Research Proposal

Vivian Wang

**Introduction**

Housing prices are deeply intertwined with the social, economic, and infrastructural dynamics that define urban communities. They serve as a reflection of social equity, access to resources, and neighborhood development patterns. Philadelphia, with its rich history, socioeconomic diversity, and distinctive neighborhood identities, offers an important case for exploring these issues. Unlike megacities such as New York, Boston, or San Francisco, Philadelphia offers a blend of urban vibrancy and accessibility, while maintaining a housing market with huge variation.

This study seeks to uncover the key spatial and socioeconomic factors influencing housing prices in Philadelphia, and how these factors vary across geographically different censuses. By integrating survey data with spatial statistical methodologies, the research aims to contribute to a nuanced understanding of the city’s housing market dynamics.

**Literature Review:**

Most of the time, houses are measured in a hedonic way in the market, which includes but not limited to the number of bedrooms, whether or not it has a central heating, and in which year the house was built. This is based on the theory of hedonic pricing by Rosen (1974) and derived from Lancaster’s (1966) customer behavior theory that the goods’ individual characteristics create utility. Nowadays, researchers have realized the heterogeneity of house values instead of a uniform combination of characters, and done spatial regressions with hedonic independent variables (Lu et al. 2014). Other researchers have used multilevel models to account for the spatial effects when estimating hedonic functions on pricing, for example, Tan et al. (2019) implemented a community-county-province three-level model to analyze house inequality in China with socioeconomic independent variables.

**Motivation:**

My personal experiences in Philadelphia sparked my interest in exploring the city’s housing market. Two years ago, while searching for an apartment in the University City area, I was struck by the stark contrasts in rent prices, sometimes even within the same block. Philadelphia also stands out as a prime city for housing market analysis due to its unique urban composition. Furthermore, the city’s housing market is less influenced by factors like dormitory housing, which often distort price patterns in college-dense cities.

These variations raised my questions about the factors influencing housing costs and how they manifest spatially across neighborhoods. Given the limited area, I will apply spatial regression models rather than the multilevel models to find the associations, and both socioeconomic and hedonic variables on a census tract level will be considered, aiming to reveal patterns on a county level that define the city’s housing landscape.

**Data:**

This study employs data from the ACS 2022 5-Year Estimates at the census tract level from Social Explorer[[1]](#footnote-1), providing a snapshot of housing prices and associated factors across Philadelphia neighborhoods. The data in this report includes demographic statistics of the residents, such as population, race and household type; social-economic indices, such as household income, education level, and means of transportation; and housing features, such as occupancy status, rent, built year. The shape file used is from OpenDataPhilly[[2]](#footnote-2) census tract 2010, and merged on FIPS code to the ACS file.

**Dependent variable:**

Data for median house value, the dependent variable, is missing for the airport, national park area, neighborhood park areas, mountains, rivers and golf courses scattering in the county of Philadelphia. These areas can be seen on the map shown below (Graph 1) as the blank white polygons. These areas are also excluded from the analysis this project conducted, using a total of 348 observations from the rest of the census tracts in the county of Philadelphia.

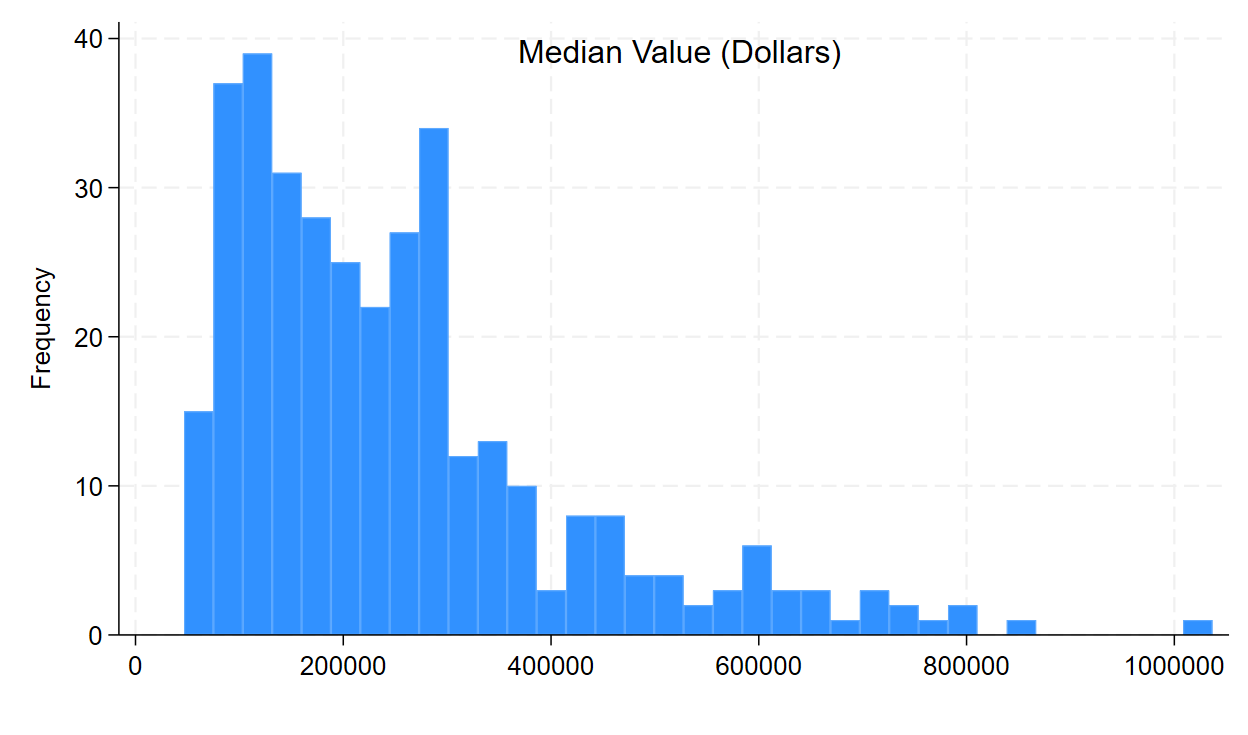
Descriptively, the median house value of each census tract has a mean of $253,444.3 with a $165,584.2 standard deviation, minimum value of $46,700 and maximum value of $1,036,700. The values are highly skewed to the right, displayed in the left-hand side of graph 2. To correct the skewed distribution, I did a log transformation on median house value to compresses large values and spreads out smaller ones, making the variable closer to normal distribution and enhancing the interpretability of the model. The distribution of the natural log of median house value can be shown on the right-hand side of Graph 2. The median house values of each census tract are classified based on magnitude numbers and filled with different blue colors, and the darker ones stand for higher values. Those with missing data were left blank.

