Empirical Evidence Based on Geographic Regression Discontinuity Analysis of Housing in Guangzhou School District

Yulu Ouyang, Keyue Sun*
School of Economics, Jinan University, Guangzhou, Guangdong 511400, China
*Corresponding author: sky944516@stu2018.jnu.edu.cn

Abstract. In recent years, the "school district housing boom" has become a general concern as it disrupted the housing market, undermining educational equity and class mobility. In view of the real problem of housing premium in school districts, this paper selects the housing data of primary school districts in Guangzhou as a research sample, quantifies it and discusses the impact of education supply on it, so as to provide references and suggestions for subsequent policy introduction. The main research conclusions of this paper are as follows: There is a premium for housing in Guangzhou's school districts. After excluding the influence of location and physical factors, it is calculated that the housing price from Longdong School District to Yongping School District has increased by about 42.3%.

Keywords: BP neural network prediction; optimization decision; risk control; AHP.

1. Introduction

With the development of society and "education equity" gradually put on the agenda, China in 1986 put forward the "nearby school" policy, "school district housing" came into being. In 2014, In response to the call of the Ministry of Education, Guangzhou launched the examination free admission for the nine-year compulsory education stage. In 2014, In response to the call of the Ministry of Education, Guangzhou launched the examination free admission for the nine-year compulsory education stage. However, due to the scarcity of educational resources, the housing prices in the school district continued to rise, which intensified the "Matthew effect" and deepened the class solidification [1], which made compulsory education evolve into "fight for father" and "fight for house".

At the opening of the fourth session of the 13th National People's Congress on March 5, 2021, Premier Li Keqiang delivered the Government Work Report, in which he pointed out: "We will continue to maintain the position that houses are for living in, not for speculation, and stabilize land prices, housing prices, and expectations. In the long run, we should encourage cities to increase the number of different types of rental housing, promote equal rights to purchase and rent, and promote the healthy development of the rental housing market." From a practical point of view, policies such as the lower limit of the down payment ratio and the reduction of the statutory deposit reserve ratio have played a certain role in restraining the growth of housing prices. However, as one of the forms of education investment, housing prices in school districts are still strong.

Based on this, this paper first studies the housing premium of key school districts in Guangzhou to non-key school districts. Existing literature studies the relationship between school districts and housing prices, but the influence of other unmeasured factors cannot be ruled out [2]. This paper selects the key school districts and non-key school districts of primary schools in Guangzhou as the research samples, and uses the breakpoint regression based on the Bayesian nonparametric method to exclude the influence of other factors such as room type and location. Furthermore, the differences in rents and house prices caused by only the school district were quantified, and the endogeneity problem was solved. It is able to robustly compare house price differences between different school districts in a statistical sense, and further analyze the relationship between high-quality education intensity, school rankings and school district housing premiums. In this way, the education department can increase high-quality degrees for specific school districts, and the housing construction department can control...
the housing prices in specific areas, and the difference in housing prices between the two school districts, which excludes other factors, can be used as a reference for parents to choose schools.

