Housing Price Analysis in Amsterdam Under Socio-demographic Factors

Jiatong Cui (2782278)
Jin Dai (113371)
Chenjun Zhang (1214837)

April 15, 2022

1 Introduction

Housing acts as a focus of economic activity, a symbol of achievement, social acceptance, and an element of urban growth [1]. There have been many geospatial studies on factors which influence housing prices and there is evidence to suggest a strong influence from the area of the house as well as the appearance of the surrounding area [5], [3]. Prior research also suggests that population density is a strong predictor for housing prices [4]. The present study aims to focus on how socio-demographic factors of the populace around an house affects the price of the house for Amsterdam in the Netherlands. Among these numerous socio-demographic factors, we aim to study the effects of resident age, and crime rate on the distribution of house prices. We hypothesize that these factors are driving home prices in the local market and explore the direction and extent of their respective impacts on home prices.

2 Study Area

Past research shows there is a significantly negative correlation between crime rate and housing price that indicates people are willing to pay more for a location with less crime [7]. In today's society, because of rising house prices, estate has become a precious "luxury". Buyers will usually go to enormous lengths to ensure the home they choose is perfect [6]. At the same time, since rising crime rate has been shown to have long-term impact on housing prices in residential areas, we can infer that neighborhoods with low crime rates will be in higher demand, causing prices to increase.

An individual's age is often linked to other characteristics of the resident, like income, marital status, and education [2]. Green (1995) defined that holding all else constant, the demand for housing tends to be flat or rise slightly with age. The present study aims to verify whether age has a significant impact on housing prices.

3 Data Collection

The dataset for house pricing comes from a real estate website scraped in August 2021. The dataset contains 924 pieces of data, which record the address of the house, the location, the number of rooms in the house, the area of the house, and the sale price in Euros. There are four records with missing values, and two locations of houses not located within Amsterdam, therefore we deleted them.

The dataset for crime counts and residential ages are obtained from the Gemeente Amsterdam website in open geodata form for the year 2021. The datasets divide the data into the same sub-regions. These sub-regions that divide the Amsterdam area are: Centrum, West, Nieuw-West, Zuid, Oost, Noord, Zuidoost. While Westpoort is also a region, there were no housing listings in this region, likely as it is primarily an industrial area. In the crime-related dataset, the total number of crimes is taken as suspects apprehended for the

year, divided in different ages. There are two different age division methods, one of which is divided into 12-17 years old, 17-24 years old, and 25+ group; the other is more comprehensive, it is divided into the 12-24 and 25+ groups.

The dataset for resident age was similarly taken from 2021. The main data of this dataset is the population counts for different age ranges separated by different regions of Amsterdam. The dataset is divided into eight age groups: 0-3, 4-12, 13-17, 18-22, 23-24, 25-49, 50-64 and 65+.

4 Method

We built a regression equation based on house price in Euros against the average age of the region, suspects apprehended in the region, population of the region, area of the house in square meters and the number of rooms the house has.

To answer the research question, we decided to create a spatial lag model and a GWR model to explore the elements which affect housing price in Amsterdam. Before modeling, we first preprocessed our data. Both 'housing price' and 'city sections' were transformed into spatial objects with their coordinates, and then spatially joined to combine them as a spatial form of data. Afterwards, we aggregated this data and merged them with the 'crime' and 'age' datasets. For crime we used the total suspects for each region, and for age we used the average age for each region. The population variable was converted into a proportion of the total population of Amsterdam to highlight the relative populations of different sections.

After processing the data, we first created a simple linear regression and then calculated Moran's indicator to examine the existence of auto correlation. After that, we created a spatial lag model to account for the spatial correlation. Furthermore, we also built a GWR model with fixed kernels, which will help us explore the discrepancy of housing price across the whole city.

5 Results

The following regression equation was used as the basis of the study:

$$\begin{aligned} Price = \hat{\beta}_0 + \hat{\beta}_1 Crimes + \\ \hat{\beta}_2 AverageAge + \hat{\beta}_3 Population + \\ \hat{\beta}_4 Area + \hat{\beta}_5 Room + \hat{\epsilon}_i \end{aligned} \tag{1}$$

The assumption of independence was checked with Variance Inflation Factor (VIF), using the threshold of five as acceptable.

Table 1: VIF values for regression equation

Crime	Average Age	Population	Area	Room
3.119926	2.937877	1.977857	3.000986	2.957103

The results in Table 2 displayed acceptable VIF values for all variables(VIF > 5), suggesting that the variables do not display multicollinearity.

Table 2: Moran's I for OLS

Moran I statistic	Expectation	Variance
0.1890325331	-0.0010905125	0.0003572126

Moran's I for the OLS was also significant ($p<.05$), suggesting the presence of spatial auto-correlation within the variables. The positive value suggests that variable are likely to be spatially clustered.