Graph 1: Median house value in Philadelphia

A map of a neighborhood

Description automatically generated

Graph2: Distribution of median house value and natural log of median house value

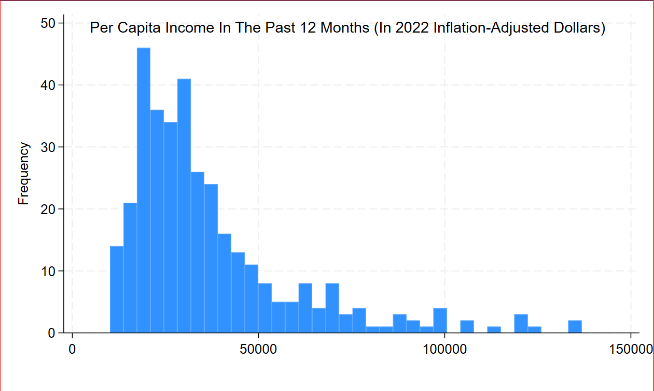
 A graph of a house value

Description automatically generated

**Independent variables:**

10 independent variables are picked from three different aspects, including the demographic features of the resident of the census, the socioeconomic factors, and the hedonic features of the houses and the neighborhood. I chose total population of the census tract and ethical components like percentage of hispanic population and percentage of black population as the demographic features. The socioeconomic factors are percentage of residents that has a higher education degree, which is defined as master and PhD degrees beyond college, natural log of per capita income, and percentage of people that has a mortgage for their houses. The natural log transformation of per capita income serves as the same function of the natural log transformation of the median house value, to adjuct for the right-skewness and improve linearity and interpretability.

Graph 3: Distribution of per capita income and natural log of per capita income

A graph of a graph

Description automatically generated

The hedonic factors of the house and neighborhood include percentage of houses that are vacant, the average age of the house calculated from the year when the houses were built, the percentage of people that work near home, which is the sum of the percentage of people that work from from home, walk to work and ride bicycles to work, and lastly the 5-year aggregated minute of transportation to work for those who is over the age of 16 and do not work from home.

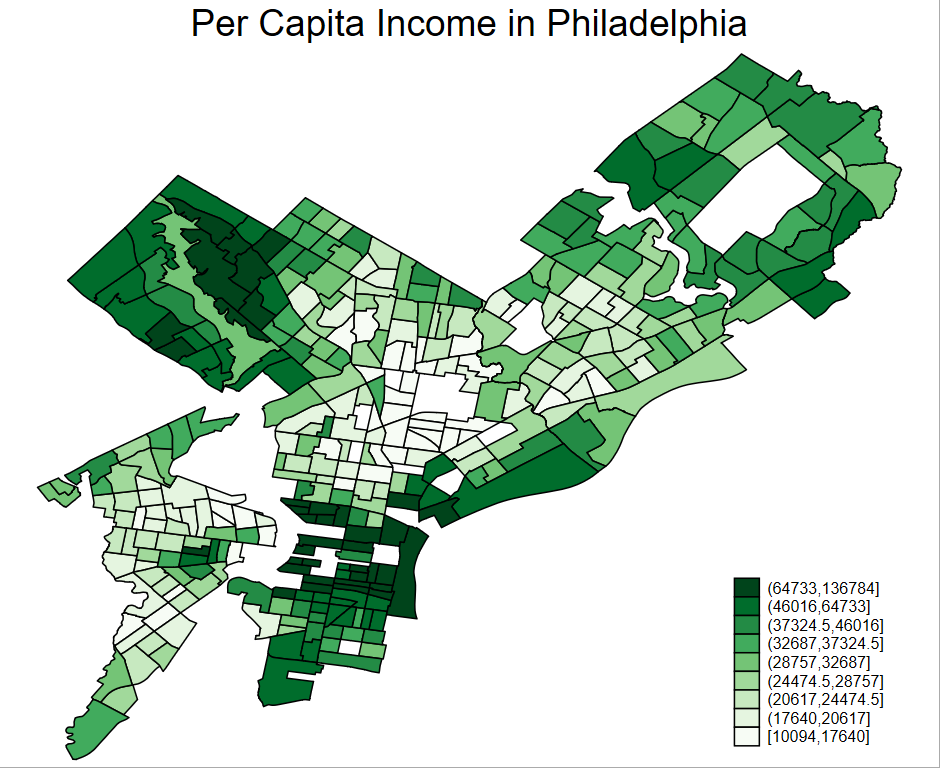
Graph 3: Geographical distribution of the independent variables

