Impact of Air Quality on Housing Prices: A Machine Learning Approach

Zulin Feng[1*, a], Xinlong Li[2, b]
* Corresponding author: [a]e0792362@u.nus.edu, [b]xinlong.li@ntu.edu.sg

[1]Department of Economics, National University of Singapore, Singapore

[2]Nanyang Business School, Nanyang Technological University, Nanyang, Singapore

Abstract: This paper investigates whether air quality could impact people's willingness to pay in the housing market. Leveraging OLS regression, Random Forest, Lasso, and Ridge regression, we find a negative correlation between housing prices and air pollution. We use Air Quality Index to measure air quality in our study. Higher AQI represents a higher pollution level. Specifically, one unit increase in AQI reduces the total price of houses by 21370 yuan. Furthermore, our findings reveal that the height of buildings is positively related to residents' concern about air quality when they make housing purchase decisions. In particular, the higher level the resident's house locates, the more attention the buyers put pay to air quality. Also, we notice that the buyers from the high total price group are more willing to pay for clean air. As for housing price prediction, we confirm that adding AQI as one of the predictors can increase the accuracy of Beijing housing price prediction even though its effect is relatively small. The findings of this paper may contribute to practical implications in environmental policies, residential satisfaction, and housing market regulations.

Keywords: AQI, housing price prediction, OLS regression, Random Forest, Lasso, Ridge

1. Introduction

This paper detects how Air Quality Index (AQI) affects the marginal willingness to pay for clean air by using housing prices and air quality data from Beijing. We find that AQI has a significant and negative impact on Beijing's housing prices. In addition, this study applies several machine learning models: Lasso, Ridge, and Random Forest to assess the impact of AQI on Beijing housing price prediction.

This research serves to deepen our understanding of air pollution's capitalization effects on housing prices in Asian metropolises. Also, quantifying the willingness to pay for a cleaner living environment could help the government to create sustainable policies and urban planning for air pollution control.

Moreover, when influencing factors for residential requirements are further cleared and realized, the health of the housing market can be enhanced.

BDEIM 2022, December 02-03, Zhengzhou, People's Republic of China

DOI 10.4108/eai.2-12-2022.2328696

This paper is organized as follows: Following the introduction, part II shows our research motivation. Part III demonstrates the data processing, and then part IV offers our OLS, robustness check, and cluster results. We discuss our findings in part V. Part VI emphasizes the economic costs of AQI from three aspects and part VII summarizes the results of machine learning models. Finally, part VIII provides the conclusion and limitations of this paper.

2. Motivation

Environmental problems arouse public attention over years. Beijing, the capital of China, suffers from severe air pollution problems caused by the rapid growth of urbanization, motorization, and industrialization in recent years [1]. Even though people cannot purchase clean air directly, property values can be affected by the air quality near the housing unit. It was confirmed that air quality becomes a more and more influential factor when people make their residential decisions [2]. Moreover, it was found that bad air quality reduces residents' visits to restaurants and shopping areas near their properties [3]. This results in an evident decline in profits, which is regarded as a "pollution tax". Thus, we are also curious about the specific value of the "pollution tax" in the Beijing housing market.

However, most existing studies in China regard single-family housing prices as the dependent variable, but they do not realize that residential communities can be considered as research objects [2]. Our study investigates 17 communities in Beijing to fill this gap. A study states that Beijing has popular centralized community development due to its urban amenities and convenient transportation [4]. Thus, it is suitable to take residential communities as one of our independent variables into regressions.

3. Data Processing

This study collected 520045 housing transactions from Beijing city during 2013-2020, including 44 columns. And details of key variables are specified in Table I. Besides, we conduct basic data cleaning by following procedures.

3.1 Data Cleaning

We delete "InteriorSize" because of its 67.9% missing values. This variable is composed of three parts: the usable area of the house inside the suite, the wall area inside the suite, and the construction area of the balcony inside the suite.

As for replicated data, if the transaction happens in the same community, and is sold on the same trade date and same price, we define the transaction as replication and only remains one of them. Finally, we delete 3576 replications.