Table 3: Linear Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4237115.7486	394550.4607	-10.74	0.0000
Crime	-165.2142	42.9005	-3.85	0.0001
Average Age	99664.5982	9178.7635	10.86	0.0000
Population	3972373.6675	367745.4563	10.80	0.0000
Area	8359.2586	255.2732	32.75	0.0000
Room	-29255.3873	9191.5649	-3.18	0.0015

The results for the OLS model showed very significant correlation ($p<.001$) between all the independent variables. Incidence of crime was negatively correlated, whilst the average age, population and area were positively correlated. Interestingly, number of rooms was also negatively correlated, suggesting that houses are worth less when they contain more rooms.

Table 4: Spatial Lag Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4057983.478	395966.404	-10.2483	0.0000
Crime	-122.566	45.028	-2.7220	0.0065
Average Age	93456.477	9325.943	10.0211	0.0000
Population	3634811.158	380798.652	9.5452	0.0000
Area	8298.325	254.346	32.6261	0.0000
Room	-28688.985	9145.245	-3.1370	0.0017

The results of the spatial lag model display similarly high significance ($p<.001$) with the same relationships found as in the OLS model.

Using a fixed bandwidth of 1.298 as determined by cross-validation, a GWR was carried out on the regression equation, yielding the following results for regression effects across Amsterdam:

The local R^2 value from the GWR displayed consistent values around the central area in Amsterdam ($\approx .75$) as shown in Figure 1. The Northern region displayed the lowest value of $\approx .55$ whilst the Eastern and Western regions

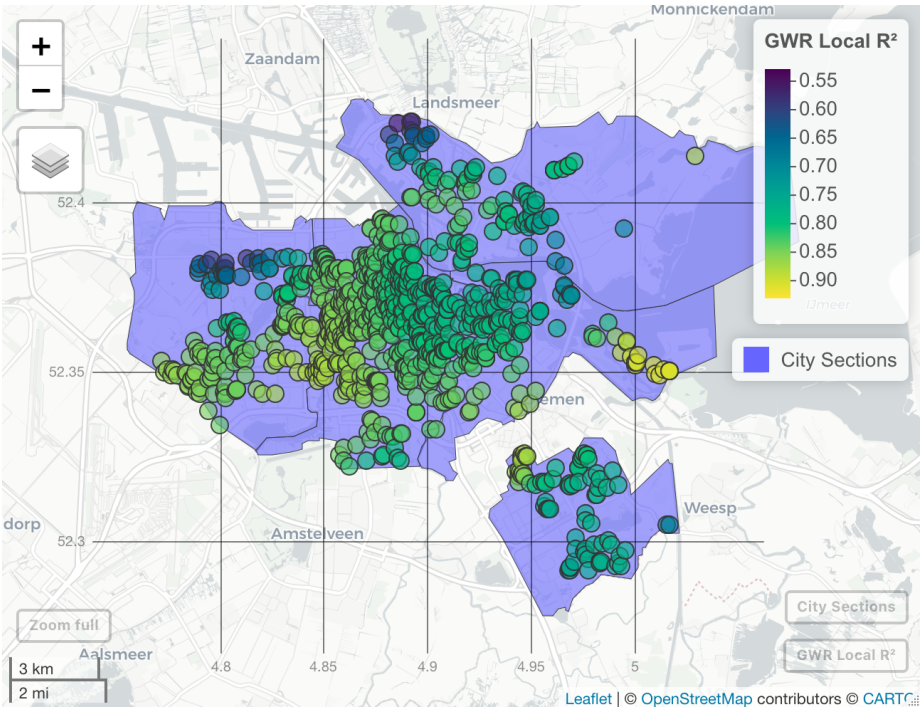

Figure 1: GWR local R^2 values for houses in Amsterdam

contained the clusters with the highest values of $\approx .90$. The quasi-global R^2 was quite high (=.85), suggesting a value of variance explained.

Table 5: AIC of models

OLS Model	Spatial Lag Model	GWR Model
25462	25453	25196

The AIC comparison between the three models suggested that the GWR was the best performing model as shown in Table 5.

Interestingly, Figure 2 shows that crime was only a predictor in certain areas of Amsterdam, though for the other factors there was significance throughout Amsterdam.