2. Literature review

2.1 The school district housing premium measuring method

Housing prices corresponding to high quality school districts may be higher due to better housing environment and location neighborhood factors, so using average price difference to estimate school district housing premium may overestimate the premium. Rosen and Fullerton [3] and Judd and Watts [4] used the characteristic price model to prove that house value is positively correlated with school quality. The hedonic price model highlights the capitalization effect of explanatory variables by adding a series of location features that affect housing value as control variables, such as the number of square meters, the age of the house, the number of bedrooms, and the availability of gardens. The influence is separated from the house character. However, Atkinson and Crocker [5] emphasized that it is difficult to fill in these housing characteristic variables, and it is difficult to obtain specific data, and this method cannot include unobservable variables. In response to this problem, Black [6] believed that the Ordinary Least Squares (OLS) estimate would be confounded by unobservable factors, so he proposed a method of selecting residential prices within a certain range at the junction of two school districts for research. Overcome the impact of some important neighborhood characteristic variables on residential prices. Subsequently, many scholars have studied the housing premium in school districts based on this method. Gibbons & Machin [7] used comprehensive breakpoint regression and instrumental variable method to find that every 1SD increase in school quality corresponds to 4%-9% increase in housing price. Bayer et al. (2007) [8] found that a 5% increase in school quality corresponds to a 1% increase in housing price after controlling the demographic characteristics of the community. Han Xuan found that the average education premium of the top 59 high-quality primary schools in Beijing was about 11% by means of breakpoint regression and dual difference regression, and this premium increased year by year, with a cumulative increase of more than 50% during the sample period. Conversion from a non-school district to a school district can significantly increase housing prices by 1.5 to 3.5 percent. In addition, many scholars also use Hedonic model [10]. Studied the influence of education distribution on school district housing price: Zhang Muyang [11] proposed the concept of rent rate discount, applied hedonic model and boundary fixed effect method to systematically analyze the rent rate discount of school district housing, and made a more accurate estimate of the willing to pay price of high-quality educational resources. The empirical results show that the rental yield of school district housing is 5% lower than that of non-school district housing on average, and can reach 19% for small-family housing. Sun Weizeng [12] adopted the Hedonic model based on fixed effects and concluded that the new primary school significantly reduced the premium rate of surrounding original school districts by 2.33 percentage points. Hu Wanyang [13] with the help of the admission system of "different rent-to-buy rights", conducted a paired regression between school district housing and adjacent non-school district housing, and accurately measured the premium of school district housing. The result showed that the premium of school district housing in Beijing's key primary schools was about 8.1% in 2011.

2.2 Research on an extension of the school district housing premium

From the perspective of the demand for educational resources, domestic and foreign research on the development of housing prices in school districts mainly focuses on the reasons for the increase in housing prices in school districts—that is, the study of capitalization of school quality. Oates [14] was the first to demonstrate that educational resources have a capitalizing effect on housing. Bogart & Cromwell found that the loss of a high-quality school in an area can reduce homebuyers' willingness to price in the area by nearly 10%. Zhang Hao [15] analyzed the phenomenon of education capitalization in my country's first-tier cities, and believed that education resources are contained in housing prices through capitalization. And it is concluded that historical inertia, urban expansion, unit
system and school district system are all important reasons for the differences in the allocation of educational resources in space and between groups. Wang Ying [16] from the perspective of the relationship between educational resource allocation and housing price, concluded that the influence coefficient of school district factors on housing price in Harbin is 0.427, while the influence coefficient of floor, hall, orientation, house age, decoration and other factors on housing price is 0.1. From the perspective of the supply of educational resources, Hilber, C. A., and C. J. Mayer [17], used the analysis of simultaneous equations to find that if an area has more developable land, the capitalization effect of school spending in housing will decrease. Similarly, Stadelmann & Billon [18] based on panel data analysis of 169 jurisdictions in Zurich, Switzerland, found that the capitalization effect of public spending decreases with the increase of construction land. Sun Weizeng and Lin Jiayu[12] used the newly built primary school in Beijing as a research sample and found that the newly built primary school significantly reduced the housing premium of the original surrounding school district by 2.33 percentage points. This change is localized and heterogeneous, and an increase in the number of houses in a school district cannot lead to a decline in housing prices when the price elasticity of demand in the school district is small.

3. Research design

3.1 Model construction

First of all, according to the district if there is a provincial key primary school, all school districts in Guangzhou city was divided into key school district and key districts, the introduction of Keele & Thitiunik[19] put forward the concept of complex processing, the adjacent province key school district room and a provincial key school district room of a pair of the research object, the provincial key school districts as treatment group, Non-provincial key school districts were set as the control group. In order to exclude the influence of housing location factors and neighborhood factors on housing prices, this paper calculates the difference in housing prices between the treatment group and the control group boundary (school district boundary), referring to the two-dimensional geographic breakpoint regression method proposed by Maxime Rischard [20]. The implementation will be divided into the following three steps: first, fit the smooth surfaces between house prices and geographic coordinates in the treatment group and control group respectively; second, use the fitting function to extrapolate the surface to the curve of the school district boundary; third, estimating treatment effects between pairwise extrapolated points on school district boundaries.

