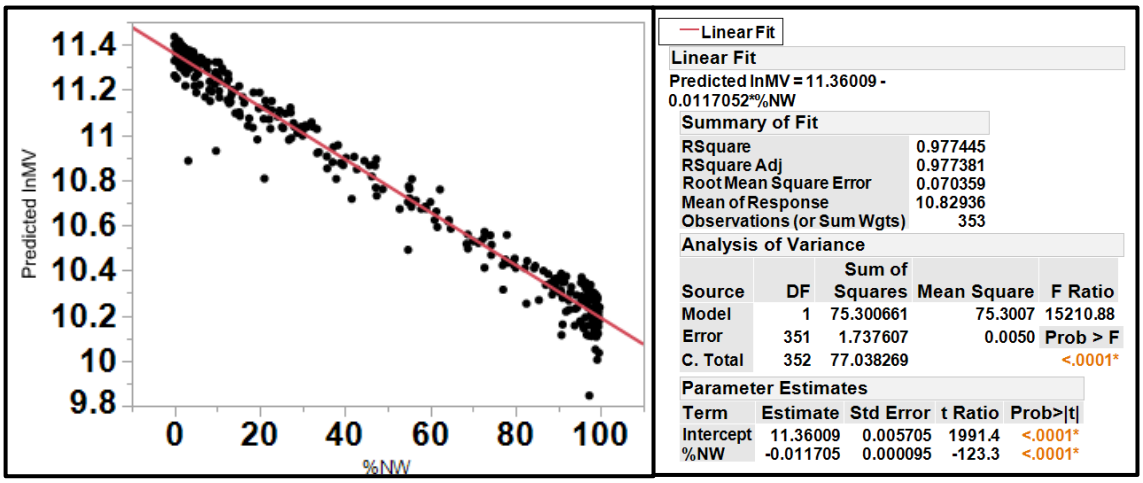


Figure 10. Predictions against Percent of Nonwhite

2.3 Test of Spatial Autocorrelation

The remaining task and also the primary objective of this analysis is to examine the spatial autocorrelation of residuals in the linear model in MATLAB using ***permutation test of spatial autocorrelation***, as the permutation procedure for G*-statistics in study 1. Recall that we have used an ***exponential*** weight matrix with ***normalized rows*** to model the spatial relations of counties in Eire that is measured by centroid distance in study 1. Accordingly, we will use another spatial weight matrix named ***nearest-neighbor*** matrix (or simply nn-matrix) and set the number of neighbors included in calculation of each tract as 4, that is, the ***first four*** nearest neighbors to a given tract (in centroid distance) will be considered to have spatial influence on (or spatially related to) this target tract. Here, weights of the four neighbors of each tract will be simply assigned 0.25 to meet the requirement of unit row sums equal to 1. Taking one census tract locating in the north Philadelphia as an example (shown in Figure 11 below), we notice that this tract (denoted in light red) is surrounded by 5 tracts, four of which are the particular first four nearest neighbors (denoted in light blue) identified by the nn-matrix here. It's reasonable to exclude the fifth nearest neighbor (denoted in grey) since it's further away from the target tract in centroid distance. However, it's indeed designated as a contiguous neighbor to the tract and will be included as well in the next report [A6] by using ***boundary-share*** weights.