For variable "Price", it represents how much the house is per square. Considering its meaning, we plan to delete those observations with extremely high or low prices. Specifically, we regard outlier

boundaries as +-3 standard deviations from the mean. Hence, we delete values that are greater than mean +3 standard deviations from the mean and less than mean -3 standard deviations, and we eventually delete 5673 rows, which accounts for 1.1% of total rows.

As for the pollutant indicator: "AQI_Month_closest", it represents the air quality index near the property during the traded month. And we observe this variable contains negative values of -1 and infinity values. Hence, we choose to exclude those irregular values and delete 30071 rows in total, which accounts for 5.84% of total rows.

3.2 Data Transformation

Firstly, we notice that the column of "Floor" in the original data only records the total level of the house and whether the house is located on a high/medium/low level. Hence, it is hard for us to do statistical analysis, and we decide to calculate the rough level of the house based on the information provided in the "Floor" column. Specifically, the "Floor" column only records the total floor level of the house and whether the house is located in the basement/low-level/medium-level/high-level/top-level. For houses located in a basement, we directly record the floor level of the house as "-1". However, if the house is located on the top level and the total floor level of the house is 6, we will regard "6" as the floor level of the house. As for houses located on low-level, we will use the total floor level divided by 3 to get the floor level of the house. Similarly, for houses located on medium levels, we will use the total floor level divided by 2 to get the floor level of the house. In terms of houses located on a high level, we use the total floor level multiplied by 5/6 to get the floor level of the house. The reason is that the minimum number of total floor levels is 6. When 6 times 5/6, we can get a floor level of 5 and record it as the floor level of the high-level house. After these simple calculations in python, we can generate the new column of "Floor_num", which means the exact number of floor levels.

Secondly, to make categorical variables available for the following numerical analysis, we create 11 dummy variables: "Elevator", "BuildingType", "HeatingType", "PropertyType", "PropertyType2", "FiveYearsProperty", "Ownership", "BuildingStructure", "Renovation", "UnitType" and "TradeYear".

3.3 Endogeneity Problem

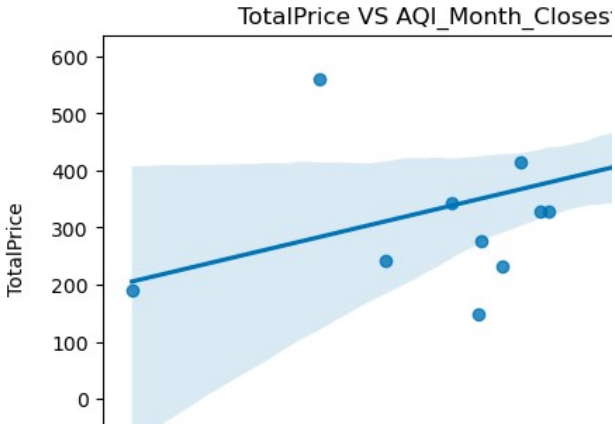

Figure 1. Scatter plot of TotalPrice VS AQI_Month_Closest

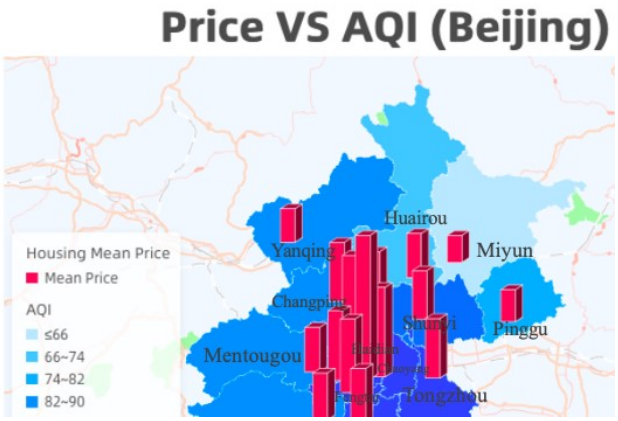

Figure 2. Picture of Price VS AQI in Beijing

Additionally, we visualize the univariate relationship between the dependent variable “TotalPrice” and the independent variable “AQI_Month_closest”. Fig.1 and Fig.2 above both show there is an obvious positive relationship between the total price of houses and AQI, which contradicts our economic intuition and implies an endogeneity problem may exist. The endogeneity problem occurs because the total price of houses is affected by many factors. And these factors may be correlated with “AQI_Month_closest”, which results in multicollinearity issues among variables and finally leads to the questionable causal inference above. Hence, we use the variance inflation factor (VIF) to check for multicollinearity. It is found that “Washingroom”, and “Bedroom” have rel-