3.1.1. Smooth surface fitting

In the model of fitting house price and housing geographic coordinates, Gaussian distribution has the advantages of being able to select the required elements and perform conditional distribution operations. Therefore, Gaussian process regression is introduced in this paper to fit the relationship between housing geographical location and house price in the treatment group and the control group. The relationship between, set the following model:

\[
Y_i = m_{Distr[i]} + \gamma_{BuildClass[i]} + f_{Distr[i]}(s_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2_{\epsilon})
Y_i \sim \mathcal{N}(0, \sigma^2_{Y}) \quad \text{for } l=1,...,L_{BuildClass},
\]

\[
m_j \sim \mathcal{N}(0, \sigma^2_{m}), f_j \sim \mathcal{GP}
\]

\[
\text{for } j=1,...,J_{Distr}
\]

Among them, \( k \) is the covariance function of \( v = \frac{1}{2}, \), \( k(s,s') = \sigma^2_G \exp(-\|s-s'\|_2/l) \), where \( Y_i \) refers to the house price Taking the logarithm of the distribution, \( s \) is the spatial covariate composed of the geographic location coordinates of each house, and \( f(s_i) \) is the spatial Gaussian process, which performs Gaussian process regression for each school district. \( L_{BuildClass} \) is the type of house, which is treated as a dummy variable and is composed of covariates of the nature of the house.
Under this assumption, the set of hyperparameters $\theta = (l, \sigma_y, \sigma_{GP}, \sigma_e)$ is consistent for the treatment group and the control group, and in order to reduce the prior information of $m_j$, we set the value of $\sigma_m^2$ to be constant 20. Then, using the Gaussian process regression based on machine learning, by maximizing the marginal likelihood $P(Y \mid l, \sigma_y, \sigma_{GP}, \sigma_e)$ of the observed value, this paper calculates the value of the hyperparameter fitting.

### 3.1.2. Estimating extrapolated point treatment effects on school district boundaries

Suppose $b$ is a set of points on the boundary, $R$ is a representative point taken in the point set, which is called "sentry point" here, $b_{1:R} = \{b_1, b_2, \ldots, b_R\}$, taking the treatment group as an example, according to the properties of the Gaussian distribution and the Bayesian formula, we can obtain the posterior distribution of the sentinel house price of the treatment group on the boundary:

$$g_T(b_{1:R})|Y_T, \theta \sim N(\mu_{b_{1:R}}|T, \sum b_{1:R}|T)$$

$$\mu_{b_{1:R}}|T = K_{BT} \sum_{T}^{-1} Y_T \sum b_{1:R}|T = K_{BB} - K_{BT} \sum_{T}^{-1} K^T_{BT}$$

In the same way, the distribution $g_C(b_{1:R})$ of the sentinel points of the control group on the boundary, according to the definition of breakpoint regression, $\tau(b) = E[Y_{1T} - Y_{1C}|s_l = b]$ We can infer the estimator of cliff height $\tau(b_{1:R})$, which is the treatment effect, is defined as follows:

$$\tau(b_{1:R})|Y, \theta \sim N(\mu_{b_{1:R}}|Y, \sum b_{1:R}|Y)$$

$$\mu_{b_{1:R}}|Y = \mu_{b_{1:R}}|Y - \sum b_{1:R}|Y \sum b_{1:R}|Y = \sum b_{1:R}|T + \sum b_{1:R}|Y$$

Here $\tau(b_{1:R})$ is an $\uparrow R$-dimensional vector consisting of the processing effect of each sentinel point in point set B. The mean of the distribution $\tau(b_{1:R})$ is the index of the housing premium of the key school district in this paper.