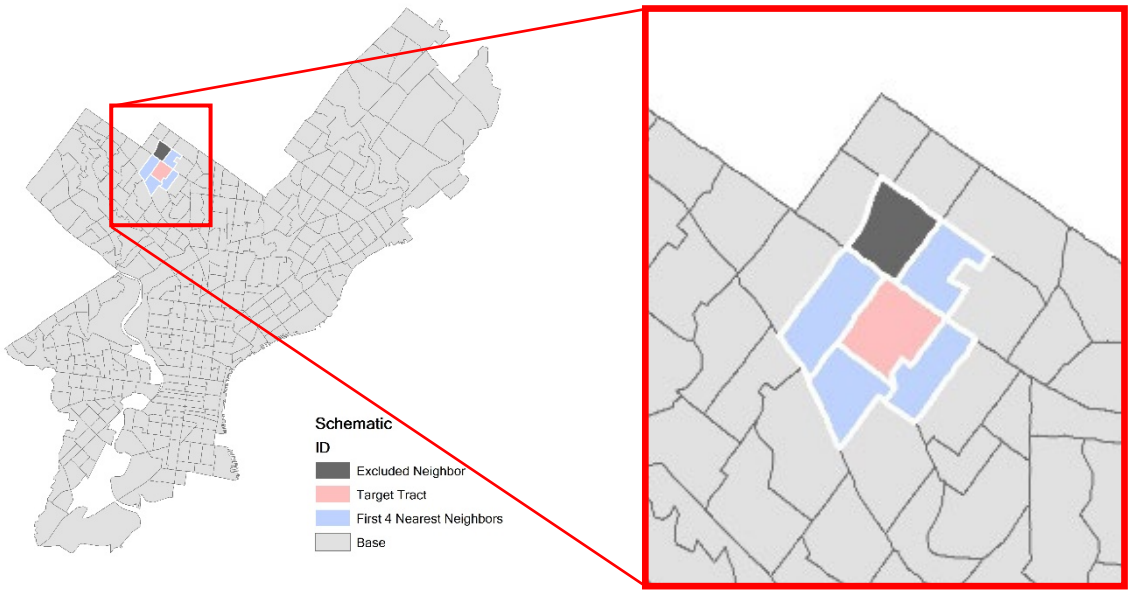


Figure 11. Target Tract and Nearest Neighbors (in Centroid Distance)

Based on the nn-matrix we have defined above, which is exactly the spatial weight matrix W in expression $u = \rho Wu + \varepsilon$ that describes the spatial influences of neighbors on each target census tract, we can test the ***null hypothesis*** of spatial autocorrelation:

(2.2.8) $H_0 : \rho = 0$ (there is no spatial autocorrelation among residuals u, that is, $u = \varepsilon$)

Usually, three alternative test statistics for spatial correlation are in common use, including ***rho statistic***, ***correlation statistic,*** and ***Moran statistic***.

2.3.1 Rho Statistic

This statistic starts by treating the OLS residuals $\hat{u}$ as a realization of u, and we can obtain the OLS estimate $\hat{\rho}$ of ρ in formula (2.2.4) $u = \rho Wu + \varepsilon$. Since for each linear model as $Y = xb + \varepsilon$, the OLS estimate of b can be evaluated by $b = (x'x)^{-1}x'y = \frac{x'y}{x'x} = \frac{x'y}{\| x \|^2}$, and $u = \rho Wu + \varepsilon$ can be regarded as a special case of it, where $b = \rho$, $Y = u$, and $x = Wu$, the OLS estimate $\hat{\rho}$ of ρ designated as ***rho statistic***, is given by:

(2.3.1) $$\rho_W = \frac{(W\hat{u})'\hat{u}}{(W\hat{u})'(W\hat{u})} = \frac{\hat{u}'W\hat{u}}{\hat{u}W'W\hat{u}} = \frac{\hat{u}'W\hat{u}}{\| W\hat{u} \|^2}$$

2.3.2 Correlation Statistic

Likewise, $\hat{u}$ is considered as a sample of u, and the same to $W\hat{u}$ and Wu. Since all sample pairs $\hat{u}_i$ and $W\hat{u}_i$ are correlated with the same (positive, negative, or zero) sign, the second test statistic is given by the sample correlation between $\hat{u}_i$ and $W\hat{u}_i$ (more detailed interpretations are given in [NoteBook, III.4-2, p.4-5]):

$$(2.3.2) \qquad r_W = r(\hat{u}, W\hat{u}) = \frac{\hat{u}'W\hat{u}}{\| \hat{u} \| \| W\hat{u} \|} \quad ,$$

which is designated as the ***correlation statistic***, or simply as ***corr***.