A map of a neighborhood

Description automatically generatedA map of a neighborhood

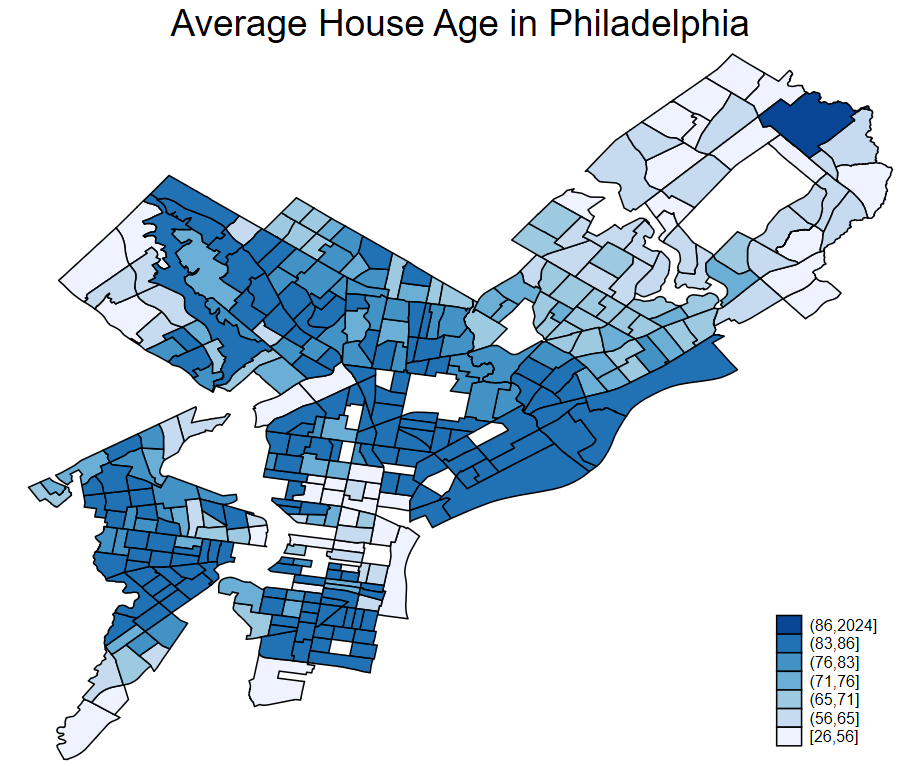
Description automatically generatedA map of different colored squares

Description automatically generatedA map of a city

Description automatically generatedA map of a neighborhood

Description automatically generated

A map of a neighborhood

Description automatically generatedA map of a neighborhood

Description automatically generatedA map of a city

Description automatically generated

**Methodologies:**

First, the spatial autocorrelations of the variables will be examined by global Moran’s I and Local Moran’s I (LISA) for the hot spots. The global Moran’s I is calculated with an inverse distance based spatial weight matrix in the Stata function of spatgsa. The clustering maps generated from LISA can be found in the appendix.

Next, I will run an Ordinary Least Squares (OLS) regression as the baseline model, with the residuals of the model examined for spatial autocorrelations with Moran’s I. Then, I will use the spatial regression models, including spatial autoregressive model (SAR), spatial error model (SEM), and a combination of spatial autoregressive model and spatial error model (SAC). The spatial models will be based on the first order queen weight matrix, illustrated in graph 4.

The models are specified as follows:

OLS model:

(1)

SAR model:

(2)

SEM model:

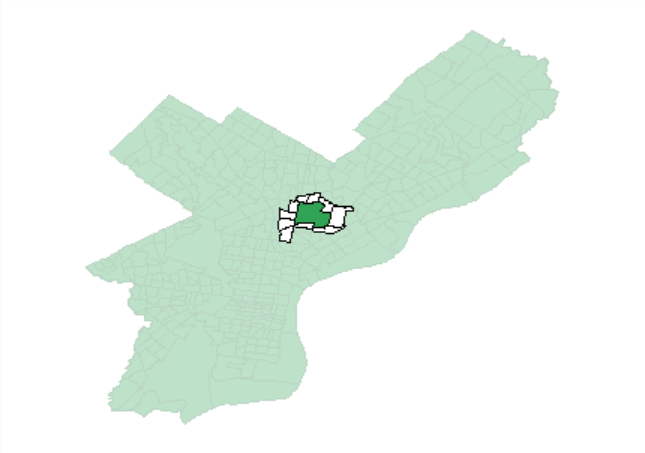
(3)

SAC model:

(4)

where , X is the matrix of independent variables with per capita income transformed to natural log. is the intercept term and is the vector of the coefficients, is the error of the regression models. is the first order queen continuity weight matrix, represents the strength of spatial dependence of the dependent variable, and represents the strength of spatial dependence of the error term.