### 3.1.3. Treatment effect based on inverse variance weighting

In the preceding part, we got the cliff height vector along the boundary, but for the convenience of processing and the need of further research, this paper will calculate the local mean weighted processing effect (LATE) and get the cliff height weighted by each value in the vector. According to the geometrical structure of knowledge on the two-dimensional space, covariance $\sum b_{1:R}|Y$ will reflect the boundary line segment on the correlation between adjacent sentinel will than those in the bending section of the correlation between the sentinel on the weak. The more relevant the sentinel, the less information it alone carries about the effect of local processing. Therefore, this paper assigns a greater weight to the sentinel points on straight line segments, and uses the following formula to calculate the weighted treatment effect of school district housing prices:

$$\tau^{INV}|Y, \theta \sim N(\mu^{INV}|Y, \sum \tau^{INV}|Y)$$

$$\mu^{INV}|Y = \frac{1}{\tau^T \sum b_{1:R}|Y} \frac{1}{\tau_T \sum b_{1:R}|Y} \tau^T \sum b_{1:R}|Y$$

### 3.2 Hypothesis testing

In this paper, the null hypothesis is set as the weighted cliff height (LATE) on the boundary of the school district to be 0. In order to carry out effective frequency testing, this paper uses the parameter guidance method based on the null hypothesis. First, the parameterized zero model $M_0$ is set. Under this setting, it can be considered that the boundaries of the treatment group and the control group are smooth and continuous, satisfying the null hypothesis. Then, in accordance with z test, this article selects $\mu^{INV}|Y$ as test statistics, in $b=1,2, \ldots, B$ iteration, $Y^{(b)}$ based on $M_0$ is obtained, calculated
using simulated rather than real spatial location information data \((s_i)\), so the ratio of the absolute value of \(\mu_{\text{INV}}|_{Y(b)}\) greater than \(\mu_{\text{INV}}|_Y\) is the p-value of the hypothesis test to be obtained in this paper:

\[
p^{\text{INV}} = p(\left| \mu_{\text{INV}}|_{Y(b)} \right| \geq \left| \mu_{\text{INV}}|_Y\right|) \approx \frac{1}{B} \sum_{b=1}^{B} \prod \left\{ \left| \mu_{\text{INV}}|_{Y(b)} \right| \geq \left| \mu_{\text{INV}}|_Y\right| \right\}
\]  

(5)

3.3 Placebo testing

The term placebo comes from randomized experiments in medicine and is often used to test the efficacy of a new drug. Based on this, a number of placebo test methods for the study of economic phenomena have been proposed in economics.

The placebo test in this paper requires multiple hypothesis tests on data without treatment effects. When the distribution of p-values calculated by multiple tests is consistent, we prove that the model is convincing and further illustrate the hypothesis inspection is valid. The specific approach of this article is:

Within each school district, the area is equally divided into two halves by drawing lines, and the angles of the lines are rotated from horizontal to counterclockwise. \(1^\circ, 3^\circ, 5^\circ, \ldots, 179^\circ\). Each time a line is drawn, \(\tau(b_1, R)\) and \(p^{\text{INV}}\) are calculated to check whether all \(p^{\text{INV}}\) value distributions are at the same level. If they are at the same level, we can think that (3.2) in Hypothesis tests are credible and price differences at school district boundaries are significant.

4. Data Processing and Empirical Analysis

4.1 Data processing and descriptive statistics

In terms of housing price data, this paper uses the python crawler tool to collect the second-hand housing data of all school districts in Guangzhou on Soufun.com, and finally obtains a total of 6120 pieces of data by deleting vacant housing prices, abnormal housing prices, and real estate data that cannot be geocoded. It contains data on the characteristics of house prices, locations, and apartment types in the school district, and maps the house addresses to the geocodes in the Baidu map one-to-one. The house price heat map is shown in Figure 1.