atively large VIF, which all exceed 10. After deleting these two variables, the highest VIF is 8.62, which is below the threshold of 10 [5], indicating that multicollinearity is not a major concern in this study. Also, we build the following OLS regression to test whether the endogeneity problem does exist or not. The result shows that after adding sufficient total price-related variables, the relationship between "TotalPrice" and "AQI_Month_closest" turns out to be significantly negative.

4. Regression Analysis

4.1 Model Specification

We selected 25 variables to establish the following OLS model:

$$y = \alpha + \sum_{i}^{25} \beta_i x_i + \varepsilon$$

where y represents the total price of houses (the unit is hundred thousand yuan), α is the unknown intercept; β_i denotes the coefficients for all variables; ε is the error term. x_i represents the 25 variables as shown in Table 1.

Table 1 The Data Description of the 25 Variables

Variable	**Description**
TradeYear	which year the house is traded
BuildingStructure	steel-concrete /composite /brick-concrete/masonry-timber/pure steel/frame/unknown structure
Floor_num	floor level of the house
Lng	longitude of the house location
Lat	latitude of the house location
Views	how many times viewed by home buyers
PriceChange	how many times of price change
RoomShowing	the number of times the house has been shown online
Followers	how many buyers show willingness to purchase the house
Elevator	whether the building the house located in has elevator
HeatingType	central heating/self-heating
PropertyType	commercial residential building/not commercial residential building (including center delivery room housing, etc)

PropertyType2	common residence/not common residence (in-cluding serviced apartment, etc)
FiveYearsProperty	building age is over/below 5 years
Ownership	common/not in common
BuildingType	slab-type/bungalows/tower building/
Livingroom	the number of living room
Parkinglot	the number of parking lot
Size_num	size of the house
Only_South	whether the house is located in south
South_North	whether the house is located in south north
StairsUnitRatio_num	the ratio of the number of elevators to the number of households
Renovation	hardcover/plain/blank/others
County	which county the house located in
AQI_Month_Closest	air quality index near the property during the traded month

Regarding OLS regression results, the R-squared of 0.684 tells us that 68.4% variations in Beijing housing prices can be explained by the 25 variables. Noticeably, Table 2 shows that the coefficient of "AQI_Month_closest" is -0.2137, which means if AQI rises by 1 unit, then housing prices will decline by 21370 yuan. The p-value of 0 implies that "AQI_Month_closest" is a statistically significant feature for predicting housing prices.

Table 2 Regression Results of Key Variables

Key Variables	**Coefficient**	**P-value**
AQI_Month_closest	-0.2137	0
Floor_num	0.8995	0
Views	0.0010	0
PriceChange	-5.5741	0
RoomShowing	0.3997	0
Followers	-0.1016	0
Elevator	50.5333	0
Heatingtype	41.5637	0

FiveYearsProperty	8.8993	0
Ownership	-5.5069	0
Livingroom	38.5887	0
Size_num	4.2772	0

4.2 Robustness Check

To do the robustness check, we eliminate counties where housing transactions are below 50. The reason is that 13 of 17 counties all have more than 4000 housing sales, however, the remaining 4 counties' sales are even lower than 50. To make sure our data is more representative, we delete these four counties, which are "Pinggu", "Miyun", "Huairou", and "Yanqing" respectively. As a result, AQI remains statistically significant for predicting housing prices. Thus, we can say that our result is robust after excluding some special values.