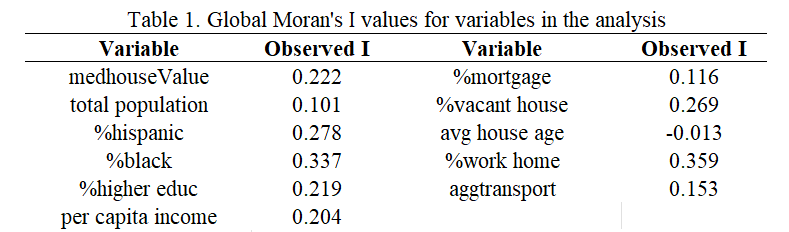
Graph 4: An illustration of the queen continuity matrix



For each model, the Moran’s I of the residuals will be checked with the inverse distance weight matrix, and the Akaike Information Criterion (AIC) will be compared for model selection and evaluation. Based on the diagnostics and a comprehensive understanding of the data, a final model will be determined, with its interpretations of the coefficients and the data of Philadelphia.

**Spatial Autocorrelation**

Global Moran’s Is are calculated with an inverse distance weight matrix, where the areas are assigned 0 when the distance is larger than 0.1 unit of distance. The 1-tail test and values of global Moran’s Is of the dependent and independent variables can be found in Table 1. All of the 11 variables have significant global spatial autocorrelation in the county of Philadelphia.



Local Moran’s Is along with the cluster maps can be found in Graph 4.

**Preliminary Results:**

In the outcome of the OLS model, percentage of people that have higher education degree, per capita income, average house age, percentage of people that work near home, and minute of transportation to work are positively related to median house value, and total population, percentage of Hispanic population, percentage of black population, percentage of people that have mortgage and percentage of house vacant are negatively associated with median house value. Among the variables, total population, average house age and minute transportation to work do not have a statistically significant relationship with median house value.

The residuals of the OLS model have a 0.039 Moran’s I and significant spatial autocorrelation. This is also the validation and motivation of the consequent spatial regression models.

For the spatial autoregressive model, percentage of people that work near home becomes insignificant compared to the OLS model, with a 0.097 p-value. The spatial autocorrelations of the residuals are not significant. For the spatial error model, the percentage of vacant house becomes insignificant compared to the OLS model. The Moran’s I of the residuals is 0.095 and is significant, indicating significant spatial autocorrelations in the residuals of the SEM model. For the combined spatial regression model, percentage of people that work near home becomes insignificant compared to the OLS model. The spatial autocorrelations of the residuals are not significant.

Residual Moran’s I, log likelihood and AIC of the models are considered for the evaluation process. OLS and SEM are having significant spatial autocorrelations in their residuals, indicating that the models failed to adequately explain the spatial dependence within the data, while SAR and SAC accounted for the spatial patterns well. Among the spatial regression models, SAR has the lowest AIC, which signals the best fit of the data. The diagnostics are shown in Table 2.

A black text with black text

Description automatically generated

**Discussions:**

According to the diagnosis, SAR is the best fit of the data. There is no spatial dependence in the unexplained errors, and the median house value of Philadelphia can be partially accounted for by spatial dependence of the neighboring census tracts’ median house value.

However, I believe the independent variables have not yet explained all the spatial dependence of the factors influencing house value in Philadelphia, and there must be other measurements such as access to education and healthcare, and proximity to green places. Regarding the difference of AICs between the SAR model and the SAC model is relatively small, the fitness and complexity of the two models are similar, additionally proved by the LR test shown in Table 3 below. In the LR test, we fail to reject the null hypothesis that the SAR model is as well as the SAC model. Therefore, despite the slightly smaller AIC of SAR, SAC is a preferred model to explain the variance and spatial patterns of the median house values of Philadelphia for unpresented factors from the model.

Table 3: LR test of SAR and SAC

A close-up of a test

Description automatically generated

Consequently, the coefficients of the combined spatial model can be interpreted in two ways, for the log transformed per capita income and other variables that are not transformed, and there will be direct effects and indirect effect, also known as spillover effects. For 1% increase in per capita income, the median house value of that census increases by 0.21%, and the median house value of neighboring census tracts increase by 0.20%. For variables not transformed, take percentage of Hispanic population as an example, for every 1 unit increase in the percentage of Hispanic population, the median house value of that census tract decreases by a factor of 0.55 or 45%, and those of the adjacent census tracts decrease by a factor of 0.56 or 44%.

In terms of prediction, for example, if people want a house that is in a neighborhood with 10% Hispanic population, 10% black population, 30% of people with a higher education, 10% people having mortgage for houses, only 5% of the house vacant and 40% of people work near home, which is not an existing setting in Philadelphia and therefore cannot be tested, the value of the house should be around $882,046 (e13.69).

Table 4: Showcase of prediction power of the model

A screenshot of a computer

Description automatically generated

**Conclusions:**

There are significant local spatial autocorrelations in the house values and demographic, socioeconomic and hedonic factors in the county of Philadelphia. Despite the slightly lower AIC of the SAR model, SAC models better fit the nature of the data, with a belief that besides the spillover effect, there are spatial patterns remaining unexplained in the residuals of the model. According to SAC, percentage of Hispanic population, percentage of black population, percentage of people with a higher education degree, per capita income, percentage of people with a mortgage, and percentage of vacant house are some of the key determining factors of house values.

**Limitations:**

First of all, the weight matrices used for the spatial regression models and the Moran’s I calculations are not identical. The Moran’s Is are based on an inverse distance weight matrix, where the width of the band is 0.1 unit, while the spatial models and LISA maps are built based on first order queen continuity matrix. Stata has different but incompatible systems when conducting spatial regressions and calculating the value of Moran’s I, which poses difficulty to users on being coherent to do both.