2.3.3 Moran Statistic

It turns out that the third statistic performs better with respect to testing H_0 (more reliable for detecting spatial autocorrelation) than the other statistics, which is designated as ***Moran statistic***, also known as ***Moran's I***:

$$(2.3.3) \qquad I_W = I(\hat{u}, W\hat{u}) = \frac{\hat{u}'W\hat{u}}{\hat{u}'\hat{u}} = \frac{\hat{u}'W\hat{u}}{\| \hat{u} \|^2}$$

2.3.4 SAC-Perm Test

Finally, we move on to the permutation procedure in MATLAB by employing essentially the same random permutation test based on Monte Carlo simulation as what we have done for G*-statistics in study 1. This testing procedure for null hypothesis H_0, which calculates ***all three alternative statistics*** given above instead of G*-statistics, is designated as the ***permutation test of spatial autocorrelation***, or more simply the ***sac-perm test***. Generally, for any random permutation $\{\pi = (\pi_1, ..., \pi_n), n = 353\}$, we compute all the associated rho statistic, correlation statistic, and Moran statistic for all permuted census tract, and compare the each observed statistic of three alternatives with the respective distribution of statistic values for *N* random simulations (here, *N* = 999).

If the observed value of rho statistic (correlation statistic or Moran statistic) has rank *k* (with rank 1 denoting the highest value) among all values (the observed value is also included in the 999 permutations), then the significance of spatial autocorrelation among residuals (in this case, predicted logged-median housing values) indicated by rho statistic (correlation statistic or Moran statistic) is again presented by the ***p-value***:

(2.3.4) $P = \frac{k}{N+1} i = \ , \ 1,...,n$

The results of sac-perm test are given as Figure 12 below:

RANGE OF RANDOM-PERMUTATION INDEX VALUES:

INDEX	Moran	corr	rho
MAX	0.1105	0.2251	0.4607
MIN	-0.1355	-0.3083	-0.7018

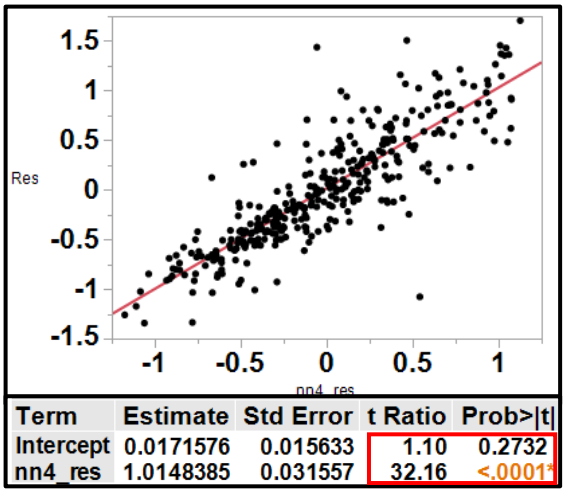
TABLE OF SIGNIFICANCE LEVELS:

INDEX	VALUE	SIGNIF
Moran	0.7357	0.0010
corr	0.8636	0.0010
rho	1.0137	0.0010

Figure 12. Results of SAC-Perm Test

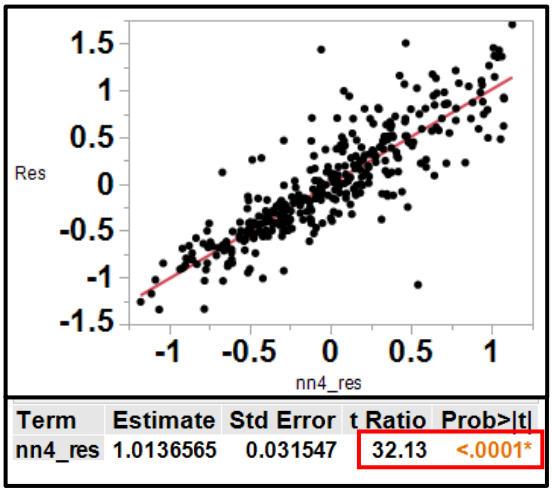
Notice that the ***significance levels*** for the three test statistics (Moran, corr, rho) are each ***maximally significant*** as expected, since all the significances (denoted as SIGNIF) are equal to = 0.001 that are well smaller than the widely used threshold ***α-level of significance*** = 0.05. Furthermore, they are ***higher*** and even ***much higher*** than the range of values displayed above for all of the 999 random permutations. For example, the rho value, 0.7357, turns out to be well above the range of values, -0.3083 to 0.2251 simulated by 999 permutations, whose associated p-value = 0. 001 indicates that there is (at most) a 1/1000 probability of observing such large a rho of 1.0137 if in fact there is no spatial autocorrelation. Same conclusion can be drawn followed by the other two statistics. To sum up, the null hypothesis of no spatial autocorrelation can be rejected for an ***alternative hypothesis*** that significantly ***positive*** spatial autocorrelation exists.