5. Findings

5.1 Total Price Classification

According to the median of the total price of all houses, we classify all transactions into two main categories, i.e., houses sold at high and relatively low prices. Under this classification, we can observe that the negative correlation between "TotalPrice" and AQI is much stronger in the high-price housing group, however, the negative correlation is weaker in the low-price housing group. The logic for this observation is that the high-price housing group may represent buyers who own high consuming ability, in contrast, the low-price housing group may consist of buyers who do not have such a high consuming ability. We can infer that residents' demand for better air quality is increasing with a higher level of income. It indicates that the rich show a higher marginal willingness to pay for cleaner air than the poor even though their research area focuses on Shanghai's housing market, as in [6]. Shanghai and Beijing are two of the most prosperous Chinese cities, it is not odd to see similar consuming behaviors in their residents. But details of this phenomenon in Beijing city may need further research.

Table 3 Total Price Classification

TotalPrice Classification	**AQI**
>=TotalPrice Median	-0.1784
<TotalPrice Median	-0.0622

5.2 Floor Level Classification

According to the three quantiles (25%, 50% & 75%) of floor level, we classify all transactions into three main categories, i.e., houses located on high level, medium level, and low level. Interestingly, Table 4 above presents that the higher the floor level is, the stronger the negative correlation is. In other words, buyers who choose high-floor living would be more concerned about air quality, and vice versa.

Table 4 Floor Level Classification

Floor Level	AQI
Low	-0.1245
Medium	-0.2609
High	-0.3616

The result caught our attention since the stereotype tells us that residents living on lower levels are frequently exposed to large amounts of exhaust emissions from street-level traffic. In contrast, residents on higher-level floors may enjoy cleaner air. But the regression result breaks this stereotype. The logic is that in high-density urban surroundings, the increase in building height would be accompanied by a growing risk of air pollution [7]. Moreover, this evidence is based on selecting a high-density urban area of Nanjing, China as the research object. And our research area: Beijing also fits the high-density urban form, which makes its result more explanatory for our observation.

6. Empirical Evidence of AQI impact & Economic Consequences

Along with growing environmental protection awareness, it is significant to qualify the economic value of air pollution, and its importance is shown from three perspectives of government, society, and the housing market.

For the government, air pollution not only carries increasing human healthcare costs but also lower economic productivity and welfare. According to a 2020 report by the Center for Research on Energy and Clean Air [8], the economic costs of air pollution from fossil fuels have already been as high as 6.6 percent of China's GDP.

As for society, Beijing has received an unprecedented level of public concern about smog ravage since 2013 [9]. In addition, considerable field survey questionnaires are collected to investigate how air quality influences college students' willingness to stay in Beijing [10]. The result shows that college graduate students are sensitive to health loss caused by air quality, which results in a decline in their willingness to stay in Beijing. It states that college students are regarded as the main force of reserve talents for Beijing's sustainable economy, however, in the long term, the decline of intention to stay would harm the economic development of Beijing, as in [10]. Therefore, we can become more confirmed that it is meaningful to know how much the marginal willingness to pay for better air quality is and accordingly adjust the environmental policies to encourage the involvement of citizens and consumers.

Furthermore, over the last decade, China's housing market experiences significant inflations. As in [4], it states that the first step to implementing more effective real-estate policies is to explore the relationship between amenity values and property values. The paper also explains that amenity value does not only mean structural or locational factors but also includes environmental factors

near the property. Therefore, it is meaningful to evaluate the effects of AQI on the housing price prediction to stabilize the real-estate economy of Beijing.

Past literature mainly aims to evaluate the direct harm of environmental pollution on public health [11]. Nowadays, more and more studies tend to examine the economic value of clean air and welfare costs for air pollution. However, most existing studies mainly focus on how AQI influences housing prices without measuring its effects on the housing price prediction. Our study would simultaneously explore both problems, and we believe that our study will make the housing market clearer about the economic value of air quality.

7. The Effect of AQI on Beijing's Housing Price Prediction

To examine the effects of AQI on Beijing's housing price prediction, we apply three machine learning models: Lasso, Ridge, and Random Forest. Table 5 shows that AQI can improve the accuracy of housing price prediction with a lower root mean squared error (RMSE). The model with the best forecast performance is Random Forest, whose RMSE only accounts for 1/4 of the total price mean. Also, its RMSE has the largest reduction after adding AQI as one of the predictors, which is from 111.93 to 111.26.