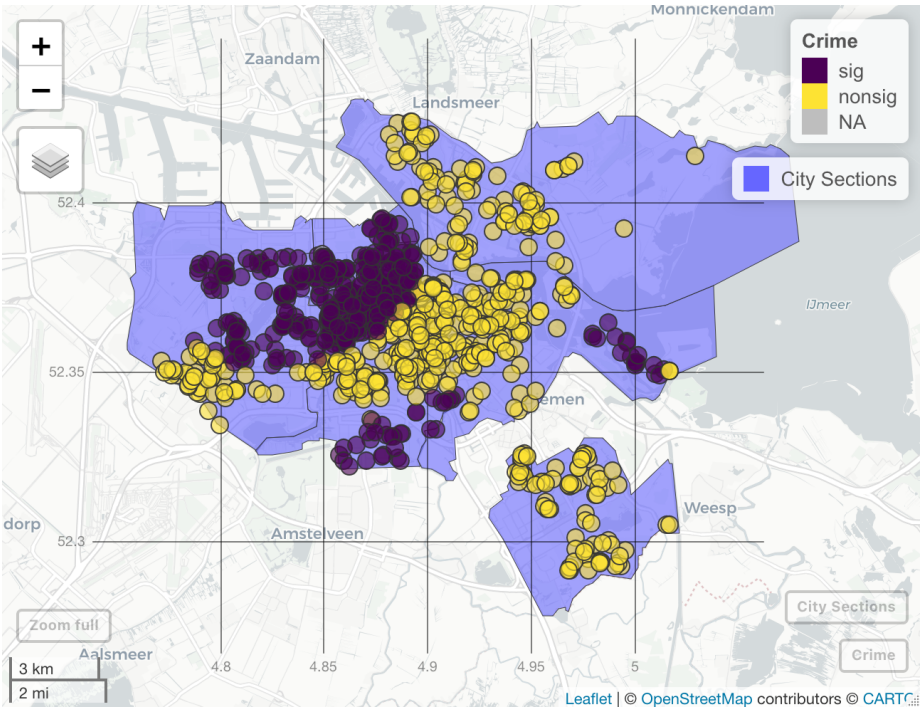

Figure 2: GWR significance of crime as a predictor

6 Discussion

The results of the analysis illustrate that the number of crimes in the neighborhood and the number of rooms of the house will affect the housing price in a negative way, indicating that when there are more crimes in the surrounding area, the housing price will decrease. This can be interpreted to mean that a house in an area with higher occurrence of crime suggests that the area is unsafe and therefore the houses are less desire-able. In terms of the number of rooms, people tend to choose house with 3 or four rooms, and when the number of rooms increase it may cause a waste of space. Therefore, houses with abundant rooms will be less popular. The area of the house, the average age, and the population in the surrounding area have a positive correlation with the housing price, which means that when the average age is larger, the price will increase and this relation is the same for the population and the housing area. Among the factors we have considered, population was the most influential variable and crime was the least influential.

A possible limitation of the current study was that the housing data was scraped from one real estate website due to many websites blocking web-scrapers. Another limitation would be that the average age of the area is difficult to in-

terpret as it does not indicate the nuance of areas containing a lot of children or elderly.

Additionally, we found that although the spatial lag model is able to account for spatial dependency, it has a higher AIC when compared to the GWR model. From the results, we can conclude that the local regression in Northern region fits the model less, whilst the Eastern and Western regions have a better fit for the regression. Given that the cluster of houses with the lowest correlation belong in a small area to the North, there could be a common explaining factor not considered such as sparse population or inaccessibility.

7 Conclusion

In the regression model, the population was the most important factor for explaining the housing price. Interestingly, this does not support the general expectation that the housing price is related more to its area and number of rooms. However, the significance and correlations of crime and population supports prior research on the subject of housing price variation [7], [4]. The results confirm the power of the location of the house aside from the pragmatic factors (i.e. distance to centre and public transport), demonstrating that the age of a region and the number of other people are also significant factors to increasing or decreasing the purchase price of the house.

References

[1] Ahmad Ariffian Bujang, Hasmah Abu Zarin, and Norhaslina Jumadi. "The relationship between demographic factors and housing affordability". In: *Malaysian Journal of Real Estate* 5.1 (2010), pp. 49–58.

[2] Richard Green and Patric H Hendershott. "Age, housing demand, and real house prices". In: *Regional Science and Urban Economics* 26.5 (1996), pp. 465–480.

[3] Yuhao Kang et al. "Understanding house price appreciation using multi-source big geo-data and machine learning". In: *Land Use Policy* 111 (2021), p. 104919.

[4] Sheng Li et al. "Understanding the Effects of Influential Factors on Housing Prices by Combining Extreme Gradient Boosting and a Hedonic Price Model (XGBoost-HPM)". In: *Land* 10.5 (2021), p. 533.

[5] Laurent Santos and Rui Jiang. "Spatial Analysis for House Price Determinants." In: (Apr. 2020).

[6] Scott Simpson. *Do local crime rates affect house prices?* Dec. 2013. URL: `https://www.flyinghomes.co.uk/blog/crime-levels-property-house-prices/`.

[7] Ran Tao, Hong Zhao, et al. "Crime Rate, Housing Price, and Value of A Statistical Case of Homicide"". In: *Economics Bulletin* 39.3 (2019), pp. 1727–1739.