![Figure 1. Guangzhou school district housing price heat map](image)

Among them, the white curve represents the boundary of the school district. The information on the division of the school district is obtained from the official website of the Guangzhou Education Bureau. It can be seen from the figure that Yuexiu District and Tianhe District are the two districts with relatively dense distribution of houses and relatively high housing prices, and the division of school districts is more detailed.
In addition, this paper preliminarily processes the sample data and counts the average house price of each school district in each district. The results are shown in Figure 2.

![Figure 2. Sample Data School District Housing Price Statistics Chart](image)

The results show that the average house price in Tianhe District is the highest at 85015.21 yuan/m², followed by Yuexiu District at 74697.56 yuan/m², and the lowest in Liwan District, at 51329.28 yuan/m².

Using the python field extraction program, this paper crawls out the indicator of the distance from the subway station in the house sale information, and makes statistics on the house price per square meter and the distance between the house and the subway station. The descriptive statistics are shown in Table 1.

**Table 1. The house price per square meter of the house in the school district and the distance from the house to the subway station**

<table>
<thead>
<tr>
<th>The distance from the house to the subway station/m</th>
<th>house price yuan/m²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>3927</td>
</tr>
<tr>
<td>Mean</td>
<td>590.75</td>
</tr>
<tr>
<td>Std</td>
<td>266.35</td>
</tr>
<tr>
<td>Min</td>
<td>34.00</td>
</tr>
<tr>
<td>25%</td>
<td>402.00</td>
</tr>
<tr>
<td>50%</td>
<td>568.00</td>
</tr>
<tr>
<td>75%</td>
<td>780.00</td>
</tr>
<tr>
<td>Max</td>
<td>1199.00</td>
</tr>
</tbody>
</table>

Overall, the sample school districts have obvious location advantages, they are close to the subway station, and the variance of the distance index from the subway station is 266.35. It can be seen that the location advantages of different school districts are quite different; for the house price per square meter data, the selected five districts. The average price per square meter is 47972.23, and the variance is 34013.62. It can be seen that the price difference of houses in different school districts is huge, but this price difference is not all caused by the different school districts where the houses are located. Housing premiums in different school districts due to quality.

### 4.2 Empirical result

This paper first selects the school district pair of Tianhe Longdong School District and Baiyun Yongping School District as the initial sample, and draws the histogram of house prices per square meter in the two school districts before performing Gaussian process regression.

As shown in Figure 3, the number of high-priced houses in Tianhe District is more than that in Baiyun District, but in general, the number of middle-priced houses in Baiyun District exceeds that in Tianhe District.
Using the unit area price data of the two school districts and the housing geographic coordinates to fit the linear regression model of the Gaussian process and the intercept term, the regression model parameters \( \theta = (l, \sigma_y, \sigma_{GP}, \sigma_e) \) correspond to \( \theta = 0.0043, 0.1278, 0.4210, 0.1273 \), we substitute the coordinates of the school district boundary sentinel point into the regression model to obtain the difference between the boundary treatment group and the control group, and draw Figure 4.

The y-axis of the left figure corresponds to the logarithmic difference of the housing price per unit area between the two school districts. A positive value means that the housing price of Tianhe Longdong School District is higher than that of Baiyun Yongping School District. The y-axis on the right corresponds to the housing price ratio between the Longdong school district and the Yongping school district. The light red in the background represents the posterior correlation of the boundary points. It can be found that for the selected 100 sentinel points, the correlation between two sentinel points is high, which explains the accuracy of the measurement to a certain extent. The picture on the right shows the location of the sentinel on the boundary between the two school districts. It can be seen that the selected 100 sentinel points are evenly distributed on the boundary of the school district, the index of the southernmost sentinel is 1, and the index of the northernmost sentinel is 2.
Next, in order to present a more intuitive single house price premium data, this paper uses the inverse variance weighting method to calculate an average value of the house price difference data at the above boundary sentinel points. It was found that the housing price from Longdong School District to Yongping School District will increase by about 42.3%. And this regression result passed the test. In the inverse variance weighted hypothesis test, it showed that the result was established at the 1% significant level. The results are shown in Table 2.