Secondly, there are missing values in either the ACS data or the shapefile. I dropped all the missing values and built the models on a relatively much smaller sample size. This will also lead to less neighbors of some of the census tracts. The missing census sometimes have schools, parks or other public infrastructures that might be a factor of pushing house values of their neighboring census tracts high.

Last but not least, there might be demographic features, socioeconomic factors and hedonic factors that are influencing house value but not included in the model. Moreover, panel data on a census tract level fail to consider factors from a larger perspective, such as segregation, social policy and health care that affect the utility or general happiness of people living in a certain area, which can only be measured by a multilevel model.

**References**

Brunauer, Wolfgang, Lang, Stefan, & Umlauf, Nikolaus. (2013). Modelling house prices using multilevel structured additive regression. Statistical Modelling, 13(2), 95–123.

Lu, B., Charlton, M., Harris, P., & Fotheringham, A. S. (2014). Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science*, *28*(4), 660–681.

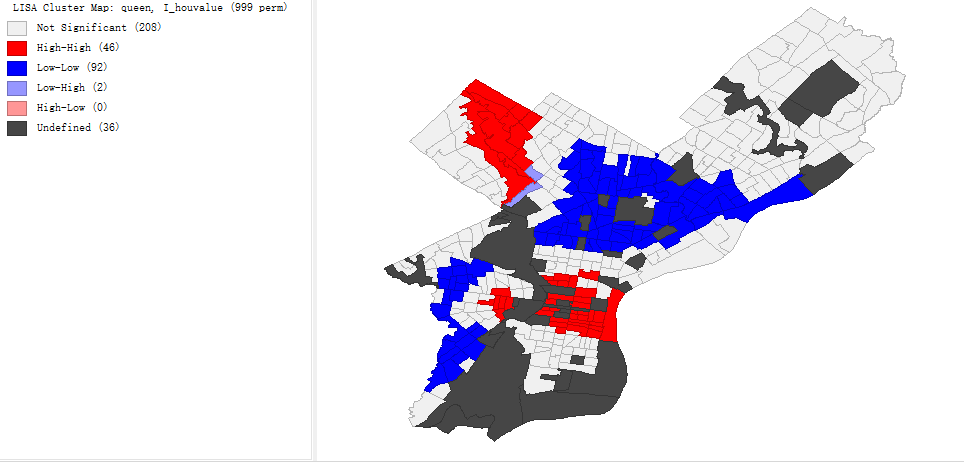
Tan, Ronghui, He, Qingsong, Zhou, Kehao, Song, Yan, & Xu, Hengzhou. (2019). Administrative hierarchy, housing market inequality, and multilevel determinants: a cross-level analysis of housing prices in China. *Journal of Housing and the Built Environment*, *34*(3), 845–868.

Helbich, Marco, Brunauer, Wolfgang, Vaz, Eric, & Nijkamp, Peter. (2014). Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria. *Urban Studies (Edinburgh, Scotland)*, *51*(2), 390–411. https://doi.org/10.1177/0042098013492234

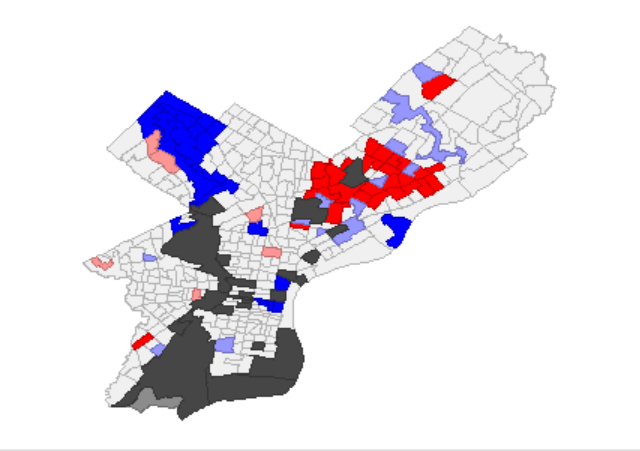
**Appendix**

Local Spatial Autocorrelation (LISA)

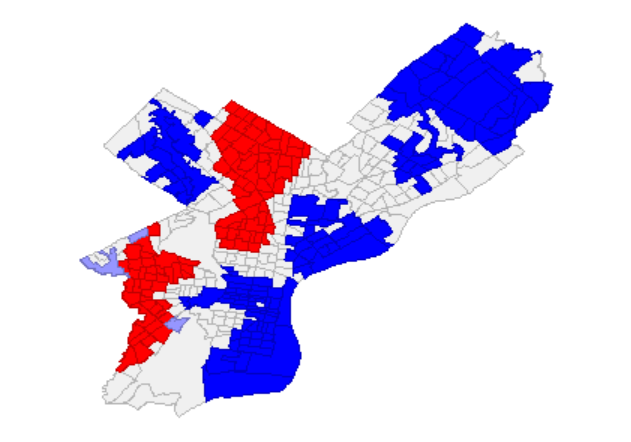
Median House Value:



Total population Percentage of Hispanic population

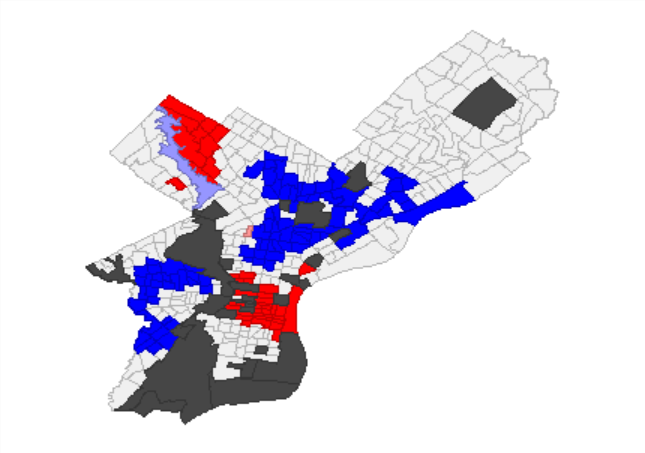
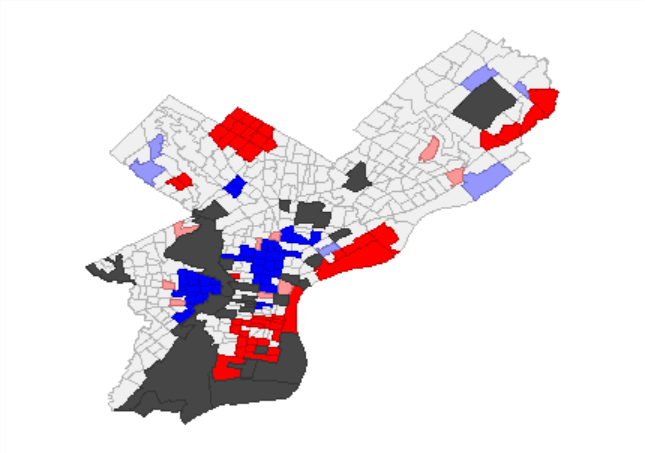
 

Percentage of black population Percentage of people with higher education degree

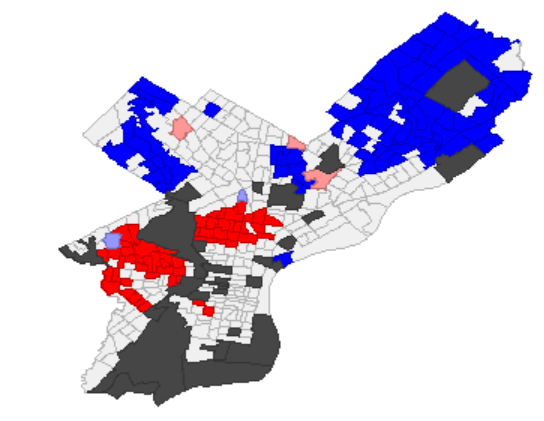
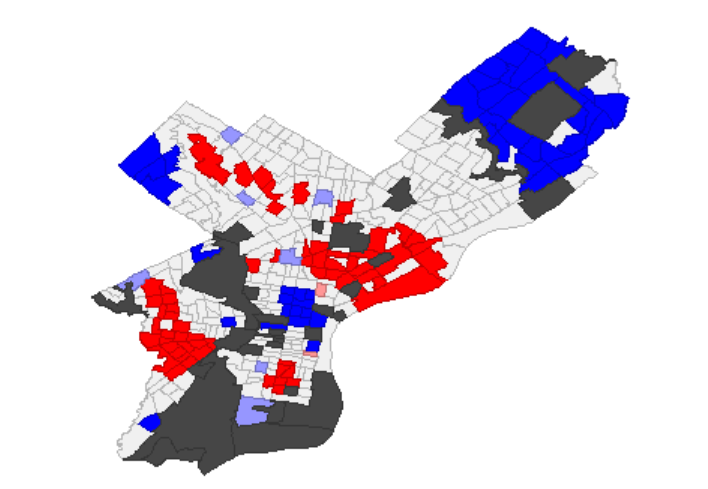
 A map of different colored areas

Description automatically generated

Per capita income Percentage of people with mortgage

Percentage of vacant house Median age of house

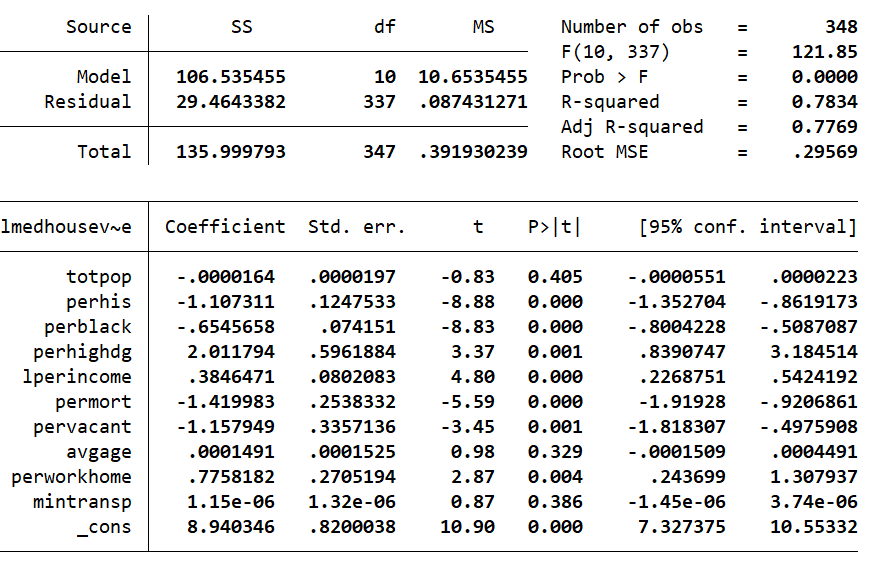
Percentage of people work near home Minutes of transportation to work