We then use JMP to repeat the analysis of spatial autocorrelation by regressing the residuals on their respective ***first four nearest-neighbor*** residuals (adding up the corresponding first four nearest-neighbor residuals of each and averaging them). The results are shown in Figure 13 below, where the statistically strong p-value < 0.0001 (denoted as Prob>|t|) indicates that these residuals are spatially correlated to their neighboring counterparts, which agrees well with the conclusion from sac-perm test.



Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0171576	0.015633	1.10	0.2732
nn4_res	1.0148385	0.031557	32.16	<.0001*

Figure 13. Regression with Intercept



Term	Estimate	Std Error	t Ratio	Prob>|t|
nn4_res	1.0136565	0.031547	32.13	<.0001*

Figure 14. Regression Without Intercept

It is also of interest to mention that such regression of residuals on nearest neighbor residuals is a special case of rho statistic, where W is chosen to be the first four nearest-neighbor matrix (i.e., $w_{ij} = 0.25$ if j is one of the first four nearest neighbors of i and $w_{ij} = 0$ otherwise). Though the intercept term must be ***zero*** in this model [NoteBook, III.4-3, p.2-3], it is usually not statistically significant. In this case, as shown in Figure 13 above, the p-value (denoted as Prob>|t|) of the intercept is designated as 0.2732, indicating its statistical insignificance. In addition, the t-value (denoted as t Ratio) of the intercept is designated as 1.10, which yields some degree of agreement with the rho statistic of 1.0137 in sac-perm test above. By contrast, results of residual regression without intercept are also shown in Figure 14 to the right. It's importantly noticeable that the ***t - values*** for the slope in both figures are almost identical.

3 DISCUSSION

In this study, we have introduced the spatial autoregressive model driving from OLS model, which performs well in capturing the structure of spatial correlation among variables of interest, and three widely used statistics for testing spatial autocorrelation based on Monte Carlo simulation. The objective of analysis we are concerned about here is median housing values in Philadelphia, which has been predicted by two independent variables percent of vacant housing and percent of nonwhite residents. It's noticeable that spatial autocorrelation is a common phenomenon with respect to spatial analysis and statistics, as we have found by observing the distribution patterns of each variable across census tracts in Philadelphia. However, it's assumed that such spatial relations should not exist among residuals. The violation of such assumption would happen when some sptial structures of the dependent variable is not included in our model, (i.e., median household income

and education levels are assumed to significantly related to the housing values). More specifically, larger errors of predictions (residuals) are more likely to emerge at areal units strongly influenced by the missing predictors. Since there predictors exhibit some degree of similarity over space, which would hence result in spatial autocorrelation among residuals. Usually, the simple OLS model cannot well capture such spatial structure. If so, spatial autoregression is a good choice to solve this problem, which is exact what we are going to do in the next study.

4 REFERENCE

Kevin C. Gillen, “Philadelphia House Price Indices”

Stephen J. McGovern, “Philadelphia’s Neighborhood Transformation Initiative: A Case Study of Mayoral Leadership, Bold Planning, and Conflict”

Pennsylvania Economy League, “Revitalizing Neighborhoods”

Lisa K. Bates, “Does Neighborhood Really Matter? - Comparing Historically Defined

Neighborhood Boundaries with Housing Submarkets”

[A1-A4] J.Wu (2019), “Assignment 1 - 4: Studies for ESE 502”