Table 5 Results of Machine Learning Models

RMSE	**Random Forest**	**Lasso**	**Ridge**
Without AQI	111.93	227.2	227.21
With AQI	111.26	227.16	227.17
TotalPrice mean	457.89		

8. Conclusion & Limitations

In summary, two issues are addressed in our study: Firstly, after excluding the endogeneity problem, AQI has a significantly negative impact on Beijing housing prices. This result is consistent with current empirical studies. But our result is based on adding "Community" as one of our research objects. Secondly, we also quantify the effect of AQI on the prediction of Beijing housing prices. And we find that even though its influence is statistically small, with the increasing social concern about environmental protection, the air quality index may occupy a larger proportion in the housing price forecast in the future, which needs further investigation. Furthermore, there are two findings from our research: Firstly, residents' demand for cleaner air may be closely associated with a higher level of income. Secondly, the height of a building is positively related to risks of air pollution, which reflects in the higher demand for clean air for high-level residents. Nevertheless, these findings could be expanded based on our research.

Besides, there are still two limitations in this study: Firstly, our study only focuses on the housing sale market in Beijing. It points out that especially in China housing market, most current literature ignores comparative studies of the effects of AQI on both selling prices and rental prices [2]. Our study also does not consider Beijing rental market due to data constraints. Secondly, we focus on the static relationship between the total price of houses and AQI. Future research could combine with Beijing's air control policies and analyze the variation of AQI and its impact on housing prices over time to generate a more dynamic analysis.

References

[1] H. Zhang, S. Wang, J. Hao, X. Wang, S. Wang, F. Chai, & M. Li, (2016). Air pollution and control action in Beijing. Journal of Cleaner Production, 112, 1519-1527.

[2] R. Liu, C. Yu, C. Liu, J. Jiang, & J. Xu, (2018). Impacts of haze on housing prices: an empirical analysis based on data from Chengdu (China). International Journal of Environmental Research and Public Health, 15(6), 1161.

[3] C. Sun, S. Zheng, J. Wang, & M. E. Kahn, (2019). Does clean air increase the demand for the consumer city? Evidence from Beijing. Journal of Regional Science, 59(3), 409-434.

[4] Y. Xiao, X. Chen, Q. Li, X. Yu, J. Chen, & J. Guo, (2017). Exploring determinants of housing prices in Beijing: An enhanced hedonic regression with open access POI data. ISPRS International Journal of Geo-Information, 6(11), 358.

[5] W. Jabr, and Z. Zheng, 2014. "Know Yourself and Know Your Enemy: An Analysis of Firm Recommendations and Consumer Reviews in a Competitive Environment," MIS Quarterly (38:3), pp. 635-654.

[6] J. Chen, Q. Hao, & C. Yoon, (2018). Measuring the welfare cost of air pollution in Shanghai: evidence from the housing market. Journal of Environmental Planning and Management, 61(10), 1744-1757.

[7] J. Yang, B. Shi, Y. Shi, S. Marvin, Y. Zheng, & G. Xia, (2020). Air pollution dispersal in high density urban areas: Research on the triadic relation of wind, air pollution, and urban form. Sustainable Cities and Society, 54, 101941. doi:10.1016/j.scs.2019.101941

[8] L. Myllyvirta, (2020). Quantifying the economic costs of air pollution from fossil fuels. Centre for Research on Energy and Clean Air.

[9] X. Chen, S. Shao , Z. Tian, Z. Xie, & P. Yin, (2017). Impacts of air pollution and its spatial spillover effect on public health based on China's big data sample. Journal of cleaner production, 142, 915-925.

[10] Y. Hao, X. Hao, Y. Li, Y. Zhang, & H. Wu, (2020). How does air quality affect the willingness of graduate students to stay? Evidence from Beijing city, China. Journal of Cleaner Production, 259, 120759.

[11] H. Kan, & B. Chen, (2004). Particulate air pollution in urban areas of Shanghai, China: health-based economic assessment. Science of the Total Environment, 322(1-3), 7