A map of a city

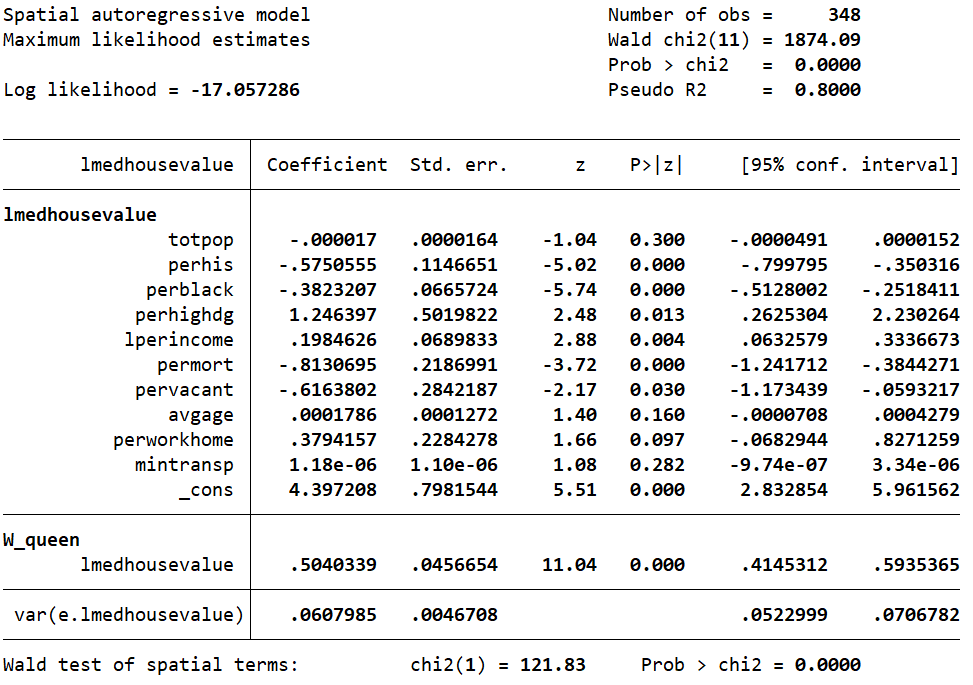
Description automatically generated A map of a city

Description automatically generated

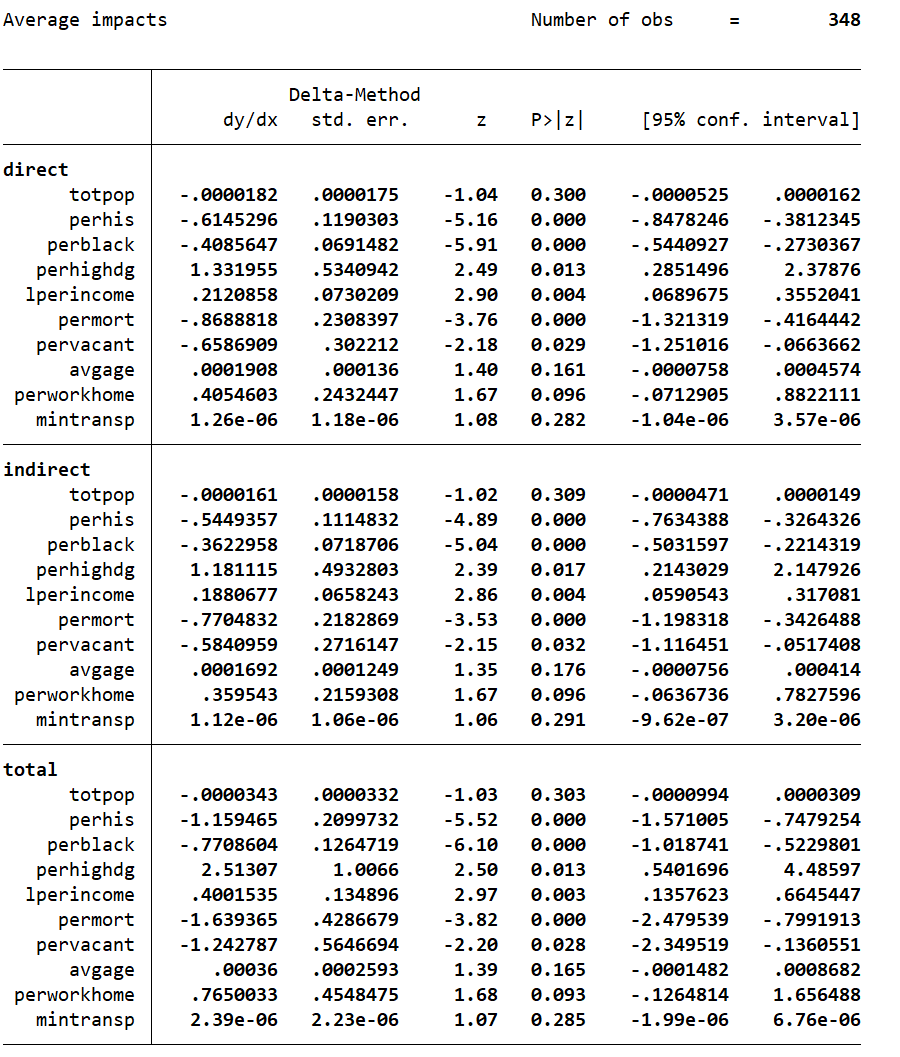
OLS regression model:



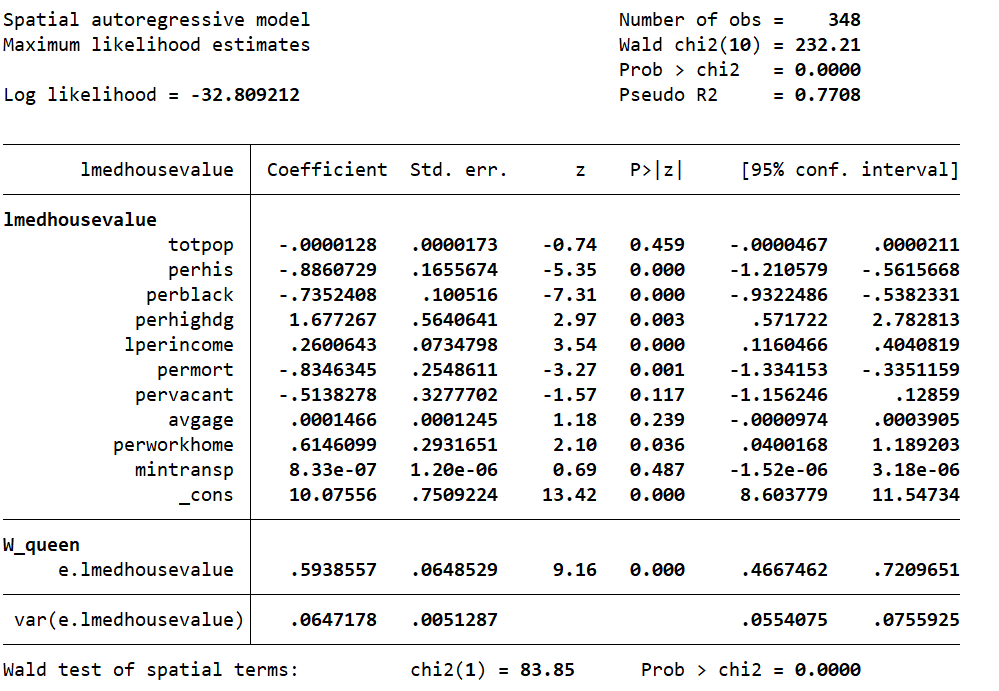
Spatial Autoregressive Model:



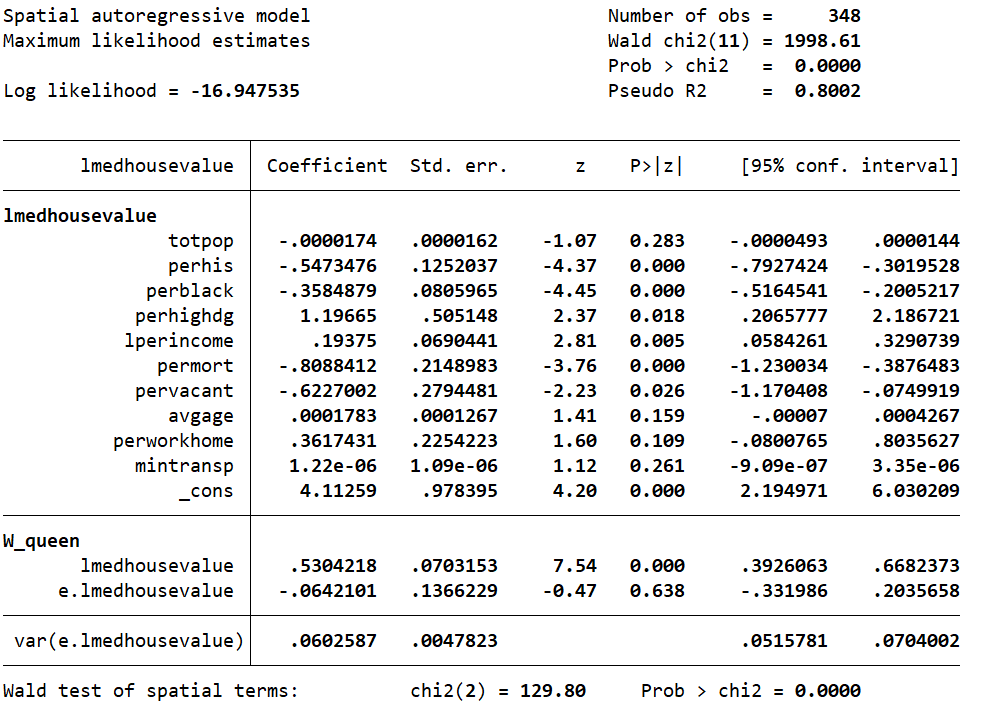
Impacts of Spatial Autoregressive Model:



Spatial Error Model:



Combined Spatial Model:



A table of numbers with black text

Description automatically generated

1. http://www.socialexplorer.com/pub/reportdata/HtmlResults.aspx?reportid=R13771388 [↑](#footnote-ref-1)
2. https://opendataphilly.org/datasets/census-tracts/ [↑](#footnote-ref-2)