**Table 2. Inverse variance weighted average and significance test results**

<table>
<thead>
<tr>
<th>indicators</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inverse variance weighted average</strong></td>
<td><strong>( \tau \mid Y \sim \text{Normal}(\mu = 0.423, \sigma = 0.160) )</strong></td>
</tr>
<tr>
<td>( P(\tau &gt; 0 \mid Y) )</td>
<td>99.596%</td>
</tr>
<tr>
<td><strong>Significance testing</strong></td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.0026</td>
</tr>
</tbody>
</table>

To further test the robustness of the results, a placebo test was performed in this paper, and the results are shown in Figure 5:

**Figure 5.** District housing premium between Longdong School District (key school district) and Yongping School District (non-key school district)

The three subgraphs from top to bottom in Figure 5 are the results of the placebo test under the chi-square distribution test, the mll bootstrap test and the inverse variance bootstrap test, respectively. The data under the labels 27 and 19 represent the placebo test results of the treatment group and the control group, respectively. If the hypothesis testing method is valid, the distribution of p-values in the treatment group and the control group in the placebo test should be roughly the same, that is to say, it is not only the treatment effects in the school district boundaries that produce specific P-values (p-values are not biased). The hypothesis testing distribution calculated for the treatment group with two unrelated regions randomly drawn should be roughly consistent with the hypothesis testing distribution calculated for the control group.

As shown in the results in Figure 5, whether it is the chi-square distribution test, the mll bootstrap test or the inverse variance weighted bootstrap test, it can be found that the p-value distributions of the treatment group and the control group are generally consistent, so the hypothesis test passes the placebo test. This further proves that the model setting of the Gaussian process regression above is correct.
5. Conclusion and suggestion

Firstly, according to the sampling plan, this paper conducts on-the-spot investigation on the resident population of the residential area in Guangzhou. The survey results show that most people think that the setting of school district housing aggravates the inequity in education, and the negative external effect brought by the setting of school district housing is greater than the positive external effect; Secondly, this paper assumes that there is indeed a premium for housing in the school districts in Guangzhou, and measures this premium. This measurement result excludes the influence of location factors such as housing location and physical factors such as property, and calculates the changes in housing prices due to differences in school districts. And the empirical results were tested for statistical significance (including inverse variance weighted test, mll bootstrap test, chi-square test) and placebo test.

According to the research conclusions of this paper, we put forward the following suggestions: On the one hand, the lack of high-quality educational resources is an important reason for the rising housing prices in China's urban school districts, and local governments' increased investment in educational resources will help improve this situation from the supply side;

At the same time, with the improvement of teachers' salaries and the increase of high-quality educational resources, the housing premium in the school district will tend to be more reasonable in the long run, and the equalization of educational resources will be truly realized from the two dimensions of quantity and quality. On the other hand, the housing market in school districts has significant localization characteristics. The allocation of educational resources needs to take into account the spatial allocation of educational resources within the city at the same time. A balanced allocation of resources can help reduce social separation caused by residential clusters.

Authors’ contributions

In view of the real problem of housing premium in school districts, this paper selects the housing data of Guangzhou primary school districts as a research sample, quantifies it and discusses the impact of education supply on it, so as to provide references and suggestions for subsequent policy introduction.

References


[12] Sun Weizeng, Lin Jiayu. “Can the supply of educational resources reduce the housing premium in school districts? —— Evidence from the new primary schools in Beijing”. [*Economics (Quarterly)*, 2020, 19(02): 499